Universitat d'Alacant
Universidad de Alicante

ellis | UNIT ALICANTE

A Sociotechnical Approach to Trustworthy AI: from Algorithms to Regulation

Adrián Arnaiz Rodríguez

Thesis presented in fulfillment of the requirements
for the degree of Doctor of Philosophy by the

UNIVERSITY OF ALICANTE

With international mention

DOCTOR OF INFORMATICS

Advised by:
Nuria Oliver Ramírez, *ELLIS Alicante*
Miguel Ángel Lozano Ortega, *University of Alicante*

The research presented in this thesis has been financed by the ELLIS unit Alicante Foundation with funding from the European Commission under the Horizon Europe Programme - Grant Agreement 101120237 - ELIAS, from a nominal grant from the Regional Government of Valencia in Spain (Convenio Singular signed with Generalitat Valenciana, Conselleria de Innovación, Industria, Comercio y Turismo, Dirección General de Innovación), from a grant by the Banco Sabadell Foundation, and from Intel via RESUMAIS, the Center of Scientific Excellence in Responsible AI. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

Universitat d'Alacant
Universidad de Alicante

ellis | UNIT ALICANTE

A Sociotechnical Approach to Trustworthy AI: from Algorithms to Regulation

Adrián Arnaiz Rodríguez

Tesis presentada para aspirar al título de doctor por la

UNIVERSIDAD DE ALICANTE

Mención de doctor internacional

DOCTORADO EN INFORMÁTICA

Dirigida por:
Nuria Oliver Ramírez, *ELLIS Alicante*
Miguel Ángel Lozano Ortega, *University of Alicante*

La investigación presentada en esta tesis ha sido financiada por la Fundación de la Comunitat Valenciana Unidad ELLIS Alicante con fondos del programa Horizon Europe de la comisión europea - Acuerdo de subvención 101120237 - ELIAS, de una subvención nominativa por parte del gobierno regional de Valencia en España (Convenio firmado con la Generalitat Valenciana, Conselleria de Innovación, Industria, Comercio y Turismo, Dirección General de Innovación), de la fundación Banco Sabadell y de Intel, en virtud de RESUMAIS, el Centro de Excelencia en Inteligencia Artificial Responsable. Las opiniones expresadas son exclusivas del autor(es) y no reflejan necesariamente el punto de vista de la Unión Europea o de la HaDEA. La Unión Europea no otorga autoridad para ello.

This document was proudly typeset with LaTeX.

Contact Details

Adrián Arnaiz Rodríguez
adrian@ellisalicante.org, arnaiztech@gmail.com

Abstract

This thesis proposes a **sociotechnical framework** for the effective implementation of *Trustworthy Artificial Intelligence (TAI)*, addressing the technical, human, and regulatory dimensions of AI-induced harms. Rather than treating TAI as a purely technical goal, we emphasize its *interdisciplinary nature*, aligning algorithmic development with societal needs and legal norms throughout the AI system life-cycle. Our approach focuses particularly on harms and investigates how they can be mitigated through algorithmic design and effective development and implementation of regulations.

From a **technical** standpoint, we focus on harms derived from *discrimination in algorithmic decisions*. We first introduce `FairShap`, a novel data valuation method that quantifies the contribution of individual training examples to group fairness decision-making metrics. This enables a more complete diagnosis and mitigation of discrimination in high-risk decision-making systems, aligning with auditing obligations under the EU AI Act.

Moving beyond the fairness definitions in decision-making, we also propose `ERG`, a graph-based approach to measure and mitigate structural disparities in social capital within social networks. This approach addresses emerging regulatory demands, such as those set out in the EU Digital Services Act, which require assessing and reducing systemic risks in online platforms.

In the context of algorithm **use**, we design and evaluate a *human-AI complementarity* framework for collaborative decision-making in high-stakes resource allocation tasks. By combining human and algorithmic matching decisions and optimizing the hand-off using bandit-based strategies, we explore how semi-automated systems can be designed to outperform humans or algorithms alone. This approach adheres to the TAI principles of technical robustness, user oversight, and minimal harm, as set out in EU GDPR and TAI guidelines.

Finally, in the **governance** sphere, we examine the use of *AI for worker management under Spanish labor law*. We identify applicable legal frameworks across the AI system life-cycle; analyze the alignment between TAI principles, the EU AI Act, and labor duties; and highlight tensions such as the gap between correlation-based models and the legal requirement for causal justification in certain decisions.

Taken together, these contributions demonstrate that ensuring trustworthiness requires more than just algorithmic improvements. Instead, it must be understood as a sociotechnical system, emerging from the interaction of data, algorithms, institutions, and regulatory constraints. This thesis provides practical insights for researchers, practitioners, and policymakers seeking to develop AI systems that are technically robust, socially aligned, and legally compliant.

Resumen

Esta tesis propone un **marco sociotécnico** para la implementación efectiva de **la Inteligencia Artificial fiable**, que aborda las dimensiones técnicas, humanas y regulatorias de los daños inducidos por la IA. En lugar de considerar la IA fiable como un objetivo puramente técnico, enfatizamos su naturaleza interdisciplinaria, alineando el desarrollo algorítmico con las necesidades sociales y las normas legales a lo largo del ciclo de vida del sistema de IA. Nuestro enfoque se centra particularmente en los daños provocados por la IA e investiga cómo promoverlos no solo mediante el diseño algorítmico, sino también mediante el correcto desarrollo y aplicación de la normativa.

Desde un punto de vista **técnico**, nos centramos en los daños derivados de la discriminación en las decisiones algorítmicas. Primero, presentamos un método de valoración de datos que cuantifica la contribución de cada datos de entrenamiento a las métricas de justicia algorítmica grupal. Esto permite un diagnóstico más completo y la mitigación de la discriminación en sistemas de toma de decisiones de alto riesgo, en consonancia con las obligaciones de auditoría de la Regulación de IA de la UE.

Más allá de la definición de justicia en la toma de decisiones proponemos un enfoque para medir y mitigar las disparidades estructurales en el capital social dentro de las redes sociales. Este enfoque aborda las demandas regulatorias emergentes, como las establecidas en la Ley de Servicios Digitales, que exigen la evaluación y reducción de riesgos sistémicos en las plataformas en línea.

En el contexto del **uso** de algoritmos, diseñamos y evaluamos un marco de colaboración humano-IA para la toma de decisiones en tareas de asignación de recursos. Al combinar decisiones humanas y algorítmicas optimizando cuántas asignaciones hace el humano mediante estrategias basadas en sistemas de *bandit*, exploramos cómo se pueden diseñar sistemas para obtener mejor rendimiento que los humanos o a los algoritmos por sí solos. Este enfoque se adhiere a los principios de robustez técnica, supervisión del usuario y daño mínimo, tal como se establece en el RGPD o las directrices de la IA.

Finalmente, en el ámbito de la **regulación**, examinamos el uso de la IA para la gestión de trabajadores bajo la legislación laboral española. Identificamos los marcos legales aplicables a lo largo del ciclo de vida del sistema de IA; analizamos la alineación entre la IA fiable, la Ley de IA y las obligaciones laborales, destacando tensiones como el dilema correlación-causalidad y el requisito legal de justificación causal en ciertas decisiones.

En conjunto, estas contribuciones demuestran que garantizar el desarrollo de una IA fiable requiere más que simples mejoras algorítmicas. Más bien, debe entenderse como un sistema sociotécnico que surge de la interacción de datos, algoritmos, instituciones y restricciones regulatorias. Esta tesis proporciona una perspectivas práctica para investigadores, profesionales y legisladores que buscan desarrollar sistemas de IA técnicamente robustos, socialmente alineados y legalmente compatibles.

Acknowledgments

Firstly, I would like to express my deepest gratitude to my supervisor, Nuria Oliver, for her scientific guidance, generous support and unwavering trust throughout my PhD journey. Your mentorship has shaped both this thesis and the researcher I have become. I am especially grateful for the freedom and confidence you gave me to develop independently, and for supporting me through unexpected challenges. Your commitment and humanity went far beyond the responsibilities of a supervisor, and I feel fortunate to have shared this path with you.

We are social animals, and science is never a solitary pursuit. I'm deeply grateful to all collaborators, co-authors, and colleagues who contributed to this thesis. I'm especially thankful to K. Schweighofer, J. Losada, F. Errica, and A. Velingker for the dynamic and often playful exchanges that rekindled the curiosity that first drew me to research. I also thank the learning-on-graphs community for its openness and collaborative spirit, which reminded me that science can still be generous and shared. Looking back, I want to thank the Admirable and Goonies labs at Universidad de Burgos for introducing me to the scientific path and planting the seed of curiosity. You opened a door and everything else followed from that beginning.

In discovering science as a joyful and collaborative pursuit, I understood what François Le Lionnais meant when he said *Science is an Art*. It is not solely driven by rigor, but by imagination, curiosity, a love of play and ideas, analytic spirit, creative synthesis, and radical non-conformism. These traits did not originate in science — which I only discovered in my twenties — but in my education and the experiences I shared with those closest to me.

To my parents and family, thank you for everything: your effort, work ethic, and unconditional quiet pride in things well done, without ever forgetting to care for the essential and joyful aspects of life. My earliest curiosity came not from textbooks, but from days spent in serigraphy, playing music, cooking or fishing in the living room. Your altruistic love of culture and your joyful perspective on life gave me more than any textbook ever could. If I've made it this far, it's because I've spent my life trying to become half the worker you are.

Although they'll probably never read this, I want to thank my friends for their down-to-earth support and the 'what do you even do at work?' conversations that reminded me this whole thing is just one version of life. Thanks for keeping it all in perspective.

They say we save the best for last. To Sara, my life partner: this journey belongs to both of us. You've walked every step beside me, and when the path became unclear, you brought clarity into my daily life and reminded me of what truly matters. I'm equally grateful to walk your path with you, witnessing your experiences, strength, and how you live with integrity and depth. Your stories could fill volumes, and each would teach something about what it means to care, endure, and live meaningfully. Through your presence and how you engage with me and the world, I've come to understand what Viktor Frankl meant when he wrote that *ultimate meaning* arises in love,

responsibility, and how we respond to suffering. You've shown me that meaning isn't something we grasp through reflection alone, but something we embody in how we stand by others, act compassionately, and find purpose even in hardship. We didn't walk this road for the destination, but because walking it together mattered: "*somos el tiempo que nos queda, la vieja búsqueda, la nueva prueba*".

This thesis is the result of years of work, where I learned that science is not just a discipline, but a form of expression — a pursuit not only of knowledge, but also of meaning, integrity, and connection. Thank you everyone.

Adrián
Arnaiz Rodríguez González Gil
Alicante, July 2025

> "And if, from the beginning of the history up to the present day, the human adventure, this amazing human adventure, has been what we all know it to be, for better or worse, it is due not only to science as science but also to the fact that science is an art."
>
> François Le Lionnais, *Science is an Art*

Contents

Papers related to this thesis

Part of the material presented in this dissertation is submitted or has appeared in peer-reviewed conference, journal publications, and tutorials. Below there is a list of the scientific publications that were published during the doctoral period, structured in three groups: *(a) Core publications*, which constitute the backbone of one or more chapters in this thesis; *(b) Supporting publications*, which stem from the same line of research and inform the thesis but are not included within the main text. They appear in the appendix for self-completeness; and *(c) Collaborative publications* that are the result of scientific collaborations where I am not the main contributor but are not part of this thesis.

We also indicate if the publications are already accepted at any venue, or instead they are submitted.

(a) Core publications

Algorithmic fairness (Chapter 2 and Chapter 3).

[21] Adrian Arnaiz-Rodriguez and Nuria Oliver. "Towards Algorithmic Fairness by means of Instance-level Data Re-weighting based on Shapley Values". In: *ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR)*. Apr. 2024. URL: https://openreview.net/forum?id=ivf1QaxEGQ

[19] Adrian Arnaiz-Rodriguez, Georgina Curto Rex, and Nuria Oliver. "Structural Group Unfairness: Measurement and Mitigation by Means of the Effective Resistance". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 19. 1. Also presented at IC2S2 2024 and TrustLOG @ WWW 2024. June 2025, pp. 83–106. DOI: 10.1609/icwsm.v19i1.35805. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/35805

Human - AI complementarity (Chapter 4).

[25] **Adrian Arnaiz-Rodriguez**, Nina Corvelo, Suhas Thejaswi, Nuria Oliver, and Manuel Gomez Rodriguez. "Towards Human-AI Complementarity in Matching Tasks". In: *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases - The Third Workshop on Hybrid Human-Machine Learning and Decision Making*. Sept. 2025. URL: https://arxiv.org/abs/2508.13285

AI and regulatory alignment (Chapter 5).

[18] Adrian Arnaiz Rodriguez and Julio Losada Carreño. "La intersección de la IA fiable y el Derecho del Trabajo. Un estudio jurídico y técnico desde una taxonomía tripartita". Spanish. In: *Revista General de Derecho del Trabajo y de la Seguridad Social* 69 (2024). EN: The Intersection of Trustworthy AI and Labour Law. A Legal and Technical Study from a Tripartite Taxonomy, p. 2. ISSN: 1969-9626. URL: https://www.iustel.com/v2/revistas/detalle_revista.asp?id_noticia=427491

[17] Adrian Arnaiz Rodriguez and Julio Losada Carreño. "Estudio de la causalidad en la toma de decisiones algorítmicas: el impacto de la IA en el ámbito empresarial." Spanish. In: *Revista Internacional y Comparada de Relaciones Laborales y Derecho del Empleo* 12.3 (Dec. 2024). EN: Studying Causality in Algorithmic Decision Making: the Impact of IA in the Business Domain. ISSN: 2282-2313. URL: https://ejcls.adapt.it/index.php/rlde_adapt/issue/view/105

(b) Supporting work

The extensive survey and theoretical work underpinning Chapter 3 has been distilled into two papers and two tutorials. These works are included as appendices in the thesis for self-completeness in Appendices D, E, F.1 and F.2.

[23] Adrian Arnaiz-Rodriguez, Ahmed Begga, Francisco Escolano, and Nuria Oliver. "DiffWire: Inductive Graph Rewiring via the Lovász Bound". In: *Proceedings of the First Learning on Graphs Conference*. Vol. 198. Proceedings of Machine Learning Research. PMLR, Dec. 2022, 15:1–15:27. URL: https://proceedings.mlr.press/v198/arnaiz-rodri-guez22a.html

[20] **Adrian Arnaiz-Rodriguez** and Federico Errica. "Oversmoothing, "Oversquashing", Heterophily, Long-Range, and more: Demystifying Common Beliefs in Graph Machine Learning". In: *22nd International Workshop on Mining and Learning with Graphs (MLG 2025) at ECML-PKDD 2025*. July 2025. URL: https://arxiv.org/abs/2505.15547

[24] Adrian Arnaiz-Rodriguez, Ahmed Begga, Francisco Escolano, Nuria Oliver, and Edwin Hancock. "Graph Rewiring: From Theory to Applications in Fairness". In: *Proceedings of the Learning on Graphs Conference (LoG 2022)*. **Tutorial**. Virtual Event, Dec. 2022. URL: https://ellisalicante.org/tutorials/GraphRewiring

[22] Adrian Arnaiz-Rodriguez and Ameya Velingker. "Graph Learning: Principles, Challenges, and Open Directions". In: *41st International Conference on Machine Learning (ICML 2024)*. **Tutorial**. Vienna, Austria, July 2024. URL: https://icml.cc/virtual/2024/tutorial/35233

(c) Additional co-authored work

[345] Kajetan Schweighofer, Adrian Arnaiz-Rodriguez, Sepp Hochreiter, and Nuria Oliver. "The Disparate Benefits of Deep Ensembles". In: *Proceedings of the 42nd International Conference on Machine Learning*. Vol. 143. Proceedings of Machine Learning Research. June 2025. URL: https://arxiv.org/abs/2410.13831

Part I

Framework

Chapter 1

Introduction: Harms and Effective Trustworthy AI Implementation

Chapter summary and context

This thesis offers a comprehensive sociotechnical approach to Trustworthy AI (TAI) by integrating algorithmic, human-centric, and legal perspectives. We argue that TAI cannot be achieved solely through better algorithms but must also consider human interactions and legal constraints. Our findings contribute to the broader discourse on TAI, informing AI researchers and policymakers on how to develop systems that align technical principles with societal needs, ensuring that AI-driven decisions remain accountable, equitable, and aligned with ethical standards.

1.1 Artificial Intelligence in High-Risk Decision-Making

Artificial Intelligence (AI) is profoundly transforming the way decisions are made. AI algorithms make it possible to analyze large amounts of data quickly and efficiently, offering rapid and optimized solutions to complex problems. This capability has made AI a key tool for public and private organizations seeking to improve efficiency, reduce costs, and personalize services.

The ability of AI algorithms to find complex patterns and handle large amounts of data makes AI-based systems an ideal choice for addressing real-world problems, including in socially critical and high-risk scenarios where the impact of decisions on people's lives can be profound, such as healthcare, employment, justice, education, finance, policing, immigration, information exposure on social media and communication. These critical use cases are not only identified by academics [41] or relevant stakeholders [202, 292]. Still, recently adopted European Regulations — such as the *Artificial Intelligence Act* (EU AI Act) [162] and the *Digital Services Act* (DSA) [158] — also reflect these cases.

In high-risk scenarios, the design, implementation, deployment, evaluation, and auditing of AI systems must be carried out with caution to **minimize the *harms*** and potential negative consequences of their use, with the ultimate goal of achieving **Trustworthy AI** systems [52, 209, 307].

1.2 Trustworthy Artificial Intelligence

Concerns about the ethical risks associated with AI systems have led to the development of technical approaches, regulatory frameworks, and guidelines to ensure their ethical and responsible use. From the initial initiatives promoted by international organizations [149, 388] and technology companies [194, 292], to the adoption of national [52, 241] and supranational [209] strategies, the concept of **Trustworthy AI** (TAI) has been established as the standard to ensure that AI systems respect fundamental rights, avoid systemic risks and harms, and have a positive social impact.

In Europe, this effort was crystallized in 2019 with the publication of the Trustworthy AI Guidelines by the European Commission's High-Level Expert Group on AI (HLEG) [209]. This framework laid the groundwork for subsequent European regulations on the responsible use of AI, such as the EU AI Act and the EU DSA.

The Digital Services Act was adopted in 2022 in Europe and aims to enhance transparency, accountability, and the protection of fundamental rights in digital environments by governing online platforms and intermediary services.

The EU AI Act entered into force in 2024, the first transversal regulation of the use of AI globally. It establishes rules for developing, deploying, and using AI systems based on their associated risk levels.

Nowadays, the legal requirements reflected in the regulations are being translated into actionable practices. The Codes of Practice for General Purpose AI models [152] and the ISO regulations on Trustworthy AI [219, 220] help comply with the regulatory frameworks by providing technical criteria for their effective implementation [355].

Components and Ethical Principles of TAI. The High-Level Expert Group on Artificial Intelligence of the European Commission defined three fundamental components of Trustworthy AI systems [209]:

(i) AI systems need to be **lawful**, fully complying with all pertinent legislation and regulations.

(ii) AI systems need to be **ethical**, demanding adherence to established principles and values.

(iii) AI systems should be **robust**, from both social and technical perspectives, since they can cause unintentional harm.

These TAI components are defined to preserve fundamental rights through **ethical principles**, "which must be respected in order to ensure that AI systems are developed, deployed and used in a trustworthy manner" [209, p. 11]. These ethical principles urge developers to build, deploy, and use AI in ways that (i) respect human autonomy, (ii) prevent harm, (iii) promote fairness, and (iv) remain explicable.

Among the ethical principles underpinning Trustworthy AI, two play a fundamental role in this thesis: the **textit**promotion of fairness and the **prevention of harm**.

The principle of *prevention of harm* requires AI systems to be auditable and demands developing safeguards to avoid adverse effects, such as biased decisions or unsafe automation. Specifically, the principle of *promotion of fairness* demands that systems neither perpetuate

nor amplify existing inequities; discrimination based on protected attributes is prohibited by law and addressed by technical standards for bias assessment [220].[1]

Requirements and key guidance for an effective TAI implementation. To operationalize the Trustworthy AI principles throughout the AI life cycle, the HLEG proposed seven concrete **requirements** to avoid AI-related harms, namely:

i. Human oversight.

ii. Technical robustness and safety.

iii. Privacy and data governance.

iv. Transparency.

v. Diversity and non-discrimination.

vi. Societal and environmental well-being.

vii. Accountability.

To effectively develop Trustworthy AI and adhere to ethical principles and requirements, they suggested several **key guidelines**, including the need for both *technical* safeguards, such as metrics and audits, and *non-technical* measures, such as regulation or governance. Moreover, they also advocated developing, deploying, and using AI systems that adhere to ethical principles, paying particular attention to situations that involve vulnerable groups, especially those at risk of exclusion, and asymmetries of power or information. Finally, they also recommended involving stakeholders throughout the AI system's life cycle to clearly and proactively communicate information about its capabilities and technical and social limitations.

In conclusion, the three components, four principles, seven requirements, and key-guidelines for Trustworthy AI, as defined by the High-Level Expert Group on AI [209], converge on one operational imperative: the systematic identification, measurement and mitigation of induced *harms* across the intertwined spheres of AI system design, use, and governance.

1.3 Algorithmic Harms

As the Trustworthy AI ethical principles aim to minimize the potential harm caused by the deployment of AI systems, it is crucial to understand the nature of these harms. Building on prior work [84, 282, 330], we group societal harms due to algorithmic malfunction into six recurring patterns.

(i) *Performance harms* arise when predictive accuracy varies across groups, as seen in the canonical example of disparate performance in facial-recognition systems between different groups of people defined by sex and race [84].

[1] For example, Title VII of the US Civil Rights Act of 1964 [387], the German General Equal Treatment Act of 2006 (AGG) [228], or Article 17 of the Spanish labor law [72] prohibit discrimination based on socially relevant characteristics or *protected attributes*, such as race, ethnic origin, gender, sexual orientation, religion, disability, or age.

(ii) *Allocation harms* emerge when the algorithm misallocates scarce opportunities, for example, credit-scoring tools that underrate women or ethnic minorities, thereby restricting finance [181].

(iii) *Stereotype harm* occurs when systematically biased outputs reinforce cultural stereotypes, for example, language models that pair "nurse" with female pronouns or "engineer" with male ones [74].

(iv) *Denigration harms* affect individuals when misclassifications affront basic respect, such as the incident in which a photo-tagging service mislabeled a group of people as animals.[2]

(v) *Representation harms* entail systematic over- or under-representation, for instance, groups of online content creators being less frequently recommended by platforms, reducing their reach and revenue [103, 138, 347, 402].

(vi) *Procedural harms* are considered when decision pipelines violate societal accepted norms in decision-making, as in ride-share drivers whose accounts are algorithmically terminated with no avenue for appeal [145].[3]

Additionally, it is essential to recognize that harms are socially constructed and can encompass intangible damage to social, cultural, and political environments. This social nature of harms may lead to novel manifestations that emerge with each technological innovation or social dynamics. For instance, different harms might arise from the use of an algorithm by humans, such as overconfidence in the decisions of the AI models or the reluctance to use an AI system if the rationale for its decisions is not clearly explained [192].

More generally, a harm resulting from this interaction can be defined as situations where an AI-based decision-support tool causes a human expert to make a worse decision than they would have made independently. The diminished human performance directly undermines the technical robustness component and requirement of the HLEG TAI report [209], which is closely linked to the TAI principle of preventing harm. In this context, the report emphasizes that high accuracy and reliability are crucial when AI systems directly affect human lives.

In conclusion, societal harms arising from AI systems can stem from various sources, including algorithmic failures, unexpected outcomes, and problematic human use. The specific manifestations of these harms depend on the task and social context.

Understanding the diverse manifestations of algorithmic harm is essential for developing Trustworthy AI, but the real challenge lies in making Trustworthy AI a reality in practice [209, Ch. 2], *i.e.*, effectively preventing and mitigating these harms across the complex sociotechnical landscape in which AI systems are deployed.

However, the effective implementation of Trustworthy AI systems that can avoid harmful outcomes, particularly in high-risk contexts, remains challenging. This is due to technical limitations and the continuous alignment between ethical principles, system design choices, human behavior, and legal constraints.

[2] https://www.bbc.com/news/technology-33347866 (Accessed May 2025)
[3] https://www.bbc.com/news/business-54698858 (Accessed May 2025)

1.4 Challenges and Research Questions

Would any individual (or group of individuals) be worse off if an AI system were deployed?

Despite a surge of academic proposals and new legislation, a trustworthy implementation of AI systems remains the exception rather than the rule in real-world deployments [27, 173, 287]. The central obstacle is AI's *sociotechnical* nature: unintended harms can surface at multiple, intertwined levels. Following the HLEG key components, principles, and requirements of Trustworthy AI (Section 1.2), we treat an AI system as a *sociotechnical system* operating in **three spheres**:

i. **Technical design**, concerning algorithms and data;

ii. **Use**, regarding human-AI interaction and the general use of AI algorithms;

iii. **Governance**, comprising regulation, standards and accountability.

These three spheres must work together to implement Trustworthy AI practically. Figure 1 illustrates the three spheres. Each sphere poses internal challenges, while tensions also emerge when they overlap.

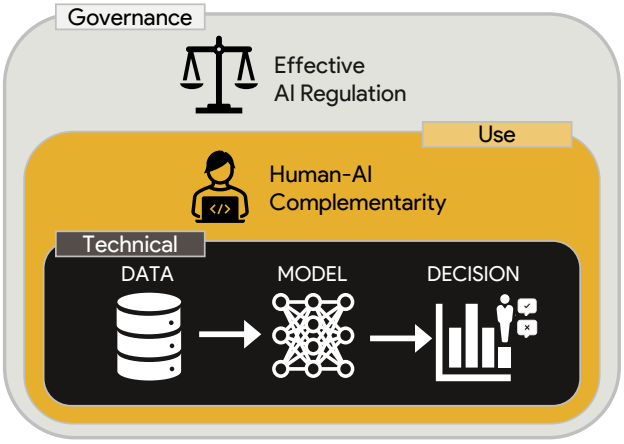


Figure 1. Mitigation of harms from three perspectives: **Technical**, **Use**, and **Governance**.

Within the *technical design* sphere, harms often manifest as discriminatory outcomes in the decisions by the AI systems, stemming from algorithmic biases [294]. In the *use sphere*, additional unintended consequences can emerge, such as a collaborative system performing less effectively than its human or algorithmic components operating in isolation [366], creating harms derived from the complete automation. In the *governance sphere*, new regulations may contradict existing ones or be challenging to align with them [3, 18].

Additionally, tension arises between the technical and legal spheres due to the difficulty of understanding and defining technical concepts in the legal domain. For example, the legal definition of an "AI system" is not trivial [168], and the concepts of "explainability" and "interpretability" have different meanings in the two spheres [312]. Finally, AI capabilities may outpace existing legal frameworks, and regulatory definitions may be misaligned with technical constructs [321].

Therefore, bridging the gap between principle and practice exposes coupled challenges that span the sociotechnical AI lifecycle: from a technically robust design and development, to the alignment of AI systems with existing and future regulations.

Thesis objective. This thesis makes several contributions in the context of Trustworthy AI with one common goal:

Mitigate AI-induced harm and
make Trustworthy AI actionable across the three sociotechnical spheres.

This thesis proposes mitigation strategies for some of the previously described harms in the three spheres. From an *algorithmic* perspective, we focus on mitigating the harm of bias and discrimination. In the machine learning literature, pre-processing, in-processing, and post-processing methods have been proposed to reduce algorithmic discrimination. From a sociotechnical perspective, these methods should also be technically robust, interpretable by humans, and aligned with existing regulations.

From a *human-AI collaboration* viewpoint, we contribute by proposing a novel method designed explicitly for human-AI complementarity that creates a robust pipeline to make decisions, while also being aligned with existing regulations.

Finally, from the *governance* perspective, we focus on the regulation. An *effective AI regulation* requires interdisciplinary collaboration between legal experts and technical specialists to develop a shared understanding of how the regulations align with the technical nature of AI systems, and to shed light on how specific AI regulations can coexist with existing ones.

The remainder of this chapter surveys mitigation strategies within each sphere and maps them to the thesis's four research questions and corresponding contributions.

1.4.1 Technical sphere: Algorithmic Fairness

Challenge

Develop technical metrics and algorithms that detect and mitigate different types of bias and systemic risks, while remaining robust, interpretable, and aligned with societal values.
RQ1: How can the fairness of machine-learning models in high-risk decision-making be improved by revealing how individual data points influence group disparities?
RQ2: How can we measure and mitigate structural discrimination in social graphs?

Algorithmic Bias

A key principle and requirement of Trustworthy AI is fairness and non-discrimination, which has been extensively studied in human decision-making [215]. For decades, numerous studies have empirically corroborated that human decisions are biased [229]. A *bias* is a systematic deviation in decision-making that unjustifiably favors or harms certain groups, whether due to cognitive, social, or structural factors.

Consequently, most countries in the Western world have implemented anti-discrimination and equal opportunity laws to prevent discriminatory decisions. For example, Title VII of the US Civil Rights Act of 1964 [387], the German General Equal Treatment Act of 2006 (AGG) [228], or Article 17 of the Spanish Workers' Statute [72] prohibit discrimination based on socially relevant characteristics or *protected attributes*, such as race, ethnic origin, gender, sexual orientation, religion, disability, or age. A protected attribute is, therefore, a personal characteristic that cannot be used as a basis for discriminatory decisions in high-stakes contexts, including employment, education, housing, or access to goods and services.

New challenges emerge, as decisions in high-risk scenarios are increasingly delegated to AI-based algorithms. Numerous studies demonstrate how AI models can perpetuate or amplify existing social biases, disproportionately impacting specific social groups, particularly the most vulnerable [41, 212]. This phenomenon, known as *algorithmic bias*, consists of the introduction or exacerbation of inequalities in automated decisions due to various causes, such as historical biases in the training data, an inappropriate choice of model or algorithm for the problem, or a biased interpretation of the results [27, 48, 288, 428].

As mentioned in Section 1.3, algorithmic bias can lead to different social harms depending on the use case. These harms include reinforcing structural inequalities and limiting certain groups' access to crucial opportunities and resources, such as education, employment, credit, and relevant information [238].

While the literature catalogues a wide range of fairness definitions based on various ethical and mathematical notions, as well as different use cases [93, 96, 215, 288, 393], our focus is on *statistical group fairness metrics* [94, 393], both for high-risk decision-making [41] and for bias in social networks [132]. The main reason for using statistical group disparity-based measures is that they are prevalent in compliance practices, they align with ethical notions of fairness [94], national anti-discrimination laws [72, 387], and underpin the EU AI Act's high-risk provisions [162]. Their adoption is further consolidated by international standards [219, 220].

Accordingly, the thesis aims to deepen and operationalize these metrics in two regulative contexts:

i. **High-risk decision-making (EU AI Act).** Canonical cases include credit-scoring tools that penalize disadvantaged groups [181], the COMPAS recidivism model that over-predicts risk for black defendants [16], and automated job applicant screener systems that downgraded female applicants.These examples would currently violate national non-discrimination laws or the non-discrimination mandate of the EU AI Act for high-risk systems. Although there are numerous fairness metrics for these decision-making processes, understanding how individual training data points influence group disparities is limited. This effect limits our understanding of the source of discrimination and actionable insights. We address this gap with a data-valuation approach that traces each training point's influence on group fairness (**RQ1**; see Chapter 2).

ii. **Online social platforms (EU DSA).** In online social networks, the use of recommender systems for content and connections can lead to effects such as popularity reinforcement, to echo chambers, political polarization, the marginalization of minorities, and unequal exposure [44, 103, 138, 163–165, 211, 402]. These risks fall within the umbrella of the EU DSA, which is responsible for monitoring and mitigating "systemic risks". These obligations are already in place for large online platforms.[4] However, no standard statistical group-level metrics exist for some systemic risks, such as disparate information access and exposure. We fill this void by proposing Structural Group Unfairness measures and an edge-augmentation strategy that reduces information access and exposure disparities (**RQ2**; see Chapter 3).

By focusing on these two legally defined arenas, the dissertation links technical advances directly to the regulatory environments where they are most urgently needed.

[4] https://digital-strategy.ec.europa.eu/en/policies/dsa-enforcement (Accesed May 2025)

Mitigation

From an algorithmic perspective, bias mitigation in machine learning models can be approached at three levels: pre-processing, modifying the input data before learning; in-processing, tuning the model's optimization function to include an algorithmic fairness term; and post-processing, correcting predictions after inference to mitigate biases in them. Depending on the system constraints, one of these approaches can be chosen or combined to improve decision fairness [206, 231, 417].

Before model training, pre-processing techniques seek to transform the data distribution to reduce underlying discrimination. This approach is applicable when the data is accessible and can be modified without restrictions. Common strategies include resampling to balance the distribution of protected groups [231], re-weighting instances in the loss function [21], and learning fair representations [417], which transforms the original features into a latent space where the protected attribute is minimized while preserving task-relevant information.

In-processing methods mitigate biases within the model optimization process by adjusting the cost function or imposing constraints that reduce disparities in predictions. This approach is beneficial when one has control over model training, but not over the input data. Among the most widely used techniques is adversarial optimization [254], in which an auxiliary model attempts to generate predictions that are indistinguishable concerning the protected attribute, thereby reducing the sensitive information that the model can exploit to bias decisions. Another common strategy is the inclusion of algorithmic fairness regularizers [232], which modify the loss function to minimize disparities in fairness metrics during training by penalizing differences in error or the distribution of predictions between different groups.

Post-processing techniques act on the output of a pre-trained model, applying adjustments to the predictions without modifying the data or optimizing the model. These methods are instrumental when the model must be treated as a black box without control over its training. The most common post-processing technique assigns different prediction thresholds across groups to ensure equivalent misclassification rates [114, 206].

Our contributions in Chapters 2 and 3 introduce metrics and methods for auditing and mitigating unfairness in high-risk decision-making processes and information access and exposure on social networks.

First, we improve the understanding of widely used decision-making fairness metrics in high-risk scenarios by proposing `FairShap`, a data-valuation approach that quantifies the influence of individual training points on group fairness scores (**RQ1**; Chapter 2). It also provides actionable insights into data re-weighting and pruning to mitigate biased decisions.

Secondly, we introduce new metrics and algorithms that measure and mitigate the harms of allocation and representation derived from disparate access and exposure to information on social networks (**RQ2**; Chapter 3).

1.4.2 Use sphere: Constructive Human-AI Collaboration

Challenge

As algorithms never operate in a vacuum, effective systems must exploit the contextual knowledge of human decision-makers and avoid the hazards of full automation.
RQ3: How should algorithms be designed to enable effective human-AI collaboration in resource allocation tasks?

Despite the perception that AI algorithms operate entirely autonomously, in many applications, especially high-risk ones, regulations enforce that final decisions must depend on human intervention [153]. In these cases, AI acts as a decision support system, assisting without replacing human control. As noted in Section 1.3, this interaction is not exempt from unintended harms arising during joint decision-making.

A growing body of work investigates how to design robust algorithms that explicitly account for human behaviour, collaboration, and oversight.[5] Proposed frameworks range from light human oversight to tightly coupled joint decision-making [261, 302], and address objectives ranging from explainability [278] to interface design that fosters user trust [92].

From the wide range of approaches in this sphere, we focus on *human-AI complementarity* systems [37, 364]. These systems are designed to achieve higher team performance than humans or AI working alone, thereby avoiding the risk of the combined system performing worse than either alone. This aligns with the HLEG principle of preventing harm by complying with the technical robustness requirement [209]. Specifically, we focus on complementarity by design [113, 123, 365, 367], where the system is *theoretically* designed to ensure that the combined decisions are always at least as accurate as those of either agent alone. For instance, to classify images, an algorithm can present a list of options to the user, whose length adapts to the user's skill [367].

In Chapter 4, we present an automated system for resource allocation tasks that inherently considers human-AI complementarity. This prevents unintended consequences and enables the use of additional human contextual knowledge (**RQ3**).

1.4.3 Governance sphere: Effective AI Regulation

Challenge

Legal incentives and safeguards will fail if the rules do not match the technical realities of AI. Governance must therefore co-evolve with algorithmic practice, and vice versa.

RQ4: What is the alignment between existing regulations and the requirements and technical realities of Trustworthy AI?

Just as algorithms do not operate autonomously and are subject to human oversight, their application occurs within a social and regulatory framework, which requires considering their impact on society. Preventing risks and harms associated with AI cannot be limited

[5] For a recent survey on human-centered AI, see [92]

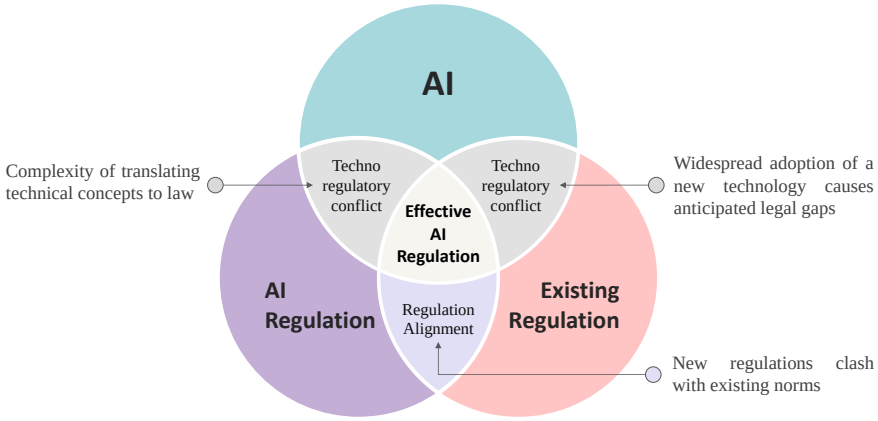


Figure 2. Effective governance requires aligning regulatory frameworks, avoiding techno-regulatory conflicts, and translating technical concepts into legal terms.

to technical solutions but must be addressed holistically, including regulation, audits, and accountability mechanisms.

Regulations and standards, such as the EU AI Act [162], seek to establish clear criteria to ensure that the deployment of AI systems does not conflict with or violate fundamental values and rights. In parallel, standardization and auditing processes make it possible to translate regulatory principles into concrete technical procedures, systematically assessing models' compliance with the regulations [219, 220, 355]. Finally, accountability is essential to ensure that companies and organizations take responsibility for the negative impacts of their models on society and take corrective measures where necessary [150].

However, although previous frameworks and regulations have addressed mitigation of harms through governance, the effective implementation of these regulatory solutions is not without challenges (see Figure 2). Inconsistencies may arise between technical methods and the current legal framework [17], contradictions between specific AI regulations and other existing legal regulations [18], or even divergences between technical AI concepts and their particular regulations [168, 312].

An example of this technical-legal conflict occurs in Spain, where the Spanish labour law [72] requires certain decisions, such as dismissals, to be causally justified. In contrast, others, such as hiring, must avoid discrimination. However, most machine learning-based AI systems operate through correlations, which makes causal identification difficult, and they are also not free from biases that can result in discriminatory decisions.

On Chapter 5, we explore how Trustworthy AI requirements overlap with the Spanish labor law regulation, and how technical challenges in the machine learning field overlap with this regulation (**RQ4**). Specifically, we identify Spanish and European regulations that apply to AI systems and analyze the connections between TAI requirements, the EU AI Act, and Spanish labor law. We then focus on a specific misalignment between the technical nature of AI systems and labor law, analyzing the implications of the technical correlation-causation dilemma in labor law. Finally, we propose technical and regulatory guidelines to align the technical and regulatory spheres to mitigate this problem.

1.5 Thesis Structure

In the following, we briefly describe each chapter in the thesis, where Chapter 2 and Chapter 3 focus on addressing **RQ1** and **RQ2** within the technical sphere; Chapter 4 tackles **RQ3** in the context of the use sphere, and Chapter 5 deals with **RQ4** in the governance sphere.

Chapter 2: FairShap – Data Valuation for Algorithmic Fairness

The core technical content of this chapter is based on the following scientific publication:

> [21] Adrian Arnaiz-Rodriguez and Nuria Oliver. "Towards Algorithmic Fairness by means of Instance-level Data Re-weighting based on Shapley Values". In: *ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR)*. Apr. 2024. URL: https://openreview.net/forum?id=ivf1QaxEGQ

Achieving algorithmic fairness is a critical challenge in machine learning, mainly as modern models rely on large-scale datasets that often reflect societal biases. These biases in the training data can lead to unfair decisions, disproportionately affecting underrepresented groups, and evaluation metrics measure them.

Statistical group fairness metrics have been widely studied in academic settings [41]. They are also closely aligned with group-based legal definitions of discrimination and have even been included in anti-discrimination standards for AI systems, such as ISO [220]. Therefore, they have the potential to be widely used to address discrimination in AI systems. However, there is a critical need for interpretable technical methods to help understand the fairness metrics and how the data used to train the model affects those fairness violations. The goal is to provide actionable insights to help avoid bias.

This chapter aims to promote algorithmic fairness through a **data-centric approach** by introducing a novel method that evaluates the contribution of individual training points to fairness metrics. We seek to improve fairness in machine learning models without compromising predictive performance.

We propose `FairShap`, a model-agnostic and interpretable instance-level data valuation method based on Shapley Values. `FairShap` quantifies each training point's influence on group fairness metrics and supports two fairness interventions: data re-weighting and data pruning. We evaluate `FairShap` on several benchmark tabular and computer vision datasets, under different training scenarios and models.

`FairShap`-guided interventions consistently improve model fairness while preserving accuracy across datasets and fairness metrics. Data pruning and re-weighting based on `FairShap` lead to significant fairness gains with minimal performance degradation. Visualizations in latent space and value histograms illustrate the method's interpretability. We also analyze the utility-fairness trade-off and characterize the computational cost of the proposed algorithm.

`FairShap` establishes data valuation as a powerful tool for algorithmic fairness. Its model-agnostic, interpretable design allows flexible integration into diverse training pipelines, enabling fairer outcomes without sacrificing utility. This work highlights the central role of data in mitigating bias and offers a practical, scalable solution for fairness-aware machine learning.

Chapter 3: Structural Group Unfairness – Fairness in Social Capital

The main technical contributions of this chapter can be found in the following scientific publication:

> [19] Adrian Arnaiz-Rodriguez, Georgina Curto Rex, and Nuria Oliver. "Structural Group Unfairness: Measurement and Mitigation by Means of the Effective Resistance". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 19. 1. Also presented at IC2S2 2024 and TrustLOG @ WWW 2024. June 2025, pp. 83–106. DOI: 10.1609/icwsm.v19i1.35805. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/35805

In socially-critical domains, algorithmic fairness must go beyond abstract principles and incorporate human context, tailored to specific scenarios. Social networks, which model the complex relationships between people, generate vast amounts of data for various purposes, including making decisions about individuals and providing targeted advertisement, recommendations, and information dissemination [132]. However, these platforms also pose systemic risks, such as polarization [211], reduced visibility for minority groups [163, 165], or online radicalization [164]. Some of these risks have been formally recognized in regulatory frameworks like the EU DSA [158].

Furthermore, social networks play a central role in distributing social capital, defined as the relationships, trust, and norms of reciprocity that support cooperation and collective action within a community. Individuals with advantageous positions in these networks benefit from faster access to diverse information and hold greater influence in the dissemination of information. While several methods have been proposed in the literature to measure social capital at the individual level, there remains a notable lack of methods to quantify social capital at the group level, particularly for groups defined by protected attributes where equity and fairness considerations are critical.

To fill this gap, we propose measuring the social capital of a group of nodes through *effective resistance* and emphasizing the importance of considering the entire network topology. Grounded in spectral graph theory, we introduce three effective resistance-based measures of group social capital, namely *group isolation*, *group diameter*, and *group control*, where the groups are defined according to the value of a protected attribute. We denote the social capital disparity among different groups in a network as *structural group unfairness*, and propose to mitigate it using a budgeted edge augmentation heuristic that systematically increases the social capital of the most disadvantaged group.

In experiments on real-world networks, we uncover significant levels of structural group unfairness when using gender as the protected attribute, with females being the most disadvantaged group in comparison to males. We also illustrate how our proposed edge augmentation approach can effectively mitigate the structural group unfairness and increase the social capital of all groups in the network.

In addition, the extensive survey and theoretical work underpinning Chapter 3 has been distilled into two papers and two tutorials. The first work presents DiffWire, an inductive graph rewiring [23] method for Graph Neural Networks (GNNs) to improve graph and node classification, especially in long-range tasks. The second work provides a systematic and critical review of some of the open challenges in GNNs that arise from parametrized

information diffusion when using the message passing mechanism. It has been released as a preprint [20].

In addition, the two tutorials were presented at LoG 2022 [24] and ICML 2024 [22]. Among these, Arnaiz-Rodriguez et al. [24] is the closest work to Chapter 3, where we presented literature on graph rewiring and its application to algorithmic fairness. These publications and tutorials on graph rewiring and GNN pathologies provide the theoretical ground for understanding the structural notions of algorithmic fairness [19]. These works will be incorporated into the thesis for self-completeness in Appendices D, E, F.1 and F.2.

Publications

[23] Adrian Arnaiz-Rodriguez, Ahmed Begga, Francisco Escolano, and Nuria Oliver. "DiffWire: Inductive Graph Rewiring via the Lovász Bound". In: *Proceedings of the First Learning on Graphs Conference*. Vol. 198. Proceedings of Machine Learning Research. PMLR, Dec. 2022, 15:1–15:27. URL: https://proceedings.mlr.press/v198/arnaiz-rodri-guez22a.html

[20] **Adrian Arnaiz-Rodriguez** and Federico Errica. "Oversmoothing, "Oversquashing", Heterophily, Long-Range, and more: Demystifying Common Beliefs in Graph Machine Learning". In: *22nd International Workshop on Mining and Learning with Graphs (MLG 2025) at ECML-PKDD 2025*. July 2025. URL: https://arxiv.org/abs/2505.15547

Tutorials

[24] Adrian Arnaiz-Rodriguez, Ahmed Begga, Francisco Escolano, Nuria Oliver, and Edwin Hancock. "Graph Rewiring: From Theory to Applications in Fairness". In: *Proceedings of the Learning on Graphs Conference (LoG 2022)*. **Tutorial**. Virtual Event, Dec. 2022. URL: https://ellisalicante.org/tutorials/GraphRewiring

[22] Adrian Arnaiz-Rodriguez and Ameya Velingker. "Graph Learning: Principles, Challenges, and Open Directions". In: *41st International Conference on Machine Learning (ICML 2024)*. **Tutorial**. Vienna, Austria, July 2024. URL: https://icml.cc/virtual/2024/tutorial/35233

Chapter 4: Human-AI Matching – Human AI Complementarity

The contributions of this chapter are presented in the following paper:

> [25] **Adrian Arnaiz-Rodriguez**, Nina Corvelo, Suhas Thejaswi, Nuria Oliver, and Manuel Gomez Rodriguez. "Towards Human-AI Complementarity in Matching Tasks". In: *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases - The Third Workshop on Hybrid Human-Machine Learning and Decision Making*. Sept. 2025. URL: https://arxiv.org/abs/2508.13285

Fully automating decision-making in high-risk scenarios may lead to unintended harm. For this reason, existing regulations, such as Article 22 of the GDPR [153], mandate human involvement in final decisions. Furthermore, human decision-makers often bring contextual knowledge that algorithms lack, and purely automated systems can overlook domain-specific factors. To address this, we focus on designing collaborative systems that leverage the complementary strengths of both humans and AI. However, these hybrid systems can also introduce new risks if not carefully designed.

In this chapter, we develop application-specific matching algorithms that explicitly consider the respective strengths of human and machine decision-makers, aiming to ensure that the combined system performs better than either component alone. Usually, data-driven algorithmic matching systems promise to improve decision-making in high-stakes domains, such as healthcare and social service provision. Yet, current systems do not reliably achieve human-AI complementarity: decisions made collaboratively are not guaranteed to outperform those produced by the human or the algorithm individually. Our work in this chapter seeks to fill this gap.

To this end, we present `CoMatch`, a data-driven algorithmic matching system that, for each matching task, does not make all matching decisions as existing systems do. Instead, it makes the matching decisions most certain and defers the remaining decisions to the human decision maker. Along the way, `CoMatch` optimizes how many decisions it makes and how many it defers to the human decision maker to maximize performance provably. We conduct a large-scale user study ($n = 800$) to validate our system. The results show that the matchings created by human participants using `CoMatch` are superior to those created either by human participants or by algorithmic matching alone.

`CoMatch` contributes to the promising field of human-AI collaboration, enabling more dynamic and adaptive decision-making, ultimately enhancing accuracy and user agency.

Chapter 5: Governance of AI in labor law - Intersection with Regulation

This chapter is partially based on the following two publications:

[18] Adrian Arnaiz Rodriguez and Julio Losada Carreño. "La intersección de la IA fiable y el Derecho del Trabajo. Un estudio jurídico y técnico desde una taxonomía tripartita". Spanish. In: *Revista General de Derecho del Trabajo y de la Seguridad Social* 69 (2024). EN: The Intersection of Trustworthy AI and Labour Law. A Legal and Technical Study from a Tripartite Taxonomy, p. 2. ISSN: 1969-9626. URL: https://www.iustel.com/v2/revistas/detalle_revista.asp?id_noticia=427491

[17] Adrian Arnaiz Rodriguez and Julio Losada Carreño. "Estudio de la causalidad en la toma de decisiones algorítmicas: el impacto de la IA en el ámbito empresarial." Spanish. In: *Revista Internacional y Comparada de Relaciones Laborales y Derecho del Empleo* 12.3 (Dec. 2024). EN: Studying Causality in Algorithmic Decision Making: the Impact of IA in the Business Domain. ISSN: 2282-2313. URL: https://ejcls.adapt.it/index.php/rlde_adapt/issue/view/105

Deploying Trustworthy AI in high-risk scenarios demands not only algorithmic robustness and fairness-aware design but also alignment with existing legal obligations and sector-specific regulatory requirements. In this chapter, we examine the interaction between the implementation of Trustworthy AI systems and Spanish labour law [72], focusing on a sensitive area where legal restrictions on discrimination, transparency, and automation are particularly strict. The labor decisions are included as a *high-risk* use case in the EU AI Act.

First, we explain the labor decisions our analysis focuses on and the applicable regulations. Next, we explain the legal definition of an AI system to understand the consensus reached by experts and regulators regarding this term.

We continue by analyzing the intersection between the requirements for an AI system to be considered trustworthy and the Spanish labor law, highlighting their areas of compatibility (*e.g.*, shared emphasis on transparency, accountability, and non-discrimination) and where labor law should focus specifically. This analysis highlights the need to move from abstract principles to operationalize **sector-specific** criteria that address both technical feasibility and legal legitimacy.

Next, we focus on a concrete case where tensions manifest: the legal *principle of sufficient reason* in employment decisions, and the use of correlation-based AI models to support or automate such decisions [17]. This second study examines how AI systems that rely on historical data and statistical correlations may fail to meet legal standards requiring explicit, individualized justifications (*e.g.*, for dismissal, contract changes, or hiring). We identify critical mismatches between how AI and law conceive "justified decision-making", and propose guidelines for technical and legal solutions, including human-AI collaboration, explainable AI, causal inference, AI literacy, and regulatory safeguards.

In conclusion, we emphasize the legal limitations of algorithmic decision-making in employment, demonstrating the importance of avoiding full automation, particularly when human rights and labor protections are at risk. This chapter ultimately emphasizes that interdisciplinary approaches are crucial for effectively deploying fair and legally compliant AI systems in the real world. Our findings inform the technical community by revealing overlooked legal constraints and the regulatory community by showcasing how technical design decisions can

better reflect labor law obligations and the requirements of Trustworthy AI.[6]

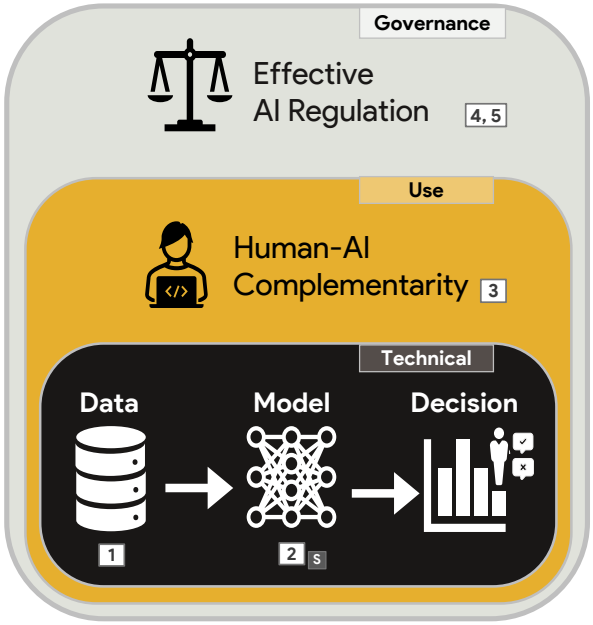
1.6 Conclusion

This dissertation demonstrates that achieving Trustworthy AI is inseparable from addressing three interconnected spheres: technical design, human use, and governance. Through the introduction of data-centric and graph-centric fairness methods, the proposal of a human-AI complementarity system for resource allocation tasks, and the analysis of the alignment between Trustworthy AI and existing labor law, the thesis translates the abstract ethical principles of fairness and harm prevention into concrete tools and guidelines. Together, these contributions demonstrate that trust can only be sustained when computational techniques and legal requirements for AI systems are developed in harmony. The result is a practical, interdisciplinary roadmap for AI systems that are both effective and aligned with societal values.

Figure 3 situates the works presented in this thesis in the proposed holistic Trustworthy AI framework, distinguishing the three proposed spheres.

Reading guide. This introduction proposes a sociotechnical framework for implementing Trustworthy AI: linking ethical principles, technical advancements, and regulatory mandates. As the following chapters cover three sociotechnical spheres and draw on diverse technical domains, each chapter includes its concise background, notation, and related work. This self-contained structure lets readers engage with any chapter independently, while the present introduction provides the unifying lens through which all contributions fit together.

[6]These analyses are part of a policy advisory project that will help regulators develop robust legislation aligned with the technical capabilities of AI systems. They are part of the Collaboration Agreement C039/23OT between Red.es and UCLM, which is focused on implementing the Digital Rights Charter in the context of digital rights in the workplace: https://www.derechosdigitales.gob.es/



[1] **DMLR @ ICLR24.** Towards Algorithmic Fairness by means of Instance-level Data Re-weighting based on Shapley Values.
***A. Arnaiz-Rodriguez**, N. Oliver.*

[2] **ICWSM 2025.** Structural Group Unfairness: Measurement and Mitigation by means of the Effective Resistance.
***A. Arnaiz-Rodriguez**, Curto. G, N. Oliver.*

[3] **HLDM @ ECML25.** Towards Human-AI Complementarity in Matching Tasks.
***A. Arnaiz-Rodriguez,** N. Corvelo, S. Thejaswi, N. Oliver, M. Gomez-Rodriguez.* Under Review.

[4] **RGDTSS.** The Intersection of Trustworthy AI and Labor Law. A Legal and Technical Study from a Tripartite Taxonomy.
***A. Arnaiz-Rodriguez**, J. Losada*

[5] **RRLDE.** Studying Causality in Algorithmic Decision Making: the Impact of IA in the Business Domain
***A. Arnaiz-Rodriguez,** J. Losada*

Supporting work

[S] **LoG 2022.** DiffWire: Inductive Graph Rewiring via the Lovász Bound.
***A. Arnaiz-Rodriguez**, A. Begga, F. Escolano, N. Oliver*

[S] **MLG @ ECML 25.** Oversmoothing, "Oversquashing", Heterophily, Long-Range, and more: Demystifying Common Beliefs in Graph ML.
***A. Arnaiz-Rodriguez** and F. Errica*

[S] **ICML 2024**, ***Tutorial.*** Graph Learning: Principles, Challenges, and Open Directions.
***A. Arnaiz-Rodriguez** and A. Velingker*

[S] **LoG 2022**, ***Tutorial***. Graph Rewiring: from theory to Fairness.
***A. Arnaiz-Rodriguez**, A. Begga, F. Escolano, E. Hancock, N. Oliver*

Figure 3. Works related to this thesis situated in the three spheres of the sociotechnical system.

Part II

A Holistic Approach to Trustworthy AI

Chapter 2

FairShap: Instance-Level Data Valuation for Algorithmic Fairness

Chapter summary and context

In this chapter, situated in the **technical** sphere, we introduce `FairShap`. This novel data valuation method quantifies the influence of each training example on group-level fairness metrics in high-risk decision-making systems. By adapting Shapley values to fairness auditing, `FairShap` enables both instance-level diagnosis and fairness-aware data interventions (*e.g.*, pruning, re-weighting). This method offers interpretable, actionable tools for practitioners seeking to align algorithmic systems with fairness requirements under the EU AI Act.
This chapter is partially based on the following publication:

> [21] Adrian Arnaiz-Rodriguez and Nuria Oliver. "Towards Algorithmic Fairness by means of Instance-level Data Re-weighting based on Shapley Values". In: *ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR)*. Apr. 2024. URL: https://openreview.net/forum?id=ivf1QaxEGQ

2.1 Introduction

Machine learning (ML) models are increasingly used to support human decision-making in diverse use cases, including in high-stakes domains, such as healthcare, education, finance, policing, or immigration. In these scenarios, algorithmic design, implementation, deployment, evaluation, and auditing should be performed cautiously to minimize the potential negative consequences of their use, and to develop fair, transparent, accountable, privacy-preserving, reproducible, and reliable systems [41, 307, 354]. To achieve algorithmic fairness, various fairness metrics that mathematically model different definitions of equality have been proposed in the literature [94]. Group fairness focuses on ensuring that other demographic groups are treated fairly by an algorithm [206, 415], whereas individual fairness aims to give a similar treatment to similar individuals [137]. In the past decade, numerous machine learning methods have been proposed to achieve algorithmic fairness [288].

Algorithmic fairness is addressed in the three stages of the ML pipeline: first, by modifying the input data (*pre-processing*) via *e.g.*, re-sampling, data cleaning, re-weighting or learning

fair representations [231, 417]; second, by including a fairness metric in the optimization function of the learning process (*in-processing*) [232, 418]; and third, by adjusting the model's decision threshold to ensure fair decisions across groups (*post-processing*) [206].

From a practical perspective, pre-processing methods for algorithmic fairness tend to be easier to understand for a diverse set of stakeholders, including legislators [166, 201]. Furthermore, to mitigate potential biases in the data, there is increased societal interest in using demographically-representative data to train ML models [186, 202, 281]. However, most available datasets used to train ML models in real-world scenarios are not demographically representative and could be biased. Moreover, datasets that are carefully created to be fair lack the required size and variety to train large-scale deep learning models. Thus, algorithmic fairness methods that focus on modeling and correcting bias in the data emerge as valuable approaches [107], especially methods that identify the value of each data point not only from the perspective of the algorithm's performance, but also from a fairness perspective [166], and methods that can leverage bias-aware curated datasets to improve fairness when learning from large-scale yet probably biased datasets. Bias-aware curated datasets are datasets where the datapoints are balanced regarding the labels (Y) and sensitive groups (A) to prevent statistical bias in $P(Y|A)$ and/or $P(Y|X,A)$, where Y are the labels or target variables, X the input variables and A the protected attributes.

Data valuation approaches are particularly well suited for this purpose. The data valuation methods proposed to date [187, 226] measure the contribution of each data point to the utility of the model —usually defined as accuracy— and use this information as a pre-processing step to improve the performance of the model. However, there is a scarcity of data valuation methods proposed for algorithmic fairness.[7]

In this chapter, we fill this gap by proposing `FairShap`, an instance-level, data re-weighting method for fair algorithmic decision-making which is model-agnostic and interpretable through data valuation. `FairShap` leverages the concept of Shapley Values [349] to measure the contribution of *each* data point to a pre-defined group *fairness* metric.

`FairShap` has several advantages:

i. it is easily interpretable, as it assigns a numeric value to each data point in the training set corresponding to its contribution to the fairness metric;

ii. it enables detecting which data points are the most important to improve fairness while preserving accuracy;

iii. it makes it possible to leverage small but bias-aware curated datasets to learn fair models from large-scale yet biased datasets;

iv. it is model agnostic and threshold independent.

Figure 4 illustrates how `FairShap` can be applied for data re-weighting and informing data pruning policies by means of dataset pruning. First, individual weights Φ_i are computed for each training point x_i by leveraging a reference dataset $\mathcal{T}$, which is typically the validation split of the original dataset $\mathcal{D}$ or, when available, an auxiliary bias-aware curated dataset. These weights can be used to: (1) re-weight the training data to train a fairer machine

[7] We clarify in Section 2.4.1 the distinction between the concept of *fairness* in data valuation (fair payouts proportional to the contribution) and algorithmic fairness (fair decisions across social groups).

learning model, or (2) inform data pruning policies by identifying the data points that contribute the least to performance and fairness.

Contributions: The main contributions of this work are four-fold:

i. We propose `FairShap`, a novel instance-level data valuation method for algorithmic fairness, which is model-agnostic and interpretable.

ii. We show how `FairShap` can guide two key fairness interventions – data re-weighting and data pruning – by identifying the data points that have the worst impact on fairness.

iii. We evaluate data valuation using FairShap across multiple state-of-the-art datasets, with different training scenarios, models, and tasks, showing that it consistently improves fairness while preserving predictive performance.

iv. We provide qualitative insights by visualizing `FairShap`'s effects through value distributions, decision boundaries, and latent representations.

2.2 Related Work

Group Algorithmic Fairness. Group bias in algorithmic decision-making is based on the conditional independence between the joint probability distributions of the sensitive attribute (A), the target variable or label (Y), and the predicted outcome ($\hat{Y}$). Barocas, Hardt, and Narayanan [41] define three concepts used to evaluate algorithmic fairness: *independence* ($\hat{Y} \perp A$), *separation* ($\hat{Y} \perp A|Y$), and *sufficiency* ($Y \perp A|\hat{Y}$). The underlying idea is that a *fair* classifier should have the same error classification rates for different protected groups. Three popular metrics to assess group algorithmic fairness are —from weaker to stronger notions of fairness— *demographic parity* (DP), *i.e.*, equal acceptance rate [137, 415]; *equal opportunity* (EOp), *i.e.*, equal true positive rate, TPR, for all groups [106, 206]; and *equalized odds* (EOdds), *i.e.*, equal TPR and false positive rate, FPR, for all groups [206, 415]. Numerous algorithms have been proposed to maximize these metrics while maintaining accuracy [288]. This chapter focuses on improving the two strongest of these group-based fairness metrics: EOp and EOdds.

Data Re-weighting for Algorithmic Fairness. Data *re-weighting* for algorithmic fairness is a pre-processing technique that assigns weights to the training data to optimize a certain fairness measure. Compared to other pre-processing approaches, data re-weighting is easily interpretable [42]. There are two broad approaches to perform data re-weighting: group and instance-level re-weighting.

In *group re-weighting*, the same weight is assigned to all data points belonging to the same group, which in the context of algorithmic fairness are defined according to their values of a sensitive attribute A (*e.g.*, gender, race, age, etc). Kamiran and Calders [231] re-weight the groups defined by A and Y based on statistics of the under-represented label(s) and the disadvantaged group(s) in a model-agnostic manner. Krasanakis et al. [250] assumes an underlying set of labels corresponding to an unbiased distribution and uses an inference model based on label error perturbation to define weights that yield better fairness performance. Jiang and Nachum [225] propose adjusting the loss function values in the sensitive groups

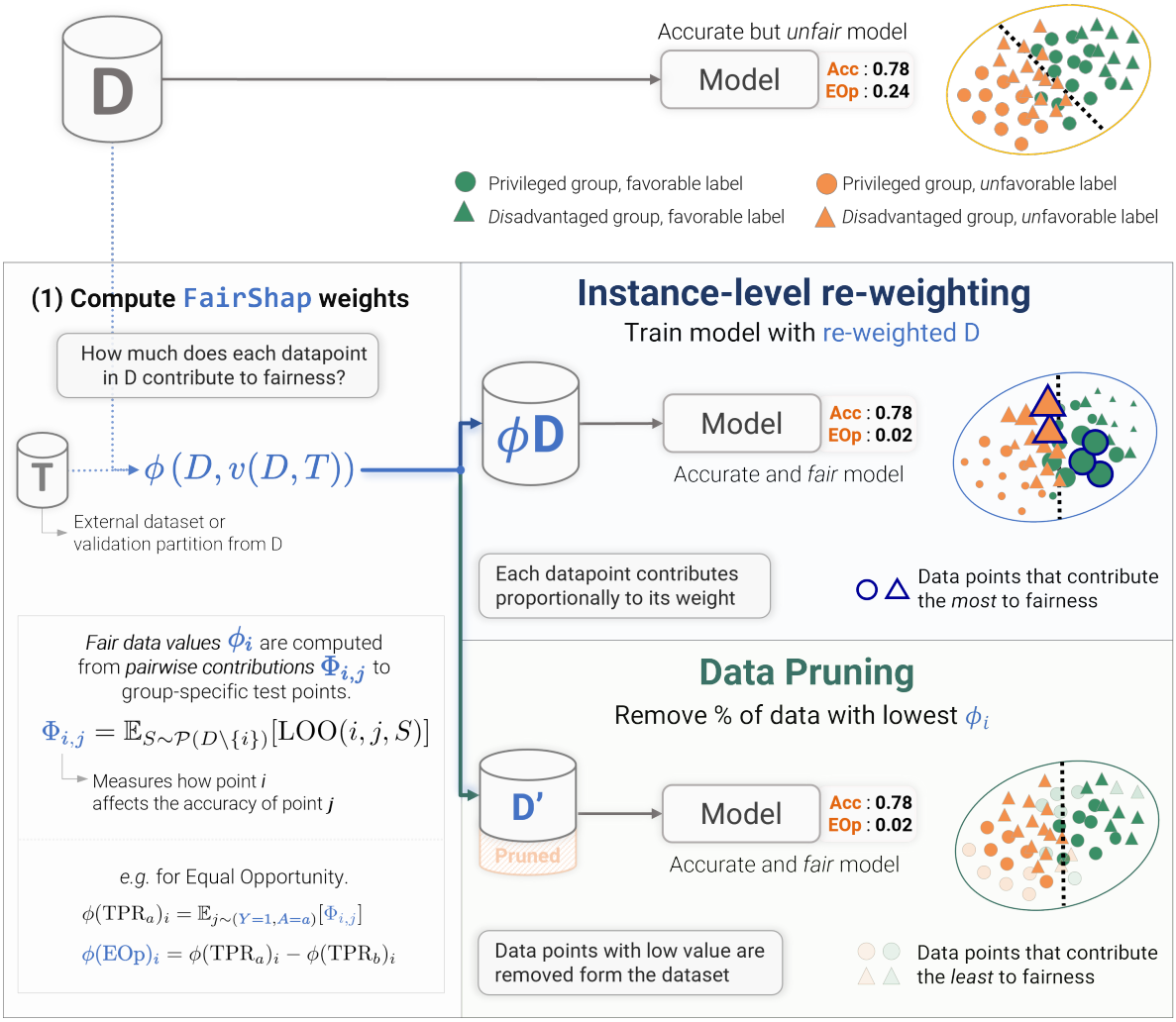


Figure 4. *Top*: Standard model training on the original dataset $\mathcal{D}$ which can lead to unfair decisions. *Bottom:* **FairShap's workflow.** *Left*: Instance-wise data values (ϕ_i) are computed using a reference dataset $\mathcal{T}$, which is typically the validation split of $\mathcal{D}$, but can also be an external, bias-aware, curated dataset. The value ϕ_i denotes how much point i contributes to a given fairness metric. FairShap computes ϕ_i using *pairwise contributions* $\Phi_{i,j}$, which quantify the impact of removing point i on the correct classification of point j. *Top right*: Illustration of the impact of instance-level re-weighting using FairShap on the data points and the resulting decision boundary. The re-weighting shifts the effective data distribution, contributing to a fairer learned model while maintaining accuracy compared to the original model. *Bottom right*: Illustration of data pruning employing FairShap: instances with the lowest contribution to fairness are pruned from the dataset before training, which contributes to a fairer model.

to iteratively learn weights that address the labeling bias and thus improve the fairness of the models, which is conceptually similar to methods based on using adversarial approaches to identify samples where the model is the most likely to make mistakes [254]. Chai and Wang [98] find the weights by solving an optimization problem that entails several rounds of model training. Finally, re-weighing has also been combined with a regularization term to adjust the weights in an iterative optimization process based on distributionally robust optimization [227].

However, note that many of these works, such as [98, 225, 227, 250, 254], propose re-weighting methods that adjust the weights repeatedly through an ongoing learning process, thus resembling in-processing rather than pre-processing approaches [97] as the computed weights depend on the model itself. This iterative process adds uncertainty to the weight computation [13]. It requires retraining the model in each iteration, which could be computationally very costly or even intractable for large datasets and/or complex models. Conversely, data-valuation methods, such as `FairShap`, are based on the concept that the value of the data should be orthogonal to the choice of the learning algorithm and hence data-valuation approaches should be purely data-driven and hence model-agnostic [352].

In contrast to group re-weighting, *instance-level re-weighting* assigns individual weights to each data point by considering the protected attributes and the sample misclassification probability. Most previous work has proposed using Influence Functions (IFs) for instance-level re-weighting. IFs estimate the changes in model performance when specific points are removed from the training set by computing the gradients or Hessian of the model [248, 314, 323, 371]. In the context of fairness, IFs have been used to estimate the impact of data points on fairness metrics. Black and Fredrikson [58] propose a leave-one-out (LOO) method to estimate such an influence. In Wang, Wang, and Liu [398], the data weights are estimated using a neural tangent kernel by leveraging a kernelized combination of training examples. Finally, Li and Liu [263] propose an algorithm that uses the Hessian of the matrix of the loss function to estimate the effect of changing the weights to identify those that most improve the model's fairness.

While promising, IFs are not exempt from limitations, including a certain level of fragility, their dependency on the model –and thus making them in-processing rather than pre-processing methods, their need for strongly convex and twice-differentiable models [45] and their limited interpretability, which is increasingly a requirement by legal stakeholders [166, 201]. Also, IFs approximate the leave-one-out score only for strongly convex loss functions, which limits the analysis by overlooking correlations between data points [205, 248, 252]. In contrast, Shapley Values have demonstrated greater stability than LOO and superior performance in data selection tasks, and in both stochastic and deterministic learning scenarios [397]. Finally, IFs do not satisfy properties that have been attributed to data valuation methods, such as the awareness of data preference, which are essential to making the methods more precise, practical, and interpretable [187, 407].

Data Valuation. Cooperative game theory concepts, such as the *Shapley Value* (SV) [349] or the *Core value* [189], measure how much a player contributes to the total utility of a team in a given coalition-based game. They have been used in the context of data valuation [187, 205], showing promise in several domains and tasks, including federated learning [395], data minimization [81], data acquisition policies, data selection for transfer learning, active learning, data sharing, exploratory data analysis and mislabeled example detection [205, 343]. In the ML literature, SVs have been also proposed to tackle a variety of tasks, such as transfer learning and counterfactual generation [11, 167], feature explainability [278, 283] and feature selection [316] by measuring the contribution of each *feature* to the individual prediction, not to be confused with computing the contribution of each *data* point, as we do in this chapter.

Our work takes inspiration from Ghorbani and Zou [187] and Jia et al. [224], who proposed using SVs to determine the contribution of each data point to the model's accuracy. In both cases, the SVs modify the training process or design data selection policies. The goal is to maximize the model's accuracy in the test set. However, we are unaware of any publication

that has leveraged SVs for data re-weighting to increase fairness while maintaining accuracy. In this chapter, we fill this gap and propose `FairShap`, an interpretable, instance-level data re-weighting method for algorithmic fairness based on SVs for data valuation. In addition to data re-weighting, `FairShap` is used to inform data pruning policies, as illustrated in Section 2.6.3.

2.3 Desiderata

In this section, we compare `FairShap` and related methods regarding their desirable properties.

Method	**D1** Data Val.	**D2** Interpretable	**D3** Pre-processing	**D4** Model agnostic	**D5** Data RW	**D6** Instance-level
`FairShap`	✓	✓	✓	✓	✓	✓
Group-RW	✗	✓	✓	✓	✓	✗
Influence Functions	✓	✗	✗	✗	✓	✗
Inpro-RW (LabelBias)	✗	✓	✗	✗	✓	✗
Massaging (OptPre)	✗	✓	✓	✓	✗	✗
Post-pro	✗	✓	✗	✓	✗	✗

Table 1. Comparative properties of related algorithmic fairness methods

The closest methods to `FairShap` in the literature are *Influence Functions* [263, 398], *In-processing re-weighting* (*e.g.*, LabelBias) [98, 225, 250, 254], *Group re-weighting* [231] and *Massaging* (*e.g.*, OptPre) [90, 166]. We define six desirable properties for fairness-aware methods (see Table 1):

- **D1 - Data Valuation**: The method quantifies the contribution of individual data points toward a target function, using leave-one-out, pairwise, or power-set-based strategies [205].
- **D2 - Interpretable**: The method's output should be understandable by both technical and non-technical stakeholders, supporting use cases such as data selection, mislabeled example detection, or federated learning.
- **D3 - Pre-processing**: The method should produce data insights applicable to a wide range of ML models during training.
- **D4 - Model Agnostic**: Data insights or transformations should not require repeated model training, to promote generality, efficiency, and robustness [352].
- **D5 - Data Re-weighting**: The method should yield data weights that enable rebalancing and fairness-aware training.
- **D6 - Instance-Level**: It should assign distinct outputs (*e.g.*, weights or values) to individual data points.

2.4 Background

2.4.1 Data Valuation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ be the dataset used to train a machine learning model M. The Shapley Value (SV) of a data point (x_i, y_i) –or i for short– that belongs to the dataset $\mathcal{D}$ is a data valuation function, $\phi_i(\mathcal{D}, v) \in \mathbb{R}$ –or $\phi_i(v)$ for short, that estimates the contribution of each data point i to the performance or valuation function $v(M, \mathcal{D}, \mathcal{T})$ –or $v(\mathcal{D})$ for short– of model M trained with dataset $\mathcal{D}$ and tested on *reference* dataset $\mathcal{T}$, which is either a subset of $\mathcal{D}$ (*e.g.*, the validation set) or possibly an external dataset. The SV is given by Equation (1). Note how its computation considers all subsets S in the powerset of $\mathcal{D}$, $\mathcal{P}(\mathcal{D})$.

$$\phi_i(\mathcal{D}, v) := \frac{1}{|\mathcal{D}|} \sum_{S \in \mathcal{P}(\mathcal{D} \setminus \{i\})} \frac{v(S \cup \{i\}) - v(S)}{\binom{|\mathcal{D}|-1}{|S|}} \tag{1}$$

The valuation function $v(\mathcal{D})$ is typically defined as the accuracy of M trained with dataset $\mathcal{D}$ and tested with $\mathcal{T}$. In this case, the SV, $\phi_i(\text{Acc})$, measures how much each data point $i \in \mathcal{D}$ contributes to the accuracy of M. The values, $\phi_i(\text{Acc})$, might be used for several purposes, including domain adaptation data re-weighting [187]. For simplicity, we denote $\phi(\mathcal{D}, v)$ as $\phi(v)$ when $\mathcal{D}$ is implicitly assumed.

Axiomatic properties of the Shapley Values. The SVs satisfy the following four axiomatic properties:

1. *Efficiency*: $v(\mathcal{D}) = \sum_{i \in \mathcal{D}} \phi_i(v)$, *i.e.*, the value of the training dataset $\mathcal{D}$ is equal to the sum of the Shapley Values of each of the data points in $\mathcal{D}$.

2. *Symmetry*: $\forall S \subseteq \mathcal{D} : v(S \cup i) = v(S \cup j) \rightarrow \phi_i = \phi_j$, *i.e.*, if two data points add the same value to the dataset, their Shapley Values must be equal.

3. *Additivity*: $\phi_i(\mathcal{D}, v_1 + v_2) = \phi_i(\mathcal{D}, v_1) + \phi_i(\mathcal{D}, v_2)$, $\phi_i(D, v_1 + v_2) = \phi_i(D, v_1) + \phi_i(D, v_2)$, *i.e.*, if the valuation function is split into additive two parts, we can also compute the Shapley values in 2 additive parts.

4. *Null Element*: $\forall S \subseteq \mathcal{D} : v(S \cup i) = v(S) \rightarrow \phi_i = 0$, *i.e.*, if a data point does not add any value to the dataset then its Shapley value is 0.

Clarification of the concept of fairness. We also clarify that the notion of fairness in SV theory differs from that in algorithmic fairness. The former ensures fair payouts based on a point's contribution to model performance, grounded in properties like efficiency and symmetry. In contrast, algorithmic fairness concerns equitable treatment across demographic groups. `FairShap` bridges these concepts by using Shapley Values to drive fairness-aware data interventions while preserving the axiomatic guarantees of the original framework.

2.4.2 Group Algorithmic Fairness

The group fairness criteria addressed in this work are based on the statistical relationships between the predicted outcomes $\hat{Y}$, the observed outcomes Y, and the protected group

attribute A. Following standard practice and without loss of generality, we consider binary classification and binary protected attributes with $\hat{Y} = Y = 1$ as the positive outcome and $A = a$ as the advantaged group. We focus on two widely used fairness metrics from the literature [288]: Equalized Odds (EOdds) and Equal Opportunity (EOp). These metrics, which are based on the separation definition ($A \perp \hat{Y} \mid Y$) [41], align well with ethical perspectives and anti-discrimination regulations [96, 288]. Their emphasis on ensuring group independence in the decision-making has led to their practical recognition, as evidenced by their inclusion in international standards [220].

Equal opportunity [206] defines fairness as predicting the positive outcome independently of the protected group attribute, conditioned on the observed outcome being positive. Formally, equal opportunity is defined as:

$$P(\hat{Y} = 1 \mid A = a, Y = 1) = P(\hat{Y} = 1 \mid A = b, Y = 1). \tag{2}$$

Equalized odds [206] is a stricter version of equal opportunity, requiring predictive independence conditioned on both positive and negative observed outcomes. Formally, the equalized odds measure is defined as:

$$P(\hat{Y} = 1 \mid A = a, Y = y) = P(\hat{Y} = 1 \mid A = b, Y = y), \quad y \in \{0, 1\}. \tag{3}$$

These fairness criteria are measured in practice with the following relaxations. For EOp, the fairness metric is the difference in TPR between different groups. For EOdds, the fairness metric considers both the TPR and FPR differences between groups:

$$\begin{aligned} \text{EOp} &:= \text{TPR}_{A=a} - \text{TPR}_{A=b}, \\ \text{EOdds} &:= \frac{1}{2}(\text{FPR}_{A=a} - \text{FPR}_{A=b}) + \frac{1}{2}(\text{TPR}_{A=a} - \text{TPR}_{A=b}). \end{aligned} \tag{4}$$

2.5 FairShap: Fair Shapley Values

We propose valuation functions that consider the model's group fairness while sharing the same axioms as the Shapley Values.

A straightforward implementation of FairShap is intractable. Thus, to address such a limitation, FairShap leverages the efficiency axiom of the SVs and the decomposability of the fairness metrics to obtain a tractable solution [198, 398]. In the following, we derive the expressions to compute the weights of a dataset according to FairShap in a binary classification case (Y is a binary variable) and with binary protected attributes. The extension to non-binary protected attributes and multi-class scenarios is provided in Appendix A.3.4.

2.5.1 Pairwise Contributions $\Phi_{i,j}$.

Group fairness depends on the disparity in a model's error rates on different groups of data points in the test set when the groups are defined according to their values of a protected attribute, A. To measure the data valuation for a training data point to the model's fairness, it is essential to identify the contribution of that training data point to the model's accuracy on the different groups of the test set defined by their protected attribute. The value of datapoint i, ϕ_i, is obtained by aggregating the pairwise accuracy contributions $\Phi_{i,j}$

across relevant test subgroups. Thus, $\Phi_{i,j}$ acts as the core building block from which group-specific expectations (as per Equation (1)) and fairness valuations (as per Equation (4)) are constructed.

Let $\mathcal{D}$ be the training dataset, $\mathcal{T}$ a reference dataset (typically the validation set in $\mathcal{D}$) and $\Phi_{i,j}$ the contribution of the training point $(x_i, y_i) \in \mathcal{D}$ (or i in short) to the probability of correctly classifying the test point $(x_j, y_j) \in \mathcal{T}$ (j in brief). Specifically, $\Phi_{i,j}$ measures the expected change in the model's prediction accuracy for j due to the inclusion of i in the dataset:

$$\Phi_{i,j} = \mathbb{E}_{S \sim \mathcal{P}(\mathcal{D} \setminus \{i\})}[\underbrace{p(y = y_j | x_j, S \cup \{i\}) - p(y = y_j | x_j, S)}_{\text{LOO}(i,j,S)}]. \tag{5}$$

The predictive distribution for datapoint j of a model trained with dataset $\mathcal{D}$ is denoted by $p(y_j|x_j, \mathcal{D})$ and the probability of correct classification of j using that model is given by $p(y = y_j|x_j, \mathcal{D})$. Therefore, $p(y = y_j \mid x_j, S \cup i) - p(y = y_j \mid x_j, S)$ represents the leave-one-out contribution (LOO) of training datapoint i to the correct classification of j when trained on subset S, denoted as $\text{LOO}(i, j, S)$. Consequently, $\Phi_{i,j}$ is the expected $\text{LOO}(i, j, S)$ over all subsets S of $\mathcal{D}$. Finally, let $\mathbf{\Phi} \in \mathbb{R}^{|\mathcal{D}| \times |\mathcal{T}|}$ be the matrix where each element corresponds to the contribution of the pairwise train-test data points, leveraging the efficiency axiom, the contribution of a point to the accuracy of the model with this formulation can be computed as

$$\phi_i(\text{Acc}) := \mathbb{E}_{j \sim p(\mathcal{T})}[\Phi_{i,j}] = \overline{\mathbf{\Phi}}_{i,:} \in \mathbb{R}. \tag{6}$$

While a direct computation of $\Phi_{i,j}$ is computationally prohibitive ($O(2^N)$), Jia et al. [223] propose an efficient implementation ($O(N \log N)$) for k-NN models (see Appendix A.3.1). This method does not depend on a trained prediction model, a decision threshold, or re-training Hammoudeh and Lowd [205] and Jiang et al. [226]. Although only theoretically exact for k-NN classifiers, it has been shown to provide accurate approximations of $\phi(\text{Acc})$ for a range of model types and data modalities, including images and text embeddings [226]. Moreover, it avoids the sampling errors inherent in Monte Carlo approximations of Shapley values and achieves superior runtime performance compared to alternative methods [226].

Note that Jia et al. [223] focus solely on the computation of $\phi(\text{Acc})$ and do not explore the formulation, interpretation, or potential applications of the pairwise contribution, $\Phi_{i,j}$, in the context of algorithmic fairness, as we do in this chapter.

Threshold-independence. In addition, this approach is threshold-independent: it computes the accuracy as the average probability of correct classification across all test points, rather than relying on a fixed decision threshold. Threshold independence is beneficial because it decouples the Shapley value computations from arbitrary decision boundaries, yielding more stable, reliable, and model-agnostic evaluations of instance-level contributions.

In the standard formulation of SV, fairness metrics depend on classification outcomes defined by a threshold t, such as: $\text{TP} = |\{\hat{Y} > t|Y = 1\}|$, $\text{TN} = |\{\hat{Y} < t|Y = 0\}|$, $\text{FP} = |\{\hat{Y} > t|Y = 0\}|$ and $\text{FN} = |\{\hat{Y} < t|Y = 1\}|$. This dependence introduces variability in the valuation. In contrast, when computing the pairwise contributions, Φ_{ij}, this approach uses the predicted probability of correct classification, thus avoiding the dependency on the threshold of `FairShap`.

2.5.2 Shapley Values for Algorithmic Fairness.

Classification rates, such as TPR and TNR, are the building blocks of EOdds and EOp. Let $\phi_i(\text{TPR})$ and $\phi_i(\text{TNR})$ be two valuation functions that measure the contribution of training point i to the TPR and TNR, respectively. Note that $\text{TPR} = \text{Acc}|_{Y=1}$ and $\text{TNR} = \text{Acc}|_{Y=0}$. Therefore, $\phi_i(\text{TPR})$ corresponds to the expected change in the model's probability of correctly predicting the positive class when point i is included in the training dataset $\mathcal{D}$, considering all possible training dataset subsets and the distribution of the reference dataset:

$$\phi_i(\text{TPR}) := \mathbb{E}_{j\sim p(\mathcal{T}|Y=1)}\left[\mathbb{E}_{S\sim\mathcal{P}(\mathcal{D}\setminus\{i\})}\left[\text{LOO}(i,j,S)\right]\right] = \mathbb{E}_{j\sim p(\mathcal{T}|Y=1)}\left[\Phi_{i,j}\right]. \tag{7}$$

The value for the entire dataset is $\boldsymbol{\phi}(\text{TPR}) = [\phi_0(\text{TPR}), \cdots, \phi_n(\text{TPR})] \in \mathbb{R}^{|\mathcal{D}|}$. $\boldsymbol{\phi}(\text{TNR})$ is obtained similarly but for $Y = 0$ and $y = 0$. In addition, $\phi_i(\text{FNR}) = \frac{1}{|\mathcal{D}|} - \phi_i(\text{TPR})$ and $\phi_i(\text{FPR}) = \frac{1}{|\mathcal{D}|} - \phi_i(\text{TNR})$. These four functions fulfill the SV axioms. More details about these metrics can be found in Appendix A.2.

Intuitively, $\boldsymbol{\phi}(\text{TPR})$ and $\boldsymbol{\phi}(\text{TNR})$ quantify how much the training data points contribute to the correct classification when $y = 1$ and $y = 0$, respectively.

To illustrate $\boldsymbol{\phi}(\text{TPR})$ and $\boldsymbol{\phi}(\text{TNR})$, Figure 5 depicts the $\boldsymbol{\phi}(\text{TPR})$ and $\boldsymbol{\phi}(\text{TNR})$ of a simple synthetic example with two normally distributed classes.

Once $\phi_i(\text{TPR})$, $\phi_i(\text{TNR})$, $\phi_i(\text{FPR})$ and $\phi_i(\text{FNR})$ have been obtained, we can compute the `FairShap` weights for a given dataset. However, there are two scenarios to consider, depending on whether the sensitive attribute (A) and the target variable or label (Y) are the same.

FairShap weights when $A \neq Y$. This is the most common scenario. In this case, EOp and EOdds use true/false positive/negative rates conditioned on Y and A. Therefore, we define $\text{TPR}_{A=a} = \text{Acc}_{Y=y,A=a}$, or TPR_a for short, and thus

$$\phi_i(\text{TPR}_a) := \mathbb{E}_{j\sim p(\mathcal{T}|Y=1,A=a)}[\Phi_{i,j}] = \overline{\boldsymbol{\Phi}}_{i,:|Y=1,A=a} \tag{8}$$

where the value for the entire dataset is $\boldsymbol{\phi}(\text{TPR}_a) = [\phi_0(\text{TPR}_a), \cdots, \phi_n(\text{TPR}_a)]$. Intuitively, $\phi_i(\text{TPR}_a)$ measures the contribution of the training point i to the TPR of the testing points belonging to a given protected group ($A = a$). $\phi_i(\text{TNR}_a)$ is obtained similarly but for $y = 0$.

Given EOp and EOdds as per Equation (4), then $\phi_i(\text{EOp})$ is given by

$$\phi_i(\text{EOp}) := \phi_i(\text{TPR}_a) - \phi_i(\text{TPR}_b) \tag{9}$$

and $\phi_i(\text{EOdds})$ is expressed as

$$\phi_i(\text{EOdds}) := \frac{1}{2}(\phi_i(\text{FPR}_a) - \phi_i(\text{FPR}_b)) + \frac{1}{2}(\phi_i(\text{TPR}_a) - \phi_i(\text{TPR}_b)) \tag{10}$$

where their corresponding $\boldsymbol{\phi}(\text{EOp})$ and $\boldsymbol{\phi}(\text{EOdds})$ vectors. A detailed view of the complete step-by-step derivation of the equations above can be found in Appendix A.2 and Appendix A.3.2. Additionally, Algorithm 1 provides the pseudo-code to compute the data weights according to `FairShap`. Finally, empirical results are presented in Section 2.5.2, which describes a synthetic example showing the impact of $\phi(\cdot)$ on the decision boundaries and the group fairness metrics.

FairShap weights when $A = Y$. The case when $A = Y$ is relevant to evaluate fairness interventions under maximal dependence between target and sensitive attribute, a critical

Algorithm 1: `FairShap`: Instance-level data re-weighting for group fairness via data valuation

Data: Training set $\mathcal{D}$, reference set $\mathcal{T}$, protected groups A, parameter k
Result: SV matrix $\mathbf{\Phi} \in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{T}|}$ and `FairShap` vectors $\boldsymbol{\phi}(\text{EOp})$, $\boldsymbol{\phi}(\text{EOdds})$

Initialize $\mathbf{\Phi} \leftarrow \mathbf{0} \in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{T}|}$;
foreach $j \in \mathcal{T}$ **do**
  Order $i \in \mathcal{D}$ by L_2 distance to j: $\rightarrow (x_1, x_2, \ldots, x_N)$;
  $\Phi_{N,j} \leftarrow \frac{\mathbb{I}[y_{x_N}=y_j]}{N}$;
  for $i = N-1$ **to** 1 **do**
    `// Marginal contribution of datapoint` i `to correct prediction of` j
    $\Phi_{i,j} \leftarrow \Phi_{i+1,j} + \mathbb{I}[y_i = y_j] - \frac{\mathbb{I}[y_{i+1}=y_j]}{\max(k,i)}$;

`// Compute fairness-aware Shapley vectors`
$\boldsymbol{\phi}(\text{TPR}_a) \leftarrow$ Equation (8);
$\boldsymbol{\phi}(\text{FPR}_a) \leftarrow [\frac{1}{|\mathcal{D}|} - \phi_i(\text{TNR}_a) : \ \forall i \in \mathcal{D}] \quad \forall a \in A$;
$\boldsymbol{\phi}(\text{EOp}) \leftarrow$ Equation (9);
$\boldsymbol{\phi}(\text{EOdds}) \leftarrow$ Equation (10);
return $\mathbf{\Phi}$, $\boldsymbol{\phi}(\text{EOp})$, $\boldsymbol{\phi}(\text{EOdds})$;

test motivated by both societal concerns and prior fairness research [53, 84, 235, 251, 362], where models that predict sensitive attributes may lead to discriminatory outcomes, *e.g.*, in gender classification.

In this case, EOp and EOdds collapse to measure the disparity between TPR and TNR or FPR and FNR for the different values of Y [53], which, in a binary classification case, may be expressed as the Equal Opportunity measure computed as $\text{EOp} := \text{TPR} - \text{FPR} \in [-1, 1]$ or its scaled version $\text{EOp} = (\text{TPR} + \text{TNR})/2 \in [0, 1]$. Thus, the $\phi_i(\text{EOp})$ of data point i may be expressed as

$$\phi_i(\text{EOp}) := \frac{\phi_i(\text{TPR}) + \phi_i(\text{TNR})}{2}. \tag{11}$$

We refer the reader to Appendix A.3.3 for more details. Empirical results for this scenario are presented in Section 2.6.2.

Instance-level data re-weighting with `FairShap`. The definition of a data valuation function states that the higher the value, the more the data point contributes to the measure. Yet, it does not necessarily mean that a higher value is more desirable: it depends on the value function of choice. In the case of accuracy, a higher value denotes a larger contribution to accuracy [187]. In the case of fairness, we prioritize points with high $-\phi(\text{EOp}) = \phi_i(\text{TPR}_B) - \phi_i(\text{TPR}_A)$, where B is the discriminated group (*i.e.*, $\text{TPR}_A > \text{TPR}_B$). Giving more weight to data points with a positive $-\phi_i(\text{EOp})$ will increase the discriminated group's TPR, balancing the difference in TPR between groups and thus yielding a smaller EOp and a fairer model. In the experimental section, we denote the re-weighting with $-\phi(\text{EOp})$ as $\phi(\text{EOp})$ for simplicity, and the same for $\phi_i(\text{EOdds})$.

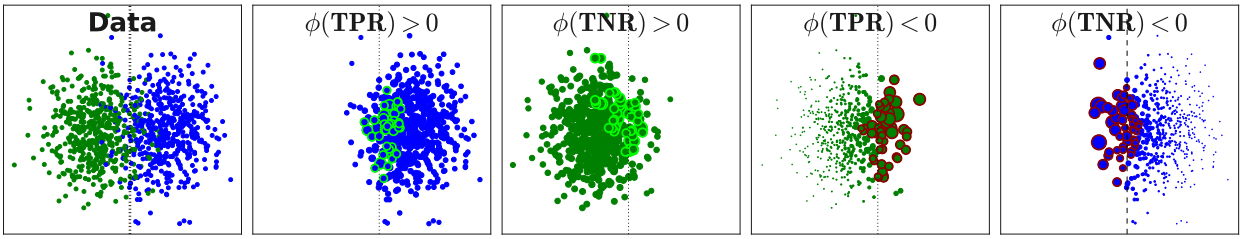
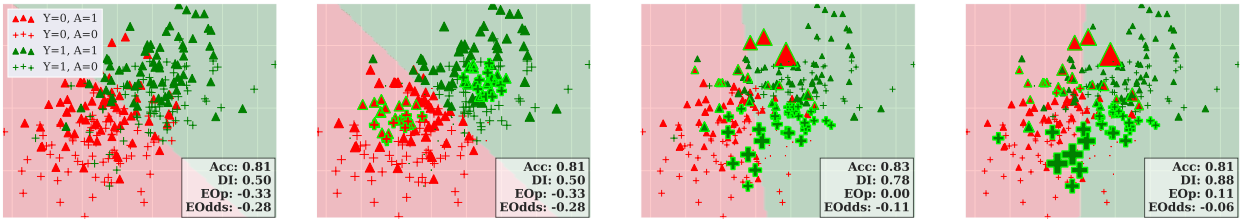


Figure 5. Synthetic example with positive ($Y = 1$, blue) and negative ($Y = 0$,green) classes. Data points with the 50 largest (green) and smallest (red) ϕ_i are highlighted. Size $\propto |\phi_i|$.



(a) **(b)** **(c)** **(d)**

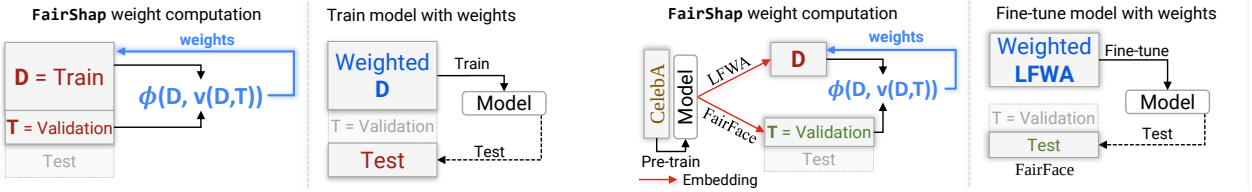
Figure 6. Synthetic example based on Case I in Zafar et al. [415], where group FPR and FNR differ in sign. (a) shows the original data; in (b–d), point size is proportional to (b) $\phi(\text{Acc})$, (c) $\phi(\text{EOp})$, and (d) $\phi(\text{EOdds})$, with the top-50 points in green. Favorable outcome (Y=1) is shown in **green**; unfavorable (Y=0) in **red**. Privileged group (A=1) is marked as triangles ▲, disadvantaged (A=0) as crosses +. Logistic Regression models are trained on re-weighted data and evaluated on the same test split; shaded regions indicate decision boundaries.

Illustration of FairShap in Synthetic Datasets

$\phi(\text{TPR})$ and $\phi(\text{TNR})$. We use a synthetic binary classification task with two Gaussian distributions to visualize $\phi(\text{TPR})$ and $\phi(\text{TNR})$. As shown in Figure 5, points with large $\phi(\text{TPR})$ are positive-class instances correctly placed near the decision boundary, while those with small $\phi(\text{TPR})$ are misaligned. The same applies to $\phi(\text{TNR})$ for negative-class instances. This highlights how different data points contribute to the respective rates.

$\mathbf{A \neq Y}$. To explore fairness-aware valuations, we generate synthetic data following Case I from Zafar et al. [415], where group FPR and FNR differ in sign: privileged groups ($A = 1$, triangles) show higher FPR, while disadvantaged groups ($A = 0$, crosses) show higher FNR. Figure 6 shows how point sizes reflect $|\phi(\cdot)|$ values across fairness metrics, with top-50 influential points highlighted in green.

We train logistic regression models on data re-weighted by $\phi(\text{Acc})$, $\phi(\text{EOp})$, or $\phi(\text{EOdds})$, and evaluate on the same test split. As seen in Figure 6, using $\phi(\text{EOp})$ and $\phi(\text{EOdds})$ shifts the decision boundaries toward fairer models while preserving or improving accuracy. These valuations primarily emphasize unfavorable-privileged (red triangles) and favorable-disadvantaged (green crosses) instances.



(a) Tabular data. $A \neq Y$ and $\mathcal{T}$ is the validation set in $\mathcal{D}$.

(b) Image data. $A = Y$ and $\mathcal{T}$ is an auxiliary and bias-aware curated dataset.

Figure 7. Pipelines of the experiments described in Section 2.6.1 (a) and Section 2.6.2 (b).

2.6 Experiments

2.6.1 Instance-level Data Re-weighing on Tabular Datasets

This section considers a common real-life scenario on benchmark tabular datasets, where the models predict a target variable $A \neq Y$. Also, a single dataset $\mathcal{D}$ is used for training, validation, and testing. Thus, the validation ($\mathcal{T}$) and test sets are obtained as a partition of $\mathcal{D}$ according to the pipeline illustrated in Figure 7a.[8]

Datasets. We test `FairShap` on three commonly used datasets in the algorithmic fairness literature: (1) the German Credit [230] dataset (German) with a binary target variable corresponding to an individual's good or bad *credit risk*, and protected attributes *age* and *sex*; (2) the Adult Income dataset [249] (Adult) where the task is to predict whether the *income* of a person is more than 50k per year, and *sex* and *race* are the protected attributes; and (3) the COMPAS [16] dataset with binary target variable *recidivism* and protected attributes *sex* and *race*. See Appendix A.4 for more details on the datasets.

Pipeline. The model in all experiments is a Gradient Boosting Classifier (GBC) [179], known for its competitive performance on tabular data and interpretability properties. The pipeline in this set of experiments is depicted in Figure 7a. The Figure shows that the reference dataset $\mathcal{T}$ is the validation set of $\mathcal{D}$. The reported results correspond to the average values of running the experiment 50 times with random splits stratified by sensitive group and label: 70% of the original dataset used for training ($\mathcal{D}$), 15% for the reference set ($\mathcal{T}$), and 15% for the test set. Train, reference, and test set are stratified by A and Y such that they have the same percentage of A - Y samples as in the original dataset.

FairShap re-weighting. Given that $A \neq Y$, `FairShap` considers two different group fairness-based valuation functions: $\boldsymbol{\phi}(\text{EOp})$ and $\boldsymbol{\phi}(\text{EOdds})$ as per Equation (9) and Equation (10), respectively. Note that in this case the weights assigned to each data point, w_i, are obtained by normalizing $\boldsymbol{\phi}(\text{EOp})$ and $\boldsymbol{\phi}(\text{EOdds})$ following a methodology similar to that described in [98]: $w_i = \frac{\phi_i}{\sum_i \phi'_i}|D|$ where $\phi'_i = \frac{\phi_i - \min(\boldsymbol{\phi})}{\max(\boldsymbol{\phi}) - \min(\boldsymbol{\phi})}$.

Baselines. To the best of our knowledge, `FairShap` is the only interpretable, instance-level model-agnostic data re-weighting approach for group algorithmic fairness (see Table 1) by means of data valuation. Thus, we compare its performance with the following six state-of-the-art algorithmic fairness approaches that only *partially* satisfy `FairShap`'s properties. *(1) Group RW*: A group-based re-weighting method that assigns the same weights to all samples from the same category or group according to the protected attribute [231]. Thus, this is not an instance-level re-weighting approach;

[8] The code of `FairShap` is publicly available at `https://github.com/ellisalicante/fair-shap`.

(2) Post-pro: A post-processing algorithmic fairness method that does not fulfill any of the desiderata, but it is broadly used in the community [206];
(3) LabelBias: A model that learns the weights in an in-processing manner and therefore it is neither a pre-processing nor a model-agnostic approach [225];
(4) Opt-Pre: A model-agnostic pre-processing approach for algorithmic fairness based on feature and label transformations which does not assign any weights to the data [90];
(5) IFs: An Influence Function (IF)-based approach, which is an in-processing retraining method, since the weights are computed from the Hessian of a pretrained model. We use the same hyperparameters reported by the authors for each of the datasets [263]; and
(6) $\boldsymbol{\phi}(Acc)$: A data re-weighting method by means of an accuracy-based Shapley valuation function [187]. The weights assigned to each data point are obtained according to the same normalization as in `FairShap`'s case.

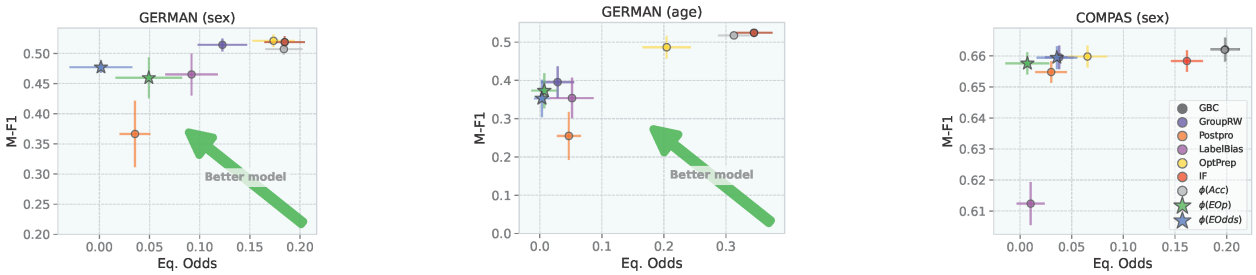


Figure 8. Performance-Fairness Pareto front. Accuracy (M-F1) vs fairness analysis. The models trained with data re-weighted via `FairShap` (depicted as stars) improve fairness while preserving competitive levels of accuracy.

Experimental setup. We adopt the experimental setup that is commonly followed in the ML community: the weights, the influence functions, and the thresholds required by the different methods are computed on the validation set. Furthermore, all the reported results correspond to the mean and standard deviation values of running 50 experiments on each dataset with random stratified train, validation and test set splits in each experiment, as previously described. Note that some previous works in the algorithmic fairness literature do not perform label-group stratification on the splits, or compute the weights or thresholds using the test set instead of the validation set. Hence, reported performances in the literature are not directly comparable to ours.

Results. The metrics used for evaluation are accuracy (Acc); Macro-F1 (M-F1), which is an extension of the F1 score that addresses class imbalance, as is the case in our datasets; EOp and EOdds. Table 2 summarizes the results, highlighting the **best** and *second-best* performing methods. The arrows indicate if the optimal result is 0 (↓) or 1 (↑).

As shown in Table 2 and Figure 8, data re-weighting with `FairShap` ($\boldsymbol{\phi}(\text{EOdds})$ and $\boldsymbol{\phi}(\text{EOp})$) generally yields better fairness results than the baselines while keeping competitive levels of accuracy. In the COMPAS (sex) and Adult (sex and race) datasets, data re-weighting via `FairShap` yields the fairest models while maintaining performance metrics that are not significantly different from the best-performing method. The improvement in fairness is notable when compared to the performance of the model built without data re-weighing (GBC). For example, in the German dataset with sex as a protected attribute, the model's Equalized Odds metric is **93x** better when re-weighting via `FairShap` ($\boldsymbol{\phi}(\text{EOdds})$) than the baseline model (GBC) and **18x** better than the most competitive baseline (PostPro). In all

	Sex				Age			
German	Accuracy ↑	M-F1 ↑	EOp ↓	EOdds ↓	Accuracy ↑	M-F1 ↑	EOp ↓	EOdds ↓
GBC	*.697*$_{\pm.006}$	*.519*$_{\pm.010}$	‡.107$_{\pm.020}$	‡.185$_{\pm.020}$	†*.704*$_{\pm.005}$	**.524**$_{\pm.010}$	‡.224$_{\pm.032}$	‡.345$_{\pm.030}$
Group RW	.695$_{\pm.006}$	.514$_{\pm.010}$	‡.062$_{\pm.019}$	‡.123$_{\pm.025}$	‡.684$_{\pm.004}$	‡.396$_{\pm.041}$	‡.040$_{\pm.025}$	‡.029$_{\pm.026}$
Postpro	‡.691$_{\pm.005}$	‡.366$_{\pm.055}$	*.013*$_{\pm.014}$	†*.036*$_{\pm.015}$	‡.686$_{\pm.005}$	‡.255$_{\pm.063}$	‡.044$_{\pm.022}$	‡.047$_{\pm.019}$
LabelBias	.695$_{\pm.006}$	‡.465$_{\pm.035}$	‡.051$_{\pm.019}$	‡.092$_{\pm.026}$	‡.690$_{\pm.004}$	‡.354$_{\pm.053}$	‡.052$_{\pm.029}$	‡.052$_{\pm.035}$
OptPrep	.694$_{\pm.006}$	**.521**$_{\pm.010}$	‡.104$_{\pm.022}$	‡.174$_{\pm.021}$	‡.693$_{\pm.007}$	‡.487$_{\pm.030}$	‡.130$_{\pm.031}$	‡.204$_{\pm.039}$
IF	*.697*$_{\pm.006}$	*.519*$_{\pm.010}$	‡.107$_{\pm.020}$	‡.185$_{\pm.020}$	†*.704*$_{\pm.005}$	**.524**$_{\pm.010}$	‡.224$_{\pm.032}$	‡.345$_{\pm.030}$
ϕ(Acc)	**.700**$_{\pm.005}$	†.507$_{\pm.009}$	‡.097$_{\pm.018}$	‡.184$_{\pm.018}$	**.706**$_{\pm.005}$	‡.517$_{\pm.010}$	‡.193$_{\pm.025}$	‡.313$_{\pm.025}$
$\boldsymbol{\phi}$(EOp)	‡.683$_{\pm.006}$	‡.460$_{\pm.034}$	.029$_{\pm.026}$	†.049$_{\pm.033}$	‡.685$_{\pm.004}$	‡.373$_{\pm.046}$	*.024*$_{\pm.023}$	*.007*$_{\pm.021}$
$\boldsymbol{\phi}$(EOdds)	‡.686$_{\pm.006}$	‡.477$_{\pm.009}$	**.002**$_{\pm.025}$	**.002**$_{\pm.031}$	‡.681$_{\pm.005}$	‡.353$_{\pm.049}$	**.019**$_{\pm.020}$	**.003**$_{\pm.013}$

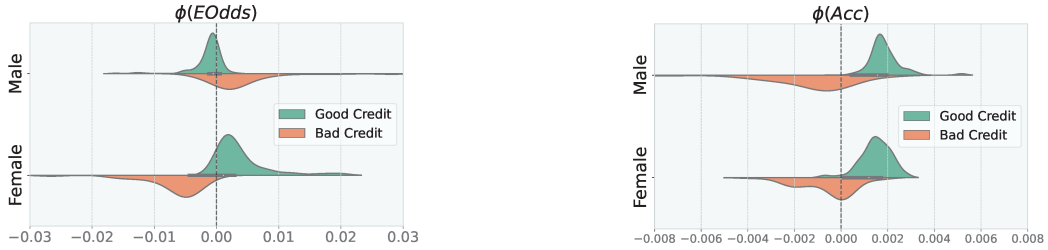
	Sex				Race			
Adult	Accuracy ↑	M-F1 ↑	EOp ↓	EOdds ↓	Accuracy ↑	M-F1 ↑	EOp ↓	EOdds ↓
GBC	*.803*$_{\pm.001}$	‡.680$_{\pm.002}$	‡.451$_{\pm.004}$	‡.278$_{\pm.003}$	**.803**$_{\pm.001}$	‡.682$_{\pm.002}$	‡.164$_{\pm.010}$	‡.106$_{\pm.006}$
Group RW	‡.790$_{\pm.001}$	**.684**$_{\pm.002}$	*.002*$_{\pm.009}$	*.001*$_{\pm.005}$	**.803**$_{\pm.001}$	‡.683$_{\pm.002}$	.010$_{\pm.009}$	.010$_{\pm.005}$
Postpro	‡.791$_{\pm.001}$	†.679$_{\pm.004}$	‡.056$_{\pm.013}$	‡.034$_{\pm.007}$	.802$_{\pm.001}$	**.688**$_{\pm.002}$	‡.061$_{\pm.011}$	‡.042$_{\pm.006}$
LabelBias	‡.781$_{\pm.001}$	‡.681$_{\pm.002}$	‡.065$_{\pm.011}$	‡.049$_{\pm.006}$	‡.800$_{\pm.001}$	*.686*$_{\pm.002}$	‡.118$_{\pm.013}$	‡.074$_{\pm.007}$
OptPrep	‡.789$_{\pm.001}$	‡.676$_{\pm.004}$	‡.064$_{\pm.029}$	‡.037$_{\pm.017}$	‡.800$_{\pm.001}$	†.685$_{\pm.002}$	‡.044$_{\pm.015}$	‡.029$_{\pm.009}$
IF	‡.787$_{\pm.002}$	†.681$_{\pm.003}$	‡.159$_{\pm.037}$	‡.092$_{\pm.022}$	‡.797$_{\pm.002}$	†.685$_{\pm.002}$	‡.042$_{\pm.020}$	‡.031$_{\pm.012}$
ϕ(Acc)	**.804**$_{\pm.001}$	‡.681$_{\pm.002}$	‡.452$_{\pm.005}$	‡.279$_{\pm.003}$	**.803**$_{\pm.001}$	‡.681$_{\pm.002}$	‡.161$_{\pm.011}$	‡.104$_{\pm.007}$
$\boldsymbol{\phi}$(EOp)	‡.790$_{\pm.001}$	**.684**$_{\pm.002}$	*.002*$_{\pm.009}$	**3e-4**$_{\pm.005}$	.802$_{\pm.001}$	‡.683$_{\pm.002}$	*.009*$_{\pm.010}$	*.009*$_{\pm.005}$
$\boldsymbol{\phi}$(EOdds)	‡.790$_{\pm.001}$	.683$_{\pm.002}$	**8e-4**$_{\pm.009}$	*.001*$_{\pm.005}$	.802$_{\pm.001}$	‡.683$_{\pm.002}$	**.007**$_{\pm.009}$	**.007**$_{\pm.005}$

	Sex				Race			
COMPAS	Accuracy ↑	M-F1 ↑	EOp ↓	EOdds ↓	Accuracy ↑	M-F1 ↑	EOp ↓	EOdds ↓
GBC	*.666*$_{\pm.004}$	**.662**$_{\pm.004}$	‡.158$_{\pm.014}$	‡.199$_{\pm.014}$	**.663**$_{\pm.004}$	**.658**$_{\pm.004}$	‡.184$_{\pm.013}$	‡.218$_{\pm.013}$
Group RW	.664$_{\pm.004}$	.660$_{\pm.004}$	.020$_{\pm.016}$	‡.038$_{\pm.014}$	‡.649$_{\pm.004}$	‡.646$_{\pm.004}$	‡.028$_{\pm.015}$	.007$_{\pm.016}$
Postpro	‡.660$_{\pm.003}$	‡.655$_{\pm.003}$	.017$_{\pm.017}$	†.030$_{\pm.015}$	‡.647$_{\pm.005}$	‡.642$_{\pm.005}$	**1e-4**$_{\pm.015}$	‡.026$_{\pm.016}$
LabelBias	‡.639$_{\pm.005}$	‡.612$_{\pm.007}$	**.006**$_{\pm.013}$	*.010*$_{\pm.014}$	‡.645$_{\pm.004}$	‡.627$_{\pm.005}$	‡.030$_{\pm.011}$	‡.045$_{\pm.014}$
OptPrep	.664$_{\pm.003}$	.660$_{\pm.003}$	‡.045$_{\pm.020}$	‡.065$_{\pm.019}$	‡.655$_{\pm.004}$	†.651$_{\pm.004}$	‡.044$_{\pm.020}$	‡.078$_{\pm.020}$
IF	.663$_{\pm.003}$	.658$_{\pm.003}$	‡.129$_{\pm.016}$	‡.161$_{\pm.015}$	.660$_{\pm.004}$	.655$_{\pm.004}$	‡.165$_{\pm.017}$	‡.198$_{\pm.015}$
ϕ(Acc)	**.667**$_{\pm.004}$	**.662**$_{\pm.004}$	‡.156$_{\pm.014}$	‡.198$_{\pm.013}$	**.663**$_{\pm.004}$	*.657*$_{\pm.004}$	‡.184$_{\pm.013}$	‡.218$_{\pm.013}$
$\boldsymbol{\phi}$(EOp)	†.661$_{\pm.003}$	†.658$_{\pm.004}$	*.013*$_{\pm.024}$	**.007**$_{\pm.021}$	‡.650$_{\pm.004}$	‡.647$_{\pm.004}$	‡*.027*$_{\pm.016}$	**.004**$_{\pm.017}$
$\boldsymbol{\phi}$(EOdds)	.663$_{\pm.004}$	.659$_{\pm.004}$	.019$_{\pm.021}$	†.036$_{\pm.020}$	‡.648$_{\pm.004}$	‡.646$_{\pm.004}$	‡.036$_{\pm.017}$	**.004**$_{\pm.018}$

Table 2. Performance of GBC with and without data re-weighting on benchmark datasets with different sensitive attributes. The **best** and *second-best* performing methods are color-coded. Statistically significant differences with the best-performing model are denoted by ‡ for $p < 0.01$ and † for $p < 0.05$. Grey shaded columns indicate fairness metrics, where low values are desired. Note how data re-weighting by means of `FairShap` yields the fairest models in most cases with a small loss in accuracy.

the experiments, data re-weighting by means of `FairShap` is the best and/or second-best method regarding fairness.

From the results, we draw several observations. First, no single method consistently yields the best performance (accuracy and/or M-F1) and group fairness results in all the datasets and for all protected attributes. Second, a simple method such as Group RW delivers very competitive results, even better than more sophisticated, recent approaches. Third, accuracy



(a) Histogram of $\phi_i(\text{EOdds})$ **(b)** Histogram of $\phi_i(\text{Acc})$

Figure 9. SV Distributions for $\phi_i(\textbf{EOdds})$ and $\phi_i(\textbf{Acc})$ on the German Credit dataset with $A = \text{sex}$.

is not an appropriate metric of the performance of the classifier due to the imbalance of the datasets, and M-F1 would be the recommended measure to assess classification performance. Finally, data re-weighting by means of `FairShap` is the method that most consistently yields competitive performance and fairness results.

Interestingly, in some cases, optimizing for EOdds yields better performance in EOp than directly optimizing for EOp (as seen in the *german-age* setting), which is probably due to two reasons: the weight computation is performed on a validation set, and the inherent stochasticity of sampling-based strategies in the datasets can introduce significant variability in the results [43]; and the fairness metrics are estimated using a k-nearest neighbor approximation, which may further contribute to fluctuations in the reported outcomes.

To shed further light on the behavior of data re-weighting by means of `FairShap`, Figure 9 depicts the histograms of $\boldsymbol{\phi}(\text{EOdds})$ and $\boldsymbol{\phi}(\text{Acc})$ on the German Credit dataset with sex as protected attribute. Note how the distribution of $\boldsymbol{\phi}(\text{Acc})$ is similar for males and females, even though the dataset is highly imbalanced: examples with good credit, irrespective of sex, receive larger weights than those with bad credit. Conversely, the $\boldsymbol{\phi}(\text{EOdds})$ values are larger for female applicants with good credit than for their male counterparts. In addition, $\boldsymbol{\phi}(\text{EOdds})$ are larger for male applicants with bad credit than for their female counterparts. These distributions of $\boldsymbol{\phi}(\text{EOdds})$ compensate for the imbalances in the raw dataset (both in terms of sex and credit risk), yielding fairer classifiers, as reflected in the results reported in Table 2.

Finally, an ablation study of the impact of the size of the reference dataset $\mathcal{T}$ on the performance of the resulting classifier and an analysis of the running time can be found in Section 2.6.4 and Section 2.6.5, respectively.

Accuracy vs fairness

Shapley Values provide a way to assign weights to individual data points based on their contributions to a particular function. While Shapley Values can be used to re-weight data points to improve fairness according to a specific fairness metric, the impact on accuracy is not guaranteed to be consistent across all datasets and scenarios. However, in many scenarios, particularly when there are many examples in the majority group, data re-weighting by means of `FairShap` maintains accuracy for the majority group while improving the model's fairness by increasing the accuracy in the disadvantaged group. We observe this behavior in all our experiments.

To further illustrate the impact of `FairShap`'s data re-weighting on the model's accuracy

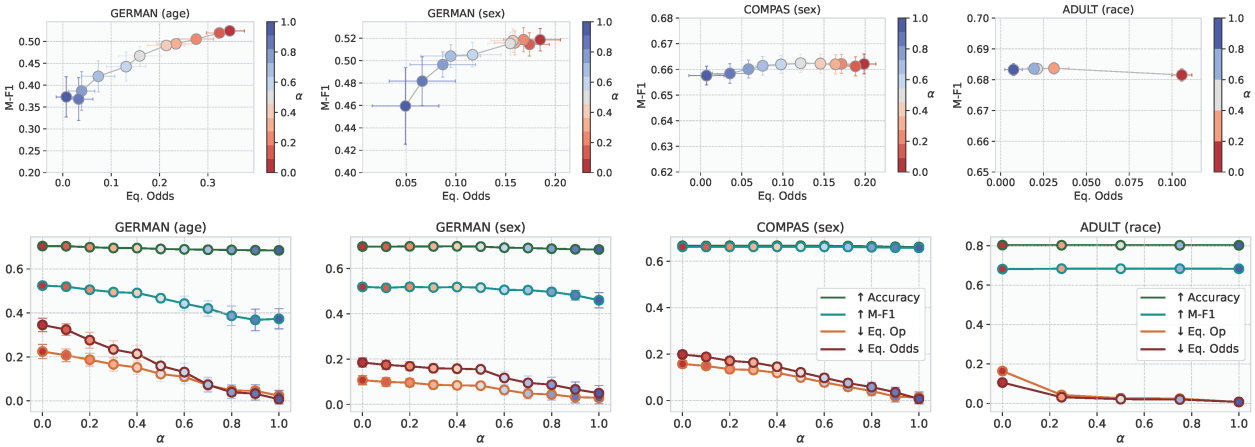


Figure 10. Accuracy vs fairness trade-off for different values of α (strength of **FairShap** re-weighting), where $\alpha = 0$ correspond to no data re-weighting and $\alpha = 1$ to re-weighting according to `FairShap`. Results show the mean and 95% CI over 50 random iterations for three datasets. $\mathbf{\Phi}(\text{EOp})$ is used to re-weight the German and COMPAS datasets, and $\mathbf{\Phi}(\text{EOdds})$ to re-weight in the Adult dataset. Top graphs show the accuracy-fairness Pareto front. The bottom graphs illustrate the Accuracy, M-F1, EOp, and EOdds for increasing values of α.

and fairness, Figure 10 depicts the utility-fairness curves on the three benchmark datasets (German, COMPAS, and Adult) with different protected attributes: age and sex (German), sex (COMPAS), and race (Adult). We define a parameter α that controls the contribution of the weights of each data point according to `FairShap`, ranging from $\alpha = 0$ (no data re-weighting) to $\alpha = 1$ (weights as given by `FairShap`). Thus, the weights of each data point i are computed as $w_i' = (1 - \alpha) + \alpha w_i$ where w_i is the weight of x_i according to `FairShap`.

As shown in the Figure, the larger the α, *i.e.*, the larger the importance of `FairShap`'s weights, the better the model's fairness. In some scenarios, such as on the German (age and sex) dataset, we observe a utility-fairness trade-off where the fairest models correspond to $\alpha = 1$ and the best performing models correspond to $\alpha = 0$. Conversely, on the COMPAS (sex) and Adult (race) datasets, larger values of α significantly increase the fairness of the model while keeping similar levels of utility (M-F1 and Accuracy).

2.6.2 Instance-level Data Re-weighting on Image Data and $A = Y$

Impact of biased-curated datasets. Models trained on biased datasets may appear fair and accurate when evaluated on similarly biased data, but their performance can degrade significantly when tested on bias-aware curated datasets. This underscores the importance of using fair reference datasets for evaluation. Biased training data can lead to models that perpetuate societal inequities, reinforcing stereotypes and producing discriminatory outcomes for underrepresented groups.

To illustrate this, we compare the performance of a model trained and tested on three dataset combinations: two biased datasets (LFWA and CelebA) and one curated for fairness (FairFace). As shown in Table 3, models trained on LFWA or CelebA perform well when tested on the same or similarly biased data. However, their accuracy and fairness drop considerably when evaluated on FairFace.

Train \ Test	FairFace	LFWA	CelebA
FairFace	90.9 | 0.01	95.7 | 0.03	96.7 | 0.09
LFWA	77.2 | 0.49	96.6 | 0.08	98.3 | 0.02
CelebA	76.1 | 0.61	96.9 | 0.09	98.2 | 0.01

Table 3. Accuracy ↑ and Accuracy Disparity ↓ for sex classification using Inception Resnet V1 across different train/test dataset combinations. Performance drops when models trained on biased datasets (LFWA, CelebA) are evaluated on the fair reference dataset (FairFace; red).

Experiments setup. In this scenario, we focus on a computer vision task to illustrate the versatility of data re-weighting via `FairShap`. In this case, the goal is to predict the sensitive attribute, *i.e.*, $A = Y$. This setting reflects a case of maximal dependence between A and Y, and a real-world scenario, as seen in commercial gender classification systems from facial images, where models directly predict sensitive attributes and have been shown to exhibit significant bias in practice [84]. Furthermore, this scenario explores the benefits of leveraging an external reference dataset $\mathcal{T}$, in contrast to Section 2.6.1 where the validation set of $\mathcal{D}$ is used as $\mathcal{T}$.

The task consists of a binary gender (male/female) classification[9] from facial images by means of a deep convolutional network (Inception Resnet V1) using `FairShap` for data re-weighting. Binary gender (male/female) is therefore both the protected attribute (A) and the target variable (Y). The pipeline of this scenario is illustrated in Figure 7b.

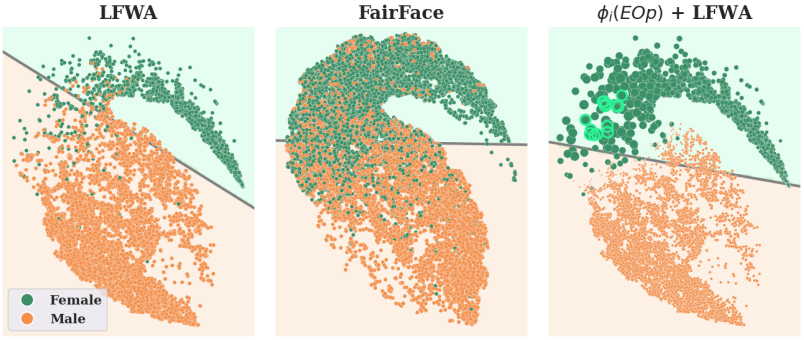
Datasets. We leverage three publicly available face datasets: CelebA, LFWA [272], and FairFace [235], where LFWA is the training set $\mathcal{D}$ (large-scale and biased) and FairFace is the reference dataset $\mathcal{T}$ (a bias-aware curated small dataset). The test split in the FairFace dataset is used for testing. CelebA is used to pre-train the Inception Resnet V1 model [373] to obtain the LFWA and FairFace embeddings that are needed to compute the Shapley Values efficiently by means of a k-NN approximation in the embedding space. In the three datasets, gender is a binary variable with two values: male and female.

Pipeline. The pipeline to obtain the `FairShap`'s weights in this scenario is depicted in Figure 7b and proceeds as follows: (1) Pre-train an Inception Resnet V1 model with the CelebA dataset; (2) Use this model to obtain the embeddings of the LFWA and FairFace datasets; (3) Compute the weights on the LFWA training set ($\mathcal{D}$) using as reference dataset ($\mathcal{T}$) the FairFace validation partition. (4) Fine-tune the pretrained model using the re-weighted data in the LFWA training set according to $\boldsymbol{\phi}$; and (5) Test the resulting model on the test partition of the FairFace dataset.

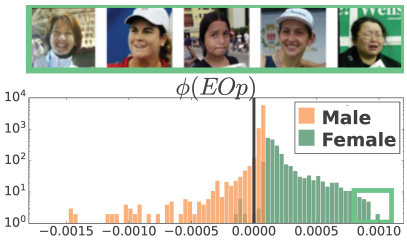
We used Binary cross-entropy loss and the Adam optimizer in both training phases. The learning rate was set to 0.001 for pre-training and reduced to 0.0005 for fine-tuning, each lasting 100 epochs. Training batches consisted of 128 images with an input shape of (160×160), and a patience parameter of 30 was employed for early stopping, saving the model with the highest accuracy on the validation set. The classification threshold for this model was set at 0.5.

Data Re-weighting with `FairShap`. In this case, as $A = Y$, EOp and EOdds are equivalent

[9] We use a binary label as this is the ground truth available in the datasets which ignores the diversity of human gender identities and expressions.



(a) Left: LFWA embeddings. Middle: FairFace embeddings. Right: LFWA embeddings with data point sizes $\propto \phi_i(\text{EOp})$.



(b) Distribution of $\phi_i(\text{EOp})$ values.

Figure 11. `FairShap` applied to image embeddings. (a) The points with the largest $\phi_i(\text{EOp})$ (highlighted in green), all correspond to images labeled as "female" near the decision boundary of the original model (a-left). As a result of the data re-weighting, the decision boundary has been shifted, yielding a fairer model. (b) Images with the largest $\phi_i(\text{EOp})$ and histogram of $\phi_i(\text{EOp})$ on the LFWA dataset.

Training Set	Acc ↑	TPR_W | TPR_M	EOp ↓
FairFace	0.909	0.906 | 0.913	0.007
CelebA	0.759	0.580 | 0.918	0.34
LFWA	0.772	0.635 | 0.896	0.26
$\boldsymbol{\phi}(\text{Acc})$	0.793	0.742 | 0.839	0.09
`FairShap` - $\boldsymbol{\phi}(\text{EOp})$	**0.799**	0.782 | 0.813	**0.03**

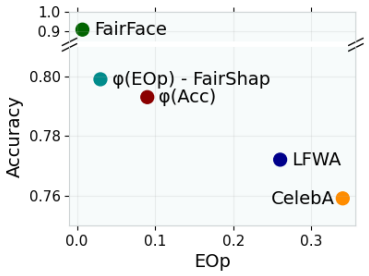


Table 4. Performance of the In-ResNet-V1 model tested on the FairFace dataset without and with re-weighting, using binary protected attribute $A = Y =$ gender. The arrows next to the metrics indicate whether the optimal result of the metric is 0 (↓) or 1 (↑). Best results are bold.

and thus we report results using $\phi_i(\text{EOp})$: $\phi_i(\text{EOp})$ quantifies the contribution of the ith data point (image) in LFWA to the fairness metric (Equal Opportunity) of the model tested on the FairFace dataset.

Baselines. The purpose of this experiment is to illustrate the versatility of `FairShap` rather than to perform an exhaustive comparison with other methods, as previously done with tabular data. Nonetheless, we compare `FairShap` with three baselines: the pre-trained model using CelebA; the fine-tuned model using LFWA without re-weighting; and a data re-weighting approach using $\boldsymbol{\phi}(\text{Acc})$ from [187]. We report two performance metrics: the accuracy of the models in correctly classifying the sex in the images (Acc) and the Equal Opportunity (EOp), measured as $\text{TPR}_M - \text{TPR}_W$ where W is the disadvantaged group (females in this case). We also report the specific TPR for males and females. A summary of the experimental setup for this scenario is depicted in Figure 7b.

Results. Note how both re-weighting approaches ($\boldsymbol{\phi}(\text{Acc})$ and `FairShap`) significantly improve the fairness metrics while *increasing the accuracy* of the model. `FairShap` yields the best results **both in fairness and accuracy**. Regarding EOp, the model trained with data re-weighted according to `FairShap` yields improvements of **88%** and **66%** when compared to the model trained without re-weighting (LFWA) and the model trained with

weights according to $\boldsymbol{\phi}(\text{Acc})$, respectively. In sum, data re-weighting with `FairShap` is able to leverage complex models trained on biased datasets and improve both their fairness and accuracy.

To gain a better understanding of the behavior of data re-weighting by means of `FairShap` in this scenario, Figure 11b (Bottom) depicts a histogram of the $\boldsymbol{\phi}(\text{EOp})$ values on the LFWA training dataset. As seen in the Figure, $\boldsymbol{\phi_i}(\text{EOp})$ are mostly positive for the examples labeled as *female* and mostly zero or negative for the examples labeled as *male*. This result makes intuitive sense given that the original model is biased against females, *i.e.*, the probability of misclassification is significantly higher for the images labeled as female than for those labeled as male. Figure 11b (top) depicts the five images with the largest $\phi_i(\text{EOp})$: they all belong to the female category and depict faces with a variety of poses, different facial expressions, and from diverse races.

Note that in this case `FairShap` behaves like a distribution shift method. Figure 11a shows how $\phi_i(\text{EOp})$ shifts the distribution of $\mathcal{D}$ (LFWA) to be as similar as possible to the distribution of the reference dataset $\mathcal{T}$ (FairFace). Therefore, biased datasets (such as $\mathcal{D}$) may be debiased by re-weighting their data according to $\phi_i(\text{EOp})$, yielding models with competitive performance both in terms of accuracy and fairness. Figure 11a illustrates how the group fairness metrics impact individual data points: critical data points are those near the decision boundary. This finding is consistent with recent work that has proposed using Shapley Values to identify counterfactual samples [11].

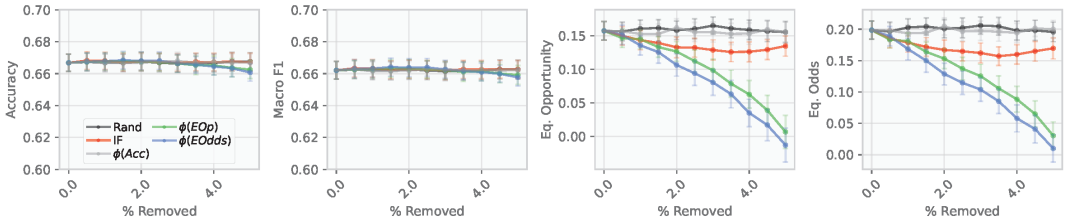
2.6.3 Data Pruning or Point Removal Experiment

Data valuation functions can also be used to guide policies for data selection. In this section, we evaluate the effectiveness of `FairShap` in data pruning (or data point removal as per [187, 253]). The goal is to identify and remove data points in the training set with the lowest valuations (negative influence) such that the model trained with the resulting data would maintain good levels of accuracy while being fairer than a model trained on the entire dataset.

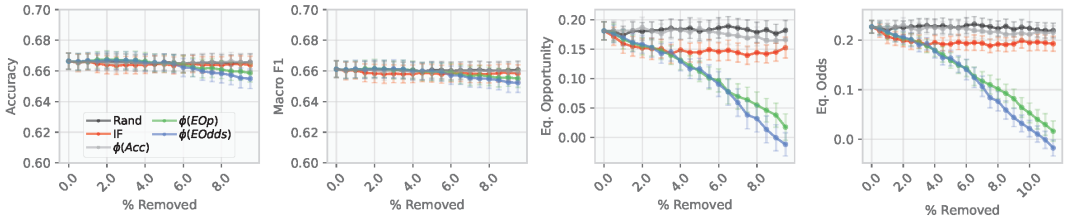
To this end, we first computed the value of each training point, ranked them in increasing order of value according to each valuation method, and iteratively removed data points in that order. At each step, we discarded 0.5% of the points, continuing this process until a total of 15% had been removed. We performed 50 random training ($\mathcal{D}$), validation ($\mathcal{T}$), and test splits for each method and dataset, reporting mean values and standard deviations.

We compared five different data valuation approaches: *Rand*, which randomly removes data points; the same *IF*-based method used as a baseline in the previously presented experiments; $\phi(\text{Acc})$, and the two `FairShap` valuations: $\phi(\text{EOp})$ and $\phi(\text{EOdds})$. A total of 75,000 models were trained for this experiment: 30 models for each pruning experiment, repeated over 50 random splits, for each of the five methods and the five datasets with two protected attributes each.

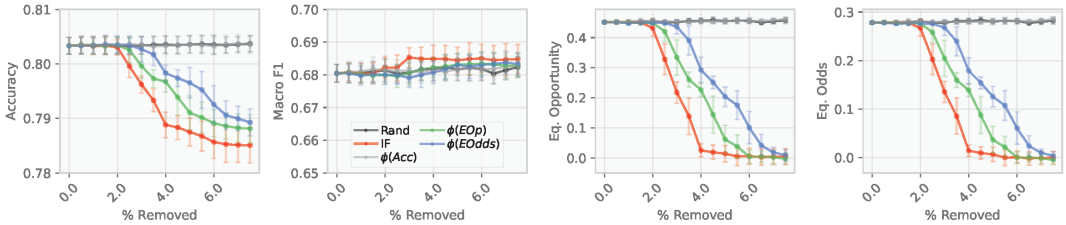
In this experiment, we included the ACSIncome dataset in addition to the COMPAS, German and Adult datasets, focusing on the two U.S. states identified by Ding et al. [130, Figure 2] that exhibit the largest levels of fairness violations: Hawaii (HI) and Alabama (AL). In both cases, the protected attribute, A, was Race with Black being the disadvantaged group. As seen in Figures 12 and 13, data removal based on `FairShap` significantly improves the fairness of the resulting classifier, even when removing a small portion of the training data: around 5% in COMPAS, 10% in German, and 15% in the ACSIncome datasets (for both the HI and AL states).



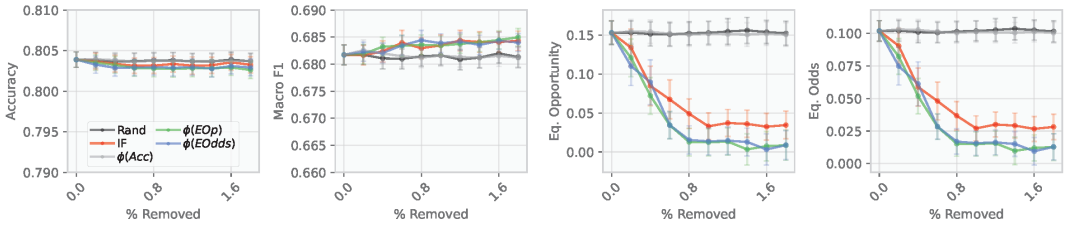
(a) COMPAS - Sex



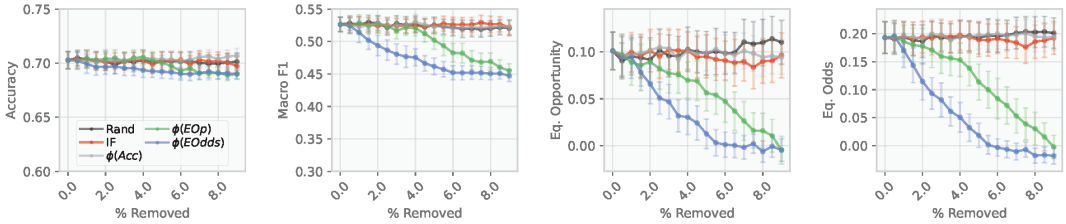
(b) COMPAS - Race



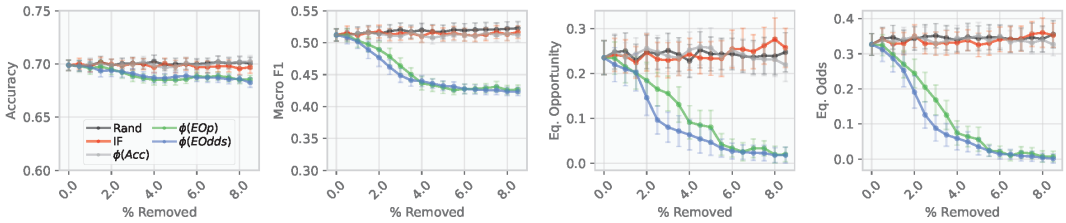
(c) Adult - Sex



(d) Adult - Race

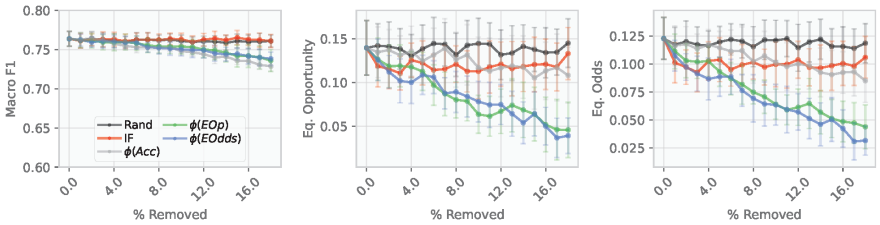


(e) German - Sex

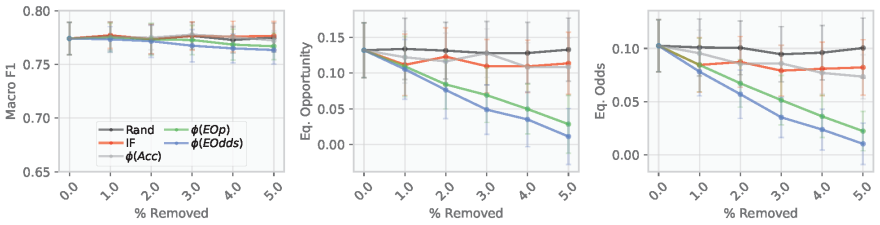


(f) German - Age

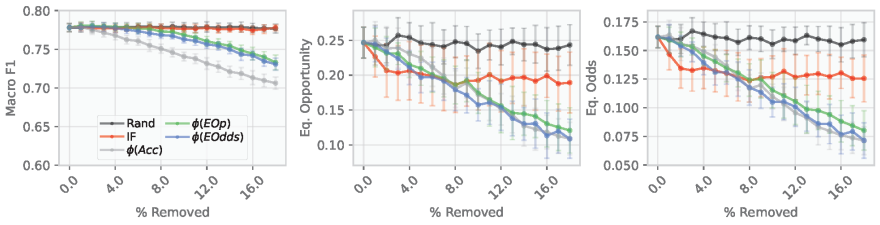
Figure 12. Performance and fairness metrics as a function of the percentage of data removed for COMPAS, Adult, and German. Subfigures depict different datasets and protected attributes.



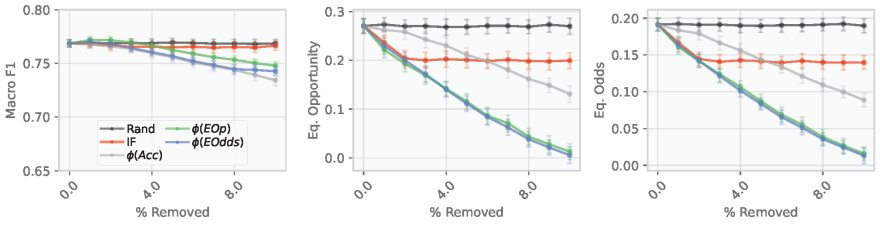
(a) ACSIncome (HI) – A=Race



(b) ACSIncome (HI) – A=Sex



(c) ACSIncome (AL) – A=Race



(d) ACSIncome (AL) – A=Sex

Figure 13. Performance and fairness metrics as a function of the percentage of data removed for ACSIncome (HI and AL).

2.6.4 Impact of the Reference Dataset's Size

In this section, we examine the influence of the size of the reference dataset, $\mathcal{T}$, and the impact of the alignment between $\mathcal{T}$ and the test set on the effectiveness of `FairShap`'s re-weighting. To do so, we perform an ablation study. We partition the three benchmark datasets (German, Adult, and COMPAS) into training (60%, $\mathcal{D}$), validation (20%), and testing (20%). We select subsets from the validation dataset –ranging from 5% to 100% of its size– and use them as $\mathcal{T}$. For each subset, we compute `FairShap`'s weights on $\mathcal{D}$ with respect to $\mathcal{T}$, train a Gradient Boosting Classifier (GBC) model, and evaluate its performance on the test set. This process is repeated 10 times with reported results comprising both mean values and standard deviations shown in Figure 14.

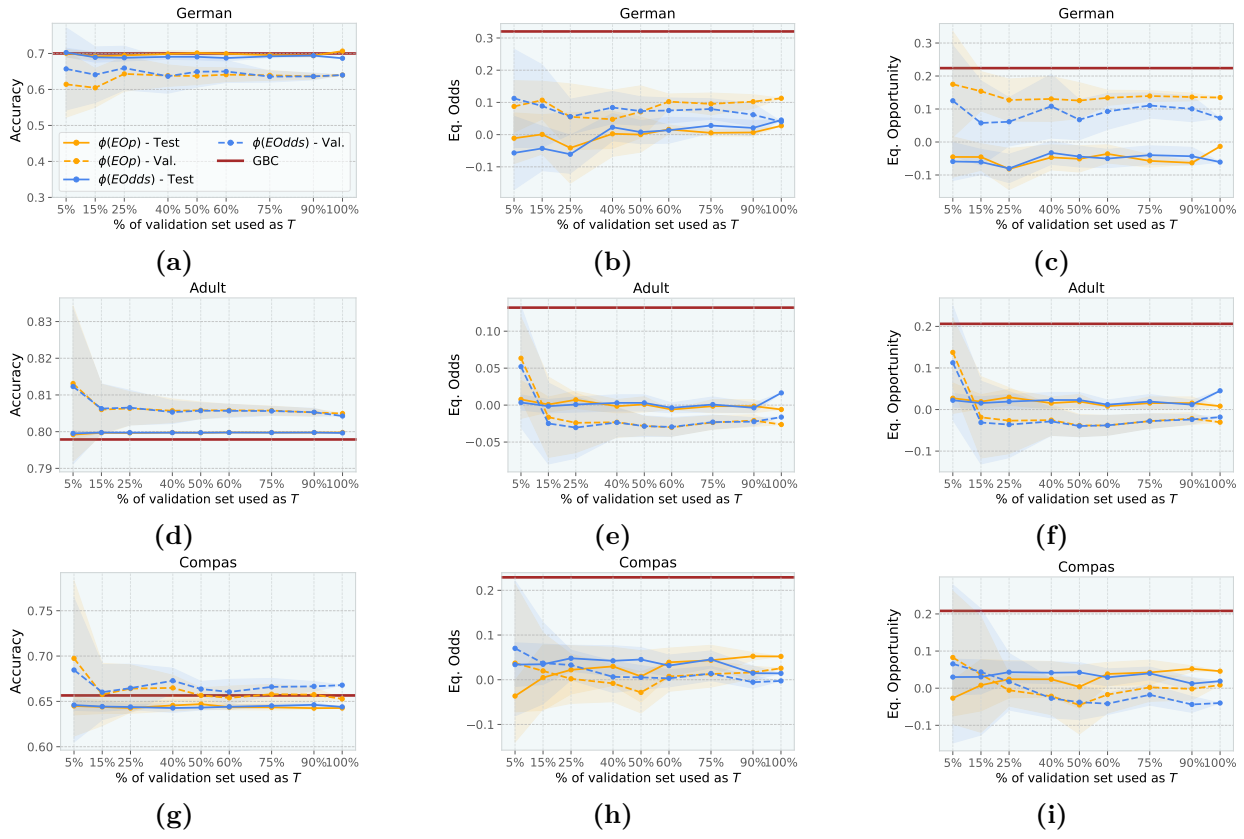


Figure 14. Accuracy and fairness metrics when applying data re-weighting via `FairShap` (ϕ(EOp) and ϕ(EOdds)) as the size of the validation sets $\mathcal{T}$ increases, evaluated on both validation (- -) and test sets (—). The performance of the baseline GBC without re-weighting is shown as a red line. From top to bottom, the rows correspond to the German, Adult, and COMPAS datasets, respectively. From left to right, the columns depict the Accuracy, Equalized Odds, and Equal Opportunity, respectively.

As shown in the Figures, the size of the validation dataset has a discernible impact on the variance of the evaluation metrics, both in terms of accuracy and fairness. Increasing the size of the reference dataset T leads to a notable reduction in the variability of the outcomes. However, averages for all the metrics remain stable across different sizes.

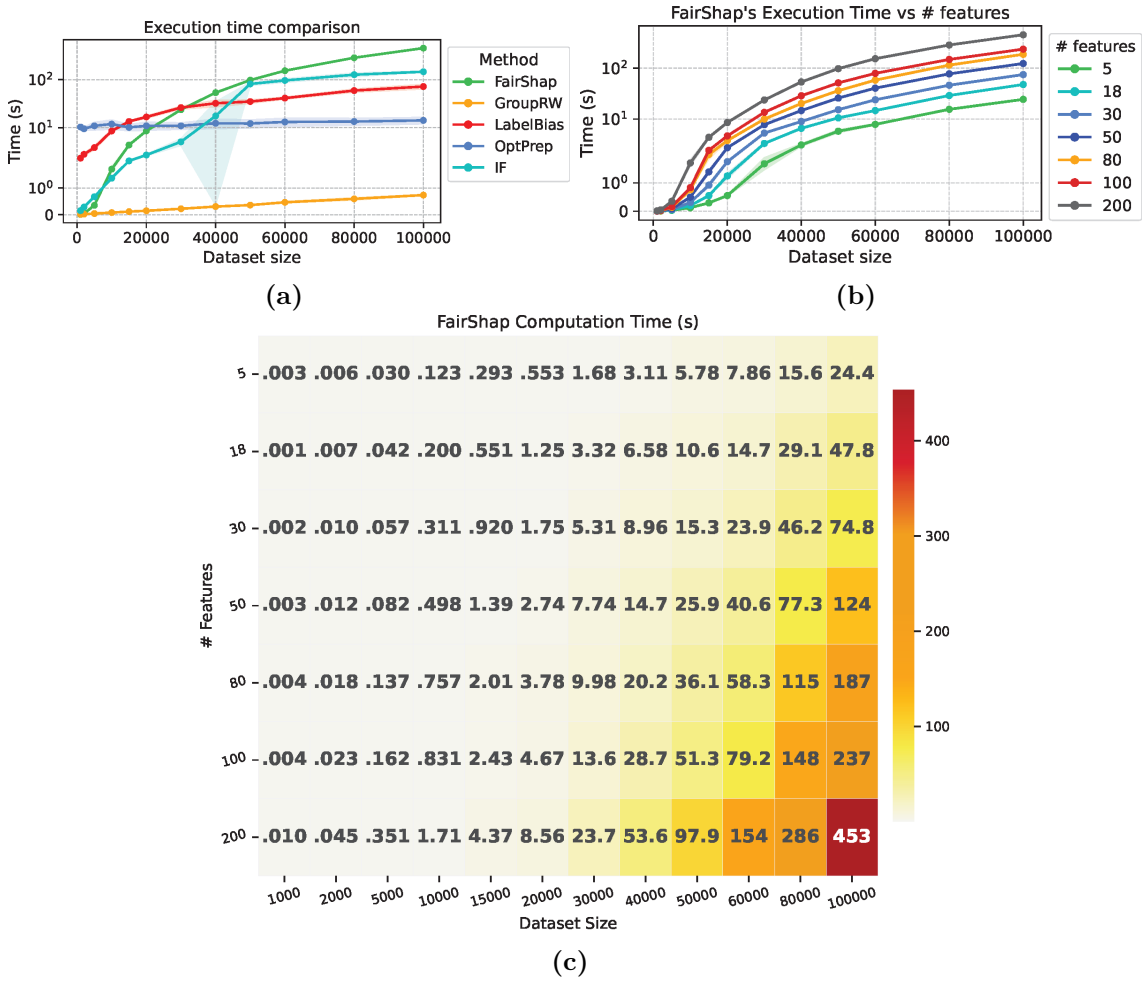


Figure 15. Run time comparison of re-weighting via `FairShap` and baselines with respect to data set size (a) and number of features (b). Datasets are split in 80% as D and 20% as T. We report mean and std run times for 10 iterations. (c) `FairShap`'s execution times (in seconds) on datasets with increasing numbers of features and sizes. In all experiments, the CPU is an Intel i7-1185G7-3.00GHz.

2.6.5 Computational Cost

We evaluate `FairShap`'s computational cost by applying data re-weighting on synthetic datasets of varying sizes (1k to 100k samples, each with 200 features). We compare the run time (in seconds) of `FairShap` with Group Re-weighting [231], OptPrep [90], LabelBias [225], and IFs [263]. Post-processing [206] is excluded, as its run time depends on the downstream model. For each configuration, we use an 80/20 train/validation split and report the mean and standard deviation over 10 runs on an Intel i7-1185G7 3.00GHz CPU. Results are shown in Figure 15.

As seen in the Figure, instance-level re-weighting via `FairShap` is computationally competitive for datasets with up to 30k data points. Group Re-weighting and LabelBias are computationally more efficient than `FairShap` on datasets with >30k data points.

Note that OptPrep [90] and IFs [263] require a hyperparameter search for each model and each dataset, yielding a significant increase on the computation time. Consequently, the actual running time for these methods would significantly increase depending on the number

of hyperparameter configurations to be tested. For example, OptPrep consistently requires $\approx$ 10 seconds regardless of the dataset's size. However, a hyperparameter grid-search scenario with 20 different hyperparameter settings and 10-fold cross-validation would increase the run time to 2,000 seconds (*i.e.*, 10s/it x 20 x 10) or 20,000s for IFs on a dataset size of 60,000 samples (*i.e.*, 100s/it x 20 x 10). These run times are significantly larger than those required to compute `FairShap`'s weights.

Finally, Figure 15 (c) also depicts `FairShap`'s execution times (in seconds) with different numbers of features in datasets of increasing sizes.

2.7 Conclusion, Discussion and Future Work

In this chapter, we have proposed `FairShap`, a novel instance-level, model-agnostic data valuation approach designed to achieve group fairness using Shapley Values. `FairShap` has been empirically validated and used to both re-weight and prune datasets across various tabular and vision datasets, demonstrating its effectiveness in different scenarios with two different types of models: GBCs and DNNs. In this chapter, we have proposed `FairShap`, a novel instance-level, model-agnostic data valuation approach designed to achieve group fairness using Shapley Values. `FairShap` has been empirically validated for use to both re-weight and prune datasets across various tabular and vision datasets, demonstrating its effectiveness in different scenarios with two different types of models: GBCs and DNNs. The experimental results reveal that models trained with data re-weighted via `FairShap` achieve competitive accuracy while delivering superior fairness outcomes compared to baseline methods. In addition, we showcased the potential of the proposed method as a data pruning policy. Furthermore, we demonstrated `FairShap`'s interpretability through histograms and latent space visualizations, providing insights into how data re-weighting impacts model decisions. Additionally, we studied the accuracy vs fairness trade-off, finding that `FairShap` achieves a favorable balance. We also explored the impact of the size of the reference dataset and assessed the computational cost of `FairShap` relative to baseline approaches. These findings demonstrate that `FairShap` is a practical and effective tool for achieving algorithmic fairness, making it suitable for real-world applications.

While data re-weighting via `FairShap` delivers strong empirical performance with various models, datasets, and tasks, developing a theoretically sound and efficient approximation to compute the Shapley Values beyond the existing approximation is an important area for future work.

From a practical perspective, the proposed fair data valuation functions align well with the interpretability criteria emphasized by legal stakeholders and emerging regulatory frameworks, ensuring that fairness interventions are not only practical but also transparent.

Chapter 3

Structural Group Unfairness: Measurement and Mitigation by means of the Effective Resistance

Chapter summary and context

In this chapter, we focus on the **technical** sphere and extend fairness analysis to graph-based systems, where harms emerge not from isolated decisions but from structural disparities in information access and visibility. We define and formalize `Structural Group Unfairness`, a class of graph-based disparities that affect marginalized groups. We propose practical measurement and mitigation methods grounded in the graph-theory, aligning our approach with systemic risk obligations under the European Digital Services Act.
This chapter is based on the following publication:

> [19] Adrian Arnaiz-Rodriguez, Georgina Curto Rex, and Nuria Oliver. "Structural Group Unfairness: Measurement and Mitigation by Means of the Effective Resistance". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 19. 1. Also presented at IC2S2 2024 and TrustLOG @ WWW 2024. June 2025, pp. 83–106. DOI: 10.1609/icwsm.v19i1.35805. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/35805

and draws theoretical foundations from [20, 22–24], included in Appendices D to F in Part V.

3.1 Introduction

Online social networks play an important role in defining and sustaining the social fabric of human communities. They allow individuals to connect, interact and share information with one another over the internet. They have opened up new opportunities for personal and professional networking, entertainment and learning. However, the formation of social networks —whether through organic growth or recommendations— can create imbalances in network positions which condition the access to resources and information [85]. These network inequalities have an impact on the social capital of its members, which exists in the

relations among individuals [111]. Better positioned network members benefit from faster access to diverse information, higher influence on information dissemination and more control of the information flow [86, 196, 199]. In practical terms, this means that individuals with a strategic position in the network will have more influence over others, and better access to information and opportunities regarding jobs, health, education or finance.

Furthermore, link recommendation algorithms that pervade social media platforms tend to connect similar users, contributing to the homophily and clustering of the network [369]. These *filter bubbles* limit the access to diverse individuals [196], exacerbate the isolation and polarization of groups, reduce the opportunities of innovation and aggravate the perpetuation of societal stereotypes [183]. In sum, the topology of the network can lead to a vicious cycle where those who are disadvantaged accumulate fewer opportunities to improve their social capital [171].

A variety of graph intervention methods have been proposed in the literature to mitigate disparities in social capital at an individual level [44]. However, there is a lack of methods that consider such disparities at a group level, which is particularly relevant when the groups correspond to socially vulnerable groups, *i.e.*, those defined on the grounds of sex, race, color, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other [389]. Focusing on the group level also supports the development of inclusive solutions at scale that benefit entire communities, promoting equity, diversity and the inclusion of disadvantaged groups. We denote the disparity in social capital among different groups in the network as *structural group unfairness*.

We consider a setting where each node in the network is a source of unique information and, therefore, access to all nodes is equally important. In this context, information flow is an integral component of the social capital and a distance metric that quantifies the total information flow in the graph, considering high-order relations that expand beyond the immediate neighbors, is of utmost importance. We propose using the effective resistance to measure the overall information flow between pairs of nodes (and thus the social capital), since it is a theoretically grounded continuous graph diffusion metric that considers both *local* and *global* properties of the network's topology [108].

This chapter is organized as follows. In Section 3.3.2, we introduce three measures of group social capital —*group isolation*, *group diameter* and *group control*— based on the effective resistance where the groups are defined according to the value of a protected attribute of interest. Using these measures of social capital, we define three measures of structural group unfairness in Section 3.3.3, and frame the challenge of mitigating structural group unfairness as a budgeted edge augmentation task in Section 3.3.4. This section also presents the Effective Resistance Group Link (`ERG-Link`) algorithm, a greedy edge augmentation algorithm that iteratively adds edges to the graph to increase the social capital of the most disadvantaged group. In experiments on real-world networks, described in Section 3.4, we uncover significant levels of structural group unfairness when using gender as the protected attribute, with females being the most disadvantaged group in comparison to males. We also illustrate how our approach is able to not only mitigate disparities in group social capital, but also increase the social capital of all the groups in the network.

3.2 Related Work

Social Capital Social capital is as a multidimensional construct that has been extensively studied in sociology, political science, economics, and more recently, computational social science [324]. It is defined as the networks, relationships, and norms of trust and reciprocity within a community or society that facilitate cooperation and collective action [111]. In simple terms, social capital is the value derived from connections between people. It can be measured and analyzed both at an individual and collective levels [76] and it has been characterized according to different criteria. Some authors propose three main dimensions of social capital, namely: structural, emphasizing the relationships among individuals, organizations and communities; cognitive, focusing on the shared values, norms and beliefs that bind members of a group or community; and relational, highlighting the intensity and quality of relationships, including reciprocity, trust and obligations among individuals [300]. Others have proposed the distinction between bonding, bridging, and linking social capital [374]. Bonding social capital captures the aspects of "inward looking" communities that reinforce exclusive identities and homogeneous groups [111]; bridging social capital refers to "outward looking" networks across different groups that do not necessarily share similar identities [85, 196]; and linking social capital characterizes the trusting relationships and norms of respect across power or authority gradients [403]. The three forms are important for the well-being of individuals and communities: bonding social capital contributes to social cohesion and support; bridging social capital to mutual understanding, solidarity and respect; and linking social capital to mobilize political resources and power.

Computational Models of Social Capital Network analysis offers a robust computational framework to examine and quantify social capital [111]. We consider a setting where all the nodes in the network may be sources of relevant information. As a consequence, access to all nodes —not just the sources or seeds of information— is equally important. In this context, *information flow* is an integral component of the social capital, and a variety of methods have been proposed to characterize it, mainly through two concepts: centrality and criticality [75].

Centrality measures the relative importance or prominence of a node in the network, quantifying its ability to reach the rest of nodes. Different approaches have been proposed in the literature to measure centrality, including the degree centrality, closeness, local clustering, the assortativity coefficient [304], Katz centrality [236] and PageRank [311]. Criticality reflects the node's level of influence or vulnerability within the network [380]. Nodes with high criticality are essential, such that their failure or disruption can have significant consequences, cascading effects or system-wide impact. Measures of criticality include effective size [85], redundancy [76] and shortest path betweenness [221].

However, previously proposed methods are insufficient to accurately quantify the overall information flow in the network for several reasons. First, they model the distance between nodes as the shortest path distance (geodesic distance [304]) which overlooks alternative routes and indirect connections that may exist between distant nodes, thereby underestimating the potential pathways for information diffusion, influence propagation [363] or resource exchange. This myopic view can lead to oversimplified representations of network dynamics, ignoring the interplay between weak ties, bridge nodes and overlapping communities that facilitate connectivity and communication across disparate components in the network.

Second, most of the proposed approaches only consider first-order —direct and local— relationships between nodes, relying on small neighborhoods of the graph. As a result, they ignore global structural information [234], such as the properties of the network topology and long-range interactions between nodes, which can lead to inaccurate insights on how information flows globally [328]. Third, popular approaches to model information flow in a network, such as the Independent Cascade model [240] assume homogeneous, deterministic and instantaneous interactions between neighboring nodes which might lead to inaccurate predictions, biased estimations and misrepresentations of actual diffusion patterns observed in complex networks [377].

Conversely, graph diffusion metrics, such as the *effective resistance* [246, 363], offer a principled approach to quantifying distances and interactions between nodes within a network, addressing the above limitations. The effective resistance accurately captures not only short-range but also long-range relationships between nodes [108] because it considers alternative pathways, including the network dynamics, and quantifies connectivity between distant nodes. Therefore, it constitutes a natural information distance metric between nodes in a graph [78, 363]. Previous work has theoretically formulated measures of node centrality and criticality based on effective resistances [78, 79, 303, 381], yet we are not aware of any work that has modeled the social capital of a group of nodes by means of the effective resistance. From a practical perspective, the concept of effective resistance has been used to measure polarization in social networks [211] and to rank user-items relations in recommender systems [174]. In this chapter, we propose quantifying the social capital of a group of nodes in the network by means of three measures derived from the effective resistance: the group isolation, group diameter and group control, explained in Section 3.3.2.

Fairness in Graphs Networks are used for a variety of purposes, including decision-making on nodes, link prediction, node embedding learning, clustering and community detection, ranking, and influence maximization. Fairness has been extensively analyzed in these scenarios [132, 341].

Regarding social capital, social status plays a role in defining the structure of a network [33] and a node's position in a network is a form of social capital [86]. Thus, there are structural advantages in information flow depending on the position that a node occupies in the network. Prior work has studied fairness from the perspective of disparities in access to information by differently positioned nodes in the graph, particularly in the case of influence maximization, *i.e.*, when a single piece of information is spread in the network [401]. However, there is a scarcity of studies that model group fairness considering that all the nodes in the network are sources of information and therefore access to all the nodes is equally important [44]. In this context and to the best of our knowledge, no previous research has considered fairness in graphs from a group perspective, when the groups are defined according to protected attributes —such as gender, ethnicity, religion or socio-economic status. In this chapter, we fill this gap by defining, measuring and mitigating *structural group unfairness*, understood as disparities in social capital between different groups and where social capital is measured by information flow.

Network Interventions to Mitigate Unfairness in Graphs Network interventions draw upon social network theory and structural analysis to understand and address the underlying mechanisms of unfairness within social networks. Interventions to mitigate struc-

tural unfairness in a network may entail redesigning network structures [211, 340] or altering (adding and/or removing) edges [382] to eliminate discriminatory barriers, reduce homophily, and foster diversity within the networks [196]. These interventions aim to enhance connectivity, promote inclusivity, and facilitate equitable access to resources, opportunities, and support networks [75].

When aiming to improve the social capital in a network, edge augmentation (*i.e.*, adding edges) is considered to be the natural intervention to mitigate disparities [44]. Several edge augmentation strategies have been proposed in the literature, such as connecting similar nodes to improve bonding social capital [420], linking nodes with the highest product of eigenvector centralities [382] or creating edges between the most disadvantaged nodes and the central node [44]. However, these strategies are defined for individual notions of social capital and they do not consider long-range interactions between nodes.

Contributions Given previous work, the main contributions of our work are:

i. We propose three effective resistance-based measures of group social capital in social networks —namely *group isolation*, *group diameter* and *group control*— that consider short- and long-range interactions between nodes.

ii. We define *structural group unfairness* as a disparity in the values of such measures by different groups in the graph, where the groups are defined according to the values of a protected attribute. This approach is particularly relevant from a social perspective when the disadvantaged group in the network corresponds to a vulnerable social group.

iii. We propose `ERG-Link`, an effective resistance-based greedy edge augmentation algorithm that iteratively adds edges to the network to maximize the social capital of the most disadvantaged *group*. We approach this objective by adding weak ties [196] between the disadvantaged group and the rest of the graph.

iv. In experiments on real-world networks, we uncover significant levels of group structural unfairness when using gender as a protected attribute, with females being the most disadvantaged group in comparison to males. We also illustrate how our approach is the most effective in reducing structural group unfairness when compared to the baselines.

3.3 Measuring Structural Group Unfairness

First, we define how to measure information flow between nodes in a social network, which forms the theoretical foundation for the proposed measures of group social capital. Next, we introduce three metrics to quantify a group's social capital within a graph and define structural group unfairness as the disparity in these metrics across different groups. Finally, we propose a greedy graph intervention (edge augmentation) algorithm designed to mitigate structural group unfairness, *i.e.*, disparities in group social capital.

3.3.1 Preliminaries

Effective Resistance and Social Capital

We focus on the structural dimension of social capital, which emphasizes the relationships among individuals or communities [300], and propose to measure it as the information flow of a node in the network. Such a measure is captured by the *effective resistance* [133, 246] of the node. Given nodes u and v in graph $G = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of nodes, $\mathcal{E} = \{(u, v) \in \mathcal{V} \times \mathcal{V} : A_{uv} = 1\}$ is the set of edges and $\mathbf{A}$ is the graph's adjacency matrix, the effective resistance R_{uv} between nodes u and v is a distance metric given by:

$$R_{uv} = (\mathbf{e}_u - \mathbf{e}_v)\mathbf{L}^\dagger(\mathbf{e}_u - \mathbf{e}_v)^\top, \tag{12}$$

where $\mathbf{e}_u$ is the unit vector with a unit value at u-th index and zero elsewhere; $\mathbf{L}^\dagger = \sum_{i>1} \frac{1}{\lambda_i}\phi_i\phi_i^\top$ is the pseudo-inverse of the graph's Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A} = \mathbf{\Phi\Lambda\Phi}^\top$, with $\mathbf{D}$ the graph's degree matrix, $D_{u,u} = \sum_{j \in V} A_{u,v}$ and 0 elsewhere; and λ_i the i-th smallest eigenvalue of $\mathbf{L}$ corresponding to the ϕ_i eigenvector. The complete matrix of all pairwise effective resistances in a graph, $\mathbf{R}$ is given by $\mathbf{R} = \mathbf{1}\,\mathrm{diag}(\mathbf{L}^\dagger)^\top + \mathrm{diag}(\mathbf{L}^\dagger)\mathbf{1}^\top - 2\mathbf{L}^\dagger$.

The effective resistance is a distance *metric* since it satisfies the symmetry, non-negativity and triangle inequality conditions [140]. In addition, R_{uv} is proportional to the commute times between u and v, *i.e.*, the expected number of steps in a random walk starting at v to reach node u and come back: $R_{uv} \propto \mathbb{E}_u[v] + \mathbb{E}_v[u]$, where $\mathbb{E}_u[v]$, $\mathbb{E}_v[u]$ are the expected number of steps that a random walker takes to go from u to v and from v to u, respectively [99, 378]. A high value of R_{uv} means that u and v generally struggle to visit each other in a random walk, *i.e.*, nodes with high effective resistance between them are unlikely to exchange information. R_{uv} can be expressed as

$$R_{uv} = \sum_{i=0}^{\infty} \left(\frac{1}{d_u}(\mathbf{A}^i)_{uu} + \frac{1}{d_v}(\mathbf{A}^i)_{vv} - \frac{1}{\sqrt{d_u d_v}} 2(\mathbf{A}^i)_{uu} \right),$$

being $\mathbf{A}^k$ the matrix that defines the number paths of length k between u and v [59]. Hence, it is able to capture both short- and long-range interactions between nodes in the graph.

The effective resistance has been characterized as the *information distance* in a network [78, 363] as it quantifies the amount of effort (distance) required to transmit information between the nodes. The total effective resistance $\mathsf{R}_{\mathsf{tot}}$ of a graph [140] – defined as the sum of all R_{uv} ($\mathsf{R}_{\mathsf{tot}} = \mathbf{1R1}^\top$) – is therefore inversely proportional to the expected ease of information flow in the graph.

The *total effective resistance of node* u, $\mathsf{R}_{\mathsf{tot}}(u)$, is given by $\mathsf{R}_{\mathsf{tot}}(u) = \sum_{v \in \mathcal{V}} R_{uv}$, *i.e.*, the sum of all the effective resistances between node u and the rest of nodes in the network. The smaller the total effective resistance of a node, the larger its information flow. In other words, the effective resistance allows to identify which nodes in a graph have limited information flow (*i.e.*, high effective resistance) and thus low social capital. Equivalent terms to denote the effective resistance in the literature include the current-flow closeness centrality [79] and the information centrality [363].

From a computational perspective, calculating R_{uv} does not require hyper-parameter tuning and can be efficiently calculated, mitigating two significant drawbacks of other diffusion or learnable graph distances [317]. An overview of the theoretical properties of R_{uv} are provided in Appendix B.1.

3.3.2 Effective Resistance-based Measures of Group Social Capital

Based on the definition of effective resistance above, we propose three metrics that characterize the social capital of a group of nodes in a graph. In the following, a graph $G = \{\mathcal{V}, \mathcal{E}\}$ is composed of a set of nodes $\mathcal{V}$ and edges $\mathcal{E}$; and S_i is a group of nodes defined as a subset of $\mathcal{V}$, *i.e.*, $S_i \subseteq \mathcal{V}$ with $|S_i|$ nodes.

1. Group Isolation The isolation of a group S_i, $\mathsf{R}_{\mathsf{tot}}(S_i)$, is given by the average of the total effective resistances of all the nodes in group. $\mathsf{R}_{\mathsf{tot}}(S_i)$ is proportional to the expected information distance when sampling one node from group S_i and another node at random. It can be interpreted as a proxy for the marginalization of a group from the perspective of information flow, such that the lower the $\mathsf{R}_{\mathsf{tot}}(S_i)$, the more the information flows and thus the less isolated the group S_i is in the network. Therefore, reducing this measure for group S_i would yield an increase in its social capital. It is given by:

$$\mathsf{R}_{\mathsf{tot}}(S_i) = \mathbb{E}_{u\sim S_i}[R_{tot}(u)] = |\mathcal{V}|\mathbb{E}_{u\sim S_i, v\sim\mathcal{V}}\left[R_{uv}\right] \tag{13}$$

where $\mathsf{R}_{\mathsf{tot}}(u) = \sum_{v\in\mathcal{V}} R_{uv}$ is the total effective resistance of node u. Computing the expectation enables comparing groups of different sizes.

Note that adding links between nodes with the highest R_{uv} —irrespective of which group they belong to— has been found to reduce the total effective resistance of a graph [59] and hence the isolation of all the graph's nodes.

2. Group Diameter The group diameter, $\mathcal{R}_{\text{diam}}(S_i)$, measures the average of the maximum distance between any node in group S_i and any node in the graph. A larger group diameter suggests that the nodes in group S_i are distant from the rest of the graph, indicating potential challenges in information exchange with the nodes outside of S_i, and hence it can be interpreted as another measure of social capital. Low diameter is important to ensure diverse information dissemination (*e.g.*, job announcements reaching a diverse pool of candidates) which depends on the entire network and not just on the average distance as *group isolation*. This measure is based on $\mathcal{R}_{\text{diam}}(G)$, which is the maximum effective resistance of the graph [99] and closely related to the cover time. We define *Group Diameter* as:

$$\mathcal{R}_{\text{diam}}(S_i) = \mathbb{E}_{u\sim S_i}[\mathcal{R}_{\text{diam}}(u)] = \mathbb{E}_{u\sim S_i}[\max_{v\in\mathcal{V}} R_{uv}] \tag{14}$$

where $\mathcal{R}_{\text{diam}}(u) = \max_{v\in\mathcal{V}} R_{uv}$ is the diameter of node u, *i.e.*, the maximum R_{uv} from u to any other node in the graph. $\mathcal{R}_{\text{diam}}(S_i)$ gives an indication of the information flow gap between the group S_i and the rest of the network [171]. Therefore, the larger the $\mathcal{R}_{\text{diam}}(S_i)$, the lower the social capital of group S_i.

3. Group Control The aforementioned concepts measure the amount of information flow through a group of nodes in the graph. Another relevant variable to assess is the *criticality* of a node for the diffusion of information in the graph, which in the literature has been measured as betweenness [177], redundancy [76] or effective size [85]. Nodes with high levels of control serve as important connectors in the network, facilitating the flow of information and enabling communication between otherwise disconnected groups of nodes [85].

The control of a node can be computed by restricting the summation of a node's total effective resistance to its direct neighbors. Thus, it is expressed as $\mathsf{B}_\mathsf{R}(u) = \sum_{v \in \mathcal{N}(u)} R_{uv}$, where $\mathcal{N}(u) = \{v : (u, v) \in \mathcal{E}\}$ are the neighbors of u. The larger the $\mathsf{B}_\mathsf{R}(u)$, the more control a node has in the network's information flow and hence the larger its social capital. The node control of a node is bounded by $1 \leq \mathsf{B}_\mathsf{R}(u) \leq d_u$, being d_u the number of neighbors of node u (see Theorem B.1). $\mathsf{B}_\mathsf{R}(u)$ is theoretically related to the current-flow betweenness [303], the node's information bottleneck [23], and the curvature of the node [125].

We define the group control or group betweenness $\mathsf{B}_\mathsf{R}(S_i)$ as the average of the controls of all the nodes in S_i, *i.e.*:

$$\mathsf{B}_\mathsf{R}(S_i) = \mathbb{E}_{u \sim S_i}[\mathsf{B}_\mathsf{R}(u)] = \frac{1}{|S_i|} \sum_{u \in S_i} \mathsf{B}_\mathsf{R}(u), \tag{15}$$

and is bounded by $1 \leq \mathsf{B}_\mathsf{R}(S_i) \leq \mathrm{vol}(S_i)/|S_i|$, being $\mathrm{vol}(S_i)/|S_i|$ the average degree of the group S_i (see Theorem B.2). Note that the sum of all R_{uv} for all nodes in a graph is constant at $|\mathcal{V}| - 1$ and it is independent of the number of edges [246]. If an edge is added, removed or modified in the graph, all R_{uv} are updated accordingly such that their sum remains constant. The sum and the average control of all nodes in the graph are also constant with values $\sum_{u \in \mathcal{V}} \mathsf{B}_\mathsf{R}(u) = 2|V| - 2$ and $\mathbb{E}_{u \sim \mathcal{V}}[\mathsf{B}_\mathsf{R}(u)] = 2 - \frac{2}{|\mathcal{V}|}$, respectively, independently of the number of edges (see Appendix B.2.1 for more details).

Consequently, the control of a node or group of nodes is distributed among the nodes in the network and cannot be optimized for every node/group in the graph by adding more edges: if a node or group of nodes increase their control over the information flow in the graph, they must do so at the cost of reducing the control of other nodes. Figure 16 illustrates the three proposed measures of individual and group social capital.

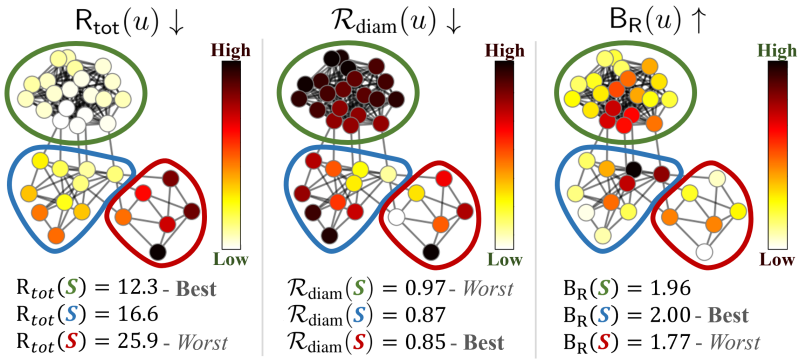


Figure 16. Illustration of the three proposed group social capital metrics on the same graph. The color of the nodes corresponds to $\mathsf{R}_{\mathsf{tot}}(u)$, $\mathcal{R}_{\mathrm{diam}}(u)$ and $\mathsf{B}_\mathsf{R}(u)$, respectively. The nodes are grouped according to three different values of the protected attribute S indicated as green, blue and red.

3.3.3 Structural Group Unfairness

To study disparities in the distribution of social capital in the network, we define groups of nodes in the network S_i according to the values of a sensitive attribute $i \in SA = \{sa_1, sa_2, \ldots, sa_{|SA|}\}$, which is a categorical variable with $|SA|$ possible values referring to a socially relevant concept, such as sex, age, gender, religion or race. We denote the value of

the sensitive attribute of a node v as $SA(v)$. For instance, if SA is sex with three possible values, SA={male, female, non-binary}, the groups S_{male}, S_{female} and $S_{\text{non-binary}}$ are the set of nodes whose sex is labeled as male, female and non-binary, respectively.

We define the *structural group unfairness* in a network as the disparity in information flow between the nodes belonging to different groups in the network. Since we have defined the groups in terms of protected attributes, the structural group unfairness is socially relevant as it informs about potential disparities in information flow (and hence social capital) between a vulnerable group and the rest of the network. We present here three metrics to characterize the structural group unfairness, namely *isolation disparity*, *diameter disparity* and *control disparity*.

1. Group Isolation Disparity Ideally, every group in the network should have the same levels of information flow and hence the same —and low— levels of group isolation, namely:

$$\mathsf{R}_{\mathsf{tot}}(S_i) = \mathsf{R}_{\mathsf{tot}}(S_j),\ \forall\, i, j \in SA. \tag{16}$$

Deviations from equality lead to isolation disparity $\Delta\,\mathsf{R}_{\mathsf{tot}}$, which is defined as the maximum over all groups in the graph of the differences in group isolation: $\Delta\,\mathsf{R}_{\mathsf{tot}} = \max_{i,j\in SA} |\,\mathsf{R}_{\mathsf{tot}}(S_i) - \mathsf{R}_{\mathsf{tot}}(S_j)|$.

Reducing the isolation disparity contributes to increasing the social capital of the most disadvantaged group and equalizes the information flow between the groups in the network.

2. Group Diameter Disparity Ideally, every social group in the network should have the same — and low — group diameter:

$$\mathcal{R}_{\text{diam}}(S_i) = \mathcal{R}_{\text{diam}}(S_j),\ \forall\, i, j \in SA. \tag{17}$$

Any deviations from equality lead to diameter disparity, $\Delta\,\mathcal{R}_{\text{diam}}$, defined as the maximum over all groups in the graph of the differences in group diameter: $\Delta\,\mathcal{R}_{\text{diam}} = \max_{i,j\in SA} |\,\mathcal{R}_{\text{diam}}(S_i) - \mathcal{R}_{\text{diam}}(S_j)|$.

Achieving equal diameter entails equalizing the worst-case scenario in information flowing to the entire network from the perspective of any group of nodes in the graph. By promoting equal group diameter, we generate a fairer information-sharing environment.

3. Group Control Disparity By striving for equalized control in all groups in the network, no particular group would dominate or be marginalized from the perspective of their control of information flow in the network:

$$\mathsf{B}_{\mathsf{R}}(S_i) = \mathsf{B}_{\mathsf{R}}(S_j) = 2 - \frac{2}{|\mathcal{V}|},\ \forall\, i, j \in SA. \tag{18}$$

Then, control disparity, $\Delta\,\mathsf{B}_{\mathsf{R}}$, is defined as the maximum over all groups in the graph of the differences in group control: $\Delta\,\mathsf{B}_{\mathsf{R}} = \max_{i,j\in SA} |\,\mathsf{B}_{\mathsf{R}}(S_i) - \mathsf{B}_{\mathsf{R}}(S_j)|$.

Control is a bounded resource to be distributed among the groups of nodes in the graph with an expected value of $2 - \frac{2}{|\mathcal{V}|}$. Hence, reducing the control disparity entails a redistribution of the control in all the groups in the graph converging to $\mathsf{B}_{\mathsf{R}}(S_i) = 2 - \frac{2}{|\mathcal{V}|}$, $\forall\, i \in SA$, and hence leading to a more equitable allocation of the control that different groups play regarding the information flow in the network.

3.3.4 Structural Group Unfairness Mitigation

Edge Augmentation Edge augmentation has been proposed as the natural intervention to mitigate information flow disparities in a network where all the nodes are sources of unique pieces of information [44]. According to Rayleigh's monotonicity principle [133], adding edges to a graph always improves information flow, which is the aim of our intervention. Socially, edge addition enhances the nodes' social capital since information reaches a larger audience. Moreover, note that edge deletion is not a suitable intervention as it could lead to a disruption of social dynamics by breaking existing connections between individuals, which is undesirable [221].

Regarding which structural group unfairness measure to optimize, we argue that we should primarily focus on improving the isolation disparity (ΔR_{tot}) of the most isolated group in the graph. Note that mitigating isolation will also yield an improvement in the diameter and control disparities, as illustrated in our experiments. The reduction of ΔR_{tot} entails creating edges between distant nodes, *i.e.*, fostering the creation of weak ties. Granoveter's work [196] provides evidence that information spreads more effectively through weak ties than through strong ties because weak ties give peripheral nodes more visibility in the network, which leads to a decrease in group isolation and diameter. Adding weak ties reduces discontinuities in the information flow, increases redundancies in the paths between nodes and improves the control of peripheral nodes while reducing the control of dominant ones [85].

Previous work has proposed connecting peripheral isolated nodes (with high isolation and low control) to salient nodes (with low isolation and high control) [382]. However, these solutions lead to a *rich-get-richer* phenomenon that benefits the best connected nodes and potentially increases disparities in information access and control [87]. Therefore, we advocate creating edges between the most distant nodes in the network without necessarily connecting them to a central node. Note that adding edges between the nodes with maximum R_{uv} theoretically leads to a minimization of $\mathsf{R}_{\mathsf{tot}}$ and $\mathcal{R}_{\text{diam}}$ for all nodes in the graph while balancing B_{R}. Hence, the choice of $R_{tot}(S)$.

Problem Definition We consider a budgeted edge augmentation intervention: given a maximum number B of allowed new connections to be created in the graph, we aim to identify the B new edges $\mathcal{E}'$ to be added to the graph G that would maximally reduce the group isolation disparity of the most disadvantaged group in the graph. This leads to a new graph G' with lower levels of structural group unfairness:

$$G' = \underset{G'=(\mathcal{V},\mathcal{E}')}{\arg\min} \quad \mathbb{E}_{u,v\sim V\times V}\left[R_{uv}\right] \quad \text{s.t. } |\mathcal{E}' \setminus \mathcal{E}| = B \text{ and } \mathcal{E} \subset \mathcal{E}' \tag{19}$$

Algorithm To tackle the problem above, we introduce `ERG-Link`, a greedy algorithm that adds edges between the nodes with the largest effective resistance between them, where at least one of the nodes belongs to the most isolated group as per Section 3.3.2, and groups in the graph are defined on the grounds of a protected attribute. Note that this strategy also reduces the isolation (total effective resistance) of the entire graph [188].

Algorithm 2 outlines the main steps of `ERG-Link`. Given a graph $G = (V, E)$, a protected attribute S and a total budget B of new edges to add, the group isolation $R_{tot}(S)$ is computed for each group according to S. The most disadvantaged group S_d is identified as the group with the largest $R_{tot}(S)$. Then, R_{uv} $\forall(u, v) \in \mathcal{V} \times \mathcal{V}$ is computed, and a ranking of all

Algorithm 2: `ERG-Link`

Data: Graph $G = (\mathcal{V}, \mathcal{E})$, a protected attribute SA, budget B of total number of edges to add

Result: New Graph $G' = (\mathcal{V}', \mathcal{E}')$ with B new edges

$\mathbf{L} = \mathbf{D} - \mathbf{A}$;

$S_d = \text{argmax}_{S_i \forall i \in SA} \mathsf{R}_{\text{tot}}(S_i)$

Repeat

1 $\quad \mathbf{L}^\dagger = \sum_{i>0} \frac{1}{\lambda_i} \phi_i \phi_i^\top = \left(\mathbf{L} + \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)^{-1} - \frac{\mathbf{1}\mathbf{1}^\top}{n}$

$\quad \mathbf{R} = \mathbf{1}\,\text{diag}(\mathbf{L}^\dagger)^\top + \text{diag}(\mathbf{L}^\dagger)\mathbf{1}^\top - 2\mathbf{L}^\dagger$

$\quad C = \{(u, v) \mid u \in S_d \text{ or } v \in S_d, (u, v) \notin \mathcal{E}'\}$

$\quad \mathcal{E}' = \mathcal{E}' \cup \arg\max_{(u,v) \in C} R_{uv}$

$\quad \mathbf{L} = \mathbf{L} + (\mathbf{e}_u - \mathbf{e}_v)(\mathbf{e}_u - \mathbf{e}_v)^\top$

Until $|\mathcal{E}' \setminus \mathcal{E}| = B$;

return G';

potential new edges in the graph is created from the highest to the lowest values of effective resistance. In each iteration, `ERG-Link` adds the new edge to the graph that yields the largest improvement in the information flow of S_d, *i.e.*, the edge that connects the two nodes with largest effective resistance between them where at least one of the nodes belongs to the disadvantaged group S_d. See Black et al. [59] and Ghosh, Boyd, and Saberi [188] for a proof that such an edge is the one that maximally improves the information flow in the graph.

`ERG-Link` leverages *Rayleigh's monotonicity principle* [133, 140], according to which the total effective resistance of a graph can only decrease when new edges are added to it, as illustrated in Figure 17. Therefore, adding an edge between nodes with maximum R_{uv} where one of the nodes belongs to the disadvantaged group not only improves the information flow between the two nodes (increasing their social capital) but it also improves the information flow of the entire graph.

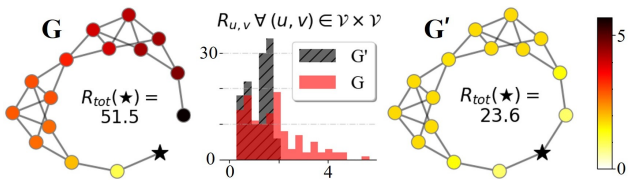


Figure 17. Impact of adding one edge on the information flow of G: all the effective resistances between the star node and the rest of nodes in the network ($R_{\star,v}$) decrease in G' even if there is no change in the geodesic distance between them.

Note that adding each new edge changes all the effective pairwise resistances between nodes in the graph, meaning they must be recomputed at each iteration. Therefore, it is not feasible to perform this type of edge augmentation by means of Independent Cascade distance estimation [44, 240], random-walk embeddings [317] or GNNs [404] since these methods require training expensive neural networks or running complex simulations for the estimation of the distances in each iteration. In addition, they only capture short-range interactions between nodes. Conversely, the effective resistance captures both short and long-range interactions between nodes in the graph and it more efficient to update. While it requires the computation of $\mathbf{L}^\dagger$, Woodbury's formula [59] can be used to avoid recomputing $\mathbf{L}^\dagger$ in line 3 of Algorithm 2, as reflected in Algorithm 4. For illustrative purposes, Appendix B.3.3 contains examples of the use of `ERG-Link` in synthetic graphs.

3.4 Experiments

3.4.1 Datasets and Set-up

To empirically evaluate `ERG-Link`, we tackle the challenge of mitigating group social capital disparities in three real-world networks (school and online social networks), where the nodes are users and the edges correspond to connections between them, *i.e.*, friendships.[10] The three datasets are commonly used in the graph fairness literature, namely:
(1) The Facebook dataset [259], a dense graph of 1,034 Facebook users ($|\mathcal{V}|$) and 26,749 edges ($|\mathcal{E}|$). It corresponds to a large ego-network where nodes are connected if they are friends in the social network;
(2) The UNC28 dataset [332], consisting of a 2005 snapshot from the Facebook network of the university of North Carolina ($|\mathcal{V}|$=3985,$|\mathcal{E}|$=65287);
(3) The Google+ dataset [259], an ego-network of G+, the social network developed by Google, with 3,508 nodes ($|\mathcal{V}|$) and 253,930 edges ($|\mathcal{E}|$).

Gender is the protected attribute in all networks with two possible values SA ={male, female}. We select the largest connected component for all the datasets. The original values of group social capital per gender are depicted in Table 5. As seen in the Table, our study unveils that the disadvantaged group according to the three defined measures corresponds to females in the three datasets.

G	$R_{tot} \downarrow$	$\mathcal{R}_{diam} \downarrow$	$\mathsf{B_R} \uparrow$
Facebook (female)	221.4	2.29	1.93
Facebook (male)	**179.8**	**2.25**	**2.03**
UNC28 (female)	608.6	2.11	1.99
UNC28 (male)	**586.3**	2.11	**2.00**
Google+ (female)	564.1	1.31	1.81
Google+ (male)	**287.7**	**1.24**	**2.32**

Table 5. Group social capital metrics in the original graphs for each of the groups.

In addition, understanding the graph topology of each social network is crucial for identifying bias and assessing the effectiveness of mitigation strategies. The following paragraphs provide a detailed breakdown of the statistics and characteristics of the datasets.

Number of male and female nodes Table 6 shows a breakdown of the number of nodes, edges, graph density, and number of males and females in each of the studied datasets. The Facebook and UNC28 datasets have a majority of males, whereas the Google+ dataset has a majority of females. Interestingly, the Google+ dataset has the highest levels of disparity in group social capital (see Table 5), which means that, although there are fewer males in the network, they are better positioned than females with respect to their access to information and thus social capital.

Edge homophily We also report the edge homophily [429] for the three datasets. Edge homophily, denoted as $h_{edge} \in [0, 1]$, represents the percentage of edges connecting nodes

[10] The code is publicly available at `https://github.com/ellisalicante/StructuralGroupUnfairness`.

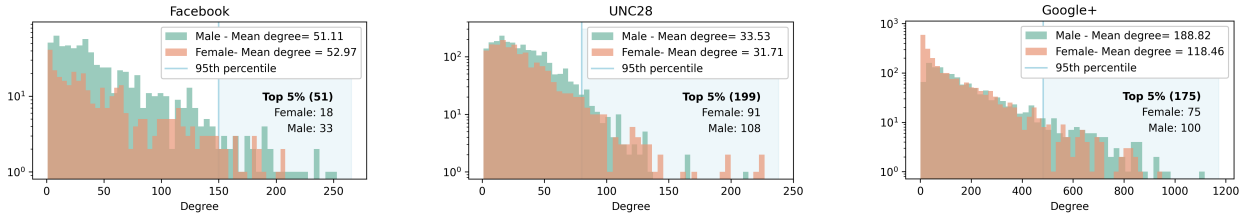


Figure 18. Original degree distribution per dataset and gender. We also include the average degree per group and the composition of the top 5% of nodes.

with the same protected attribute. The larger the group homophily, the more internally connected the groups are, with fewer connections between different groups. The obtained values of edge homophily suggest that the groups are well interconnected, and thus, there is no significant community polarization between males and females. In consequence, the differences between groups in access and control of information flow, as shown in Table 5, stem from disparities in roles and network positions, rather than from a polarized social structure.

Degree Distribution Figure 18 shows the degree distribution for the three different datasets grouped by protected attribute. As seen in the Figure, the number of high degree nodes is larger for males than females in the Facebook and Google+ datasets. Note that although there are fewer males than females in the Google+ dataset, the majority of nodes with high degree correspond to males, which explains their larger group social capital.

Centrality Metrics Figure 19 depicts the per-group closeness and betweenness centrality metrics [38] for the three datasets, providing an insight into the structural properties of the networks. However, these classical metrics are based on geodesic distances and are therefore limited in their ability to capture the entire network topology. For example, the geodesic distance is a poor estimator of long-range connections. In contrast, the proposed metrics, which are built on random walks and consider the entire network, offer a more complete view of information flow within the network. Nevertheless, even with classical metrics, we observe notable differences between groups, particularly in the Google+ dataset, where males exhibit larger closeness centrality despite being a minority.

Dataset	Facebook	UNC28	Google+
$\|V\|$	1,034	3,985	3,508
$\|\mathcal{E}\|$	26,749	65,287	253,930
Density	0.05	0.008	0.04
# Males	685	2307	1312
# Females	349	1678	2196
h_{edge}	0.58	0.55	0.51

Table 6. Dataset Statistics

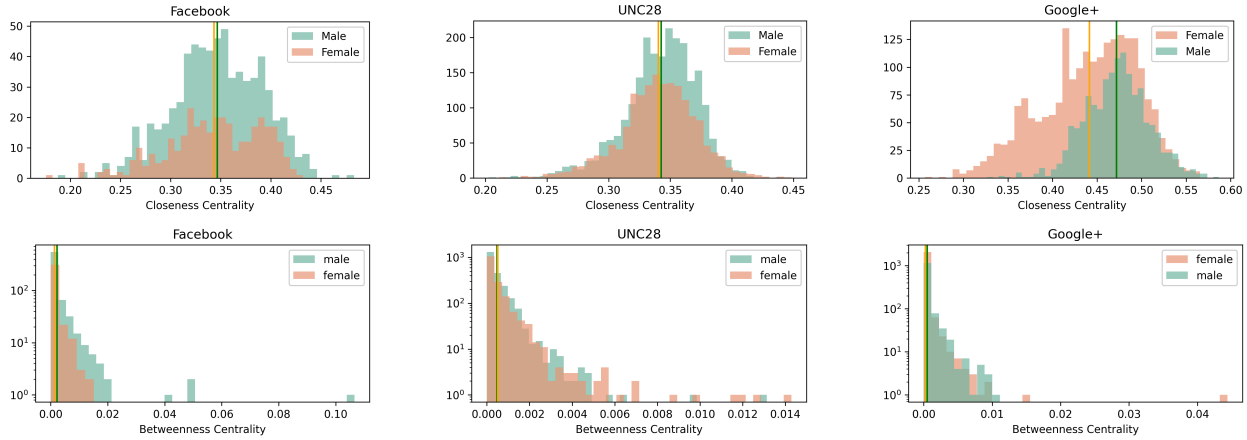


Figure 19. Per-group centrality metrics in the original datasets.

3.4.2 Baselines

We compare edge augmentation by means of `ERG-Link` with five baselines:
(1) *Random*, which adds edges at random to the graph.
(2) *DW*, which adds edges with the lowest dot product similarity of *DeepWalk* embeddings [317].
(3) *Cos*, a greedy algorithm that adds edges with the lowest cosine similarity of the rows of the adjacency matrix [341]. This is an example of a classic method based on neighborhood similarity.
(4) *SDRF*, an edge augmentation method originally proposed to mitigate over-squashing [383]. It identifies the edge with the minimum Ricci curvature and adds the edges that maximally improve the Ricci curvature of that edge. The Ricci curvature of an edge is related to R_{uv} and $\mathsf{B}_{\mathsf{R}}(u)$ as per Equation (46).
(5) *FOSR*, a graph rewiring algorithm to increase the spectral gap (λ_2) and hence avoid over-squashing [234].

Note that *DW* and *Cos* correspond to an algorithm similar to Algorithm 2 (lines 2 and 6 remain the same) with one difference: instead of using the effective resistances to quantify the distances between nodes, they consider DW or cosine distances, respectively. As previously noted, we do not include any Independent Cascade distance estimation, random-walk embedding or GNN-based method as baselines because they require training a neural network or a simulation to estimate the pairwise distances in the graph, which is computationally unfeasible in our task as it would entail retraining the neural network every time a new edge is added [404].

3.4.3 Experimental Methodology

We set a budget B of 5,000 new edges to be added to the UNC28 and Google+ datasets, which corresponds to approximately 0.05% of the number of all potential edges in the graph. We also run experiments with a maximum of 50 new edges for the Facebook dataset to evaluate the performance of the algorithms with extremely low budgets. We compute the social capital for each group (male and female) and the structural group unfairness on both the original and the augmented graphs (after all edges have been added) based on the defined

measures. Moreover, we compute them at each step of edge addition to shed light on the evolution of the structural group unfairness as new edges are added.

(a) Facebook (B=50)

	$\Delta\,\mathsf{R}_{\text{tot}}$	$\Delta\,\mathcal{R}_{\text{diam}}$	$\Delta\,\mathsf{B}_{\mathsf{R}}$
G (original)	41.62	0.042	0.107
Random	38.7	0.039	0.108
SDRF	41.6	0.042	0.106
FOSR	34.5	0.027	0.109
DW	36.3	0.031	0.104
Cos	28.7	0.029	0.120
ERG	**10.3**	**0.009**	**0.098**
S-ERG	41.6	0.042	0.107

(b) UNC28 (B=5000)

	$\Delta\,\mathsf{R}_{\text{tot}}$	$\Delta\,\mathcal{R}_{\text{diam}}$	$\Delta\,\mathsf{B}_{\mathsf{R}}$
G (original)	22.4	0.006	0.009
Random	19.8	0.005	0.014
SDRF	22.2	0.006	0.007
FOSR	19.7	0.005	0.017
DW	22.2	0.006	0.004
Cos	19.1	0.005	0.102
ERG	**8.8**	**0.002**	**0.003**
S-ERG	22.3	0.006	0.004

(c) Google+ (B=5000)

	$\Delta\,\mathsf{R}_{\text{tot}}$	$\Delta\,\mathcal{R}_{\text{diam}}$	$\Delta\,\mathsf{B}_{\mathsf{R}}$
G (original)	276.4	0.078	0.51
Random	129.4	0.037	0.47
SDRF	276.1	0.079	0.52
FOSR	240.7	0.068	0.50
DW	274.1	0.078	0.51
Cos	86.8	0.025	0.47
ERG	**37.1**	**0.011**	**0.29**
S-ERG	276.4	0.079	0.52

Table 7. Structural group unfairness, *i.e.*, differences in group social capital between males and females, before and after the graph interventions. Best result highlighted in bold.

3.4.4 Structural Group Unfairness Mitigation

Table 7 depicts the three structural group unfairness measures on the original graph G and after adding 50 edges to the Facebook dataset and 5,000 edges to the UNC28 and Google+ datasets. The groups are defined based on gender (male, female) and the disadvantaged group are females according to the three structural group unfairness measures, as depicted in Table 5 and top row of Table 7. The disparities in social capital between groups are particularly large in the Google+ network with a $\Delta\,\mathsf{R}_{\text{tot}}$ of 276.4, meaning that females have significantly lower levels of information flow than males in this network. $\Delta\,\mathsf{B}_{\mathsf{R}}$ also shows a difference of 0.51 on the control of the network, which is a large difference given that the values of B_{R} are in the range $[0, 2 - 2/|\mathcal{V}|]$.

Regarding the results of the graph intervention algorithms, we observe how edge augmentation via `ERG-Link` outperforms all the baseline methods on the three datasets in terms of reducing disparities in group social capital. Interestingly, the larger the unfairness in the original graph, the larger the improvement after the intervention with `ERG-Link`. For example, the isolation disparity, $\Delta\,\mathsf{R}_{\text{tot}}$, improves by **75%**, **61%** and **87%** after `ERG-Link`'s intervention in the Facebook, UNC28 and Google+ networks, respectively. A similar behavior is observed for the other structural group unfairness measures.

For illustration purposes, we also report the results of applying Algorithm 2 but where the added edges connect the nodes with the smallest —instead of the largest— R_{uv}, similar to how link recommendation algorithms work. We refer to this method as the "Strong" version and denote it with an "S-". As expected, strong edges do not improve the structural group unfairness since such methods are not designed to improve the information flow in the most isolated nodes in the graph, but to connect nodes that are already structurally close to each other (foster strong ties).

Note how the edge augmentation when using DW and Cos distances or via SDRF (based on curvature) and FOSR (based on λ_2) methods yields graphs with significant levels of structural group unfairness. Furthermore, in the case of using DW distances, the improvement in performance worsens as the graph gets larger (Google+). While using Cos distance for edge augmentation improves $\Delta\,\mathsf{R}_{\text{tot}}$ and $\Delta\,\mathcal{R}_{\text{diam}}$, it is unable to always improve $\Delta\,\mathsf{B}_{\mathsf{R}}$ due to the inherently more intricate nature of control disparity optimization. Unlike $\Delta\,\mathsf{R}_{\text{tot}}$

and $\Delta\,\mathcal{R}_{\text{diam}}$, minimizing $\Delta\,\mathsf{B}_{\mathsf{R}}$ requires the precise identification of network gaps which are difficult to detect using a cosine similarity distance.

In addition, over-squashing oriented methods (SDRF and FOSR) aim to improve the information flow by focusing only on the information bottlenecks of the network, which leads to a small effect on improving the overall information flow and reducing disparities. SDRF identifies where to add the edges partly based on the control $\mathsf{B}_{\mathsf{R}}(u)$ (Equation (46)) rather than based on the diameter $\mathcal{R}_{\text{diam}}$ as `ERG-Link` does. FOSR's goal is based solely on increasing the spectral bottleneck (λ_2), while `ERG-Link` considers the entire spectrum, leading to a better characterization and improvement of the information flow.

Finally, edge augmentation with `ERG-Link` is not only more effective in mitigating the group structural unfairness in the graph, but also more computationally efficient for middle-size graphs than most of the baselines as shown in Appendix B.4.

3.4.5 Overall Social Capital Improvement

(a) Facebook (50) Female

	$R_{tot}\downarrow$	$\mathcal{R}_{diam}\downarrow$	B_{R}
G	221.4	2.29	1.927
Random	211.5	2.26	1.927
SDRF	220.7	2.28	1.928
FOSR	202.1	2.15	1.926
DW	199.4	1.87	1.929
Cos	190.4	1.64	1.918
ERG	**138.7**	**0.43**	**1.933**

(b) UNC28 (5,000) Female

	$R_{tot}\downarrow$	$\mathcal{R}_{diam}\downarrow$	B_{R}
G	608.6	2.11	1.994
Random	435.0	1.23	1.992
SDRF	599.1	2.11	1.996
FOSR	509.9	1.33	1.990
DW	583.2	1.81	1.997
Cos	429.6	0.42	1.940
ERG	**316.8**	**0.10**	**2.001**

(c) Google+ (5,000) Female

	$R_{tot}\downarrow$	$\mathcal{R}_{diam}\downarrow$	B_{R}
G	564.1	1.31	1.807
Random	305.1	1.07	1.824
SDRF	561.1	1.31	1.805
FOSR	498.9	1.25	1.811
DW	558.7	1.31	1.806
Cos	230.9	0.29	1.827
ERG	**145.5**	**0.07**	**1.889**

(d) Facebook (50) Male

	$R_{tot}\downarrow$	$\mathcal{R}_{diam}\downarrow$	B_{R}
G	179.8	2.25	2.034
Random	172.8	2.22	2.035
SDRF	179.1	2.24	2.034
FOSR	167.6	2.13	2.035
DW	163.1	1.83	2.033
Cos	161.7	1.61	2.039
ERG	**128.5**	**0.41**	**2.031**

(e) UNC28 (5,000) Male

	$R_{tot}\downarrow$	$\mathcal{R}_{diam}\downarrow$	B_{R}
G	586.3	2.11	2.004
Random	415.2	1.23	2.005
SDRF	576.9	2.10	2.002
FOSR	490.2	1.33	2.007
DW	561.0	1.81	2.001
Cos	410.6	0.41	2.043
ERG	**308.0**	**0.09**	**1.998**

(f) Google+ (5,000) Male

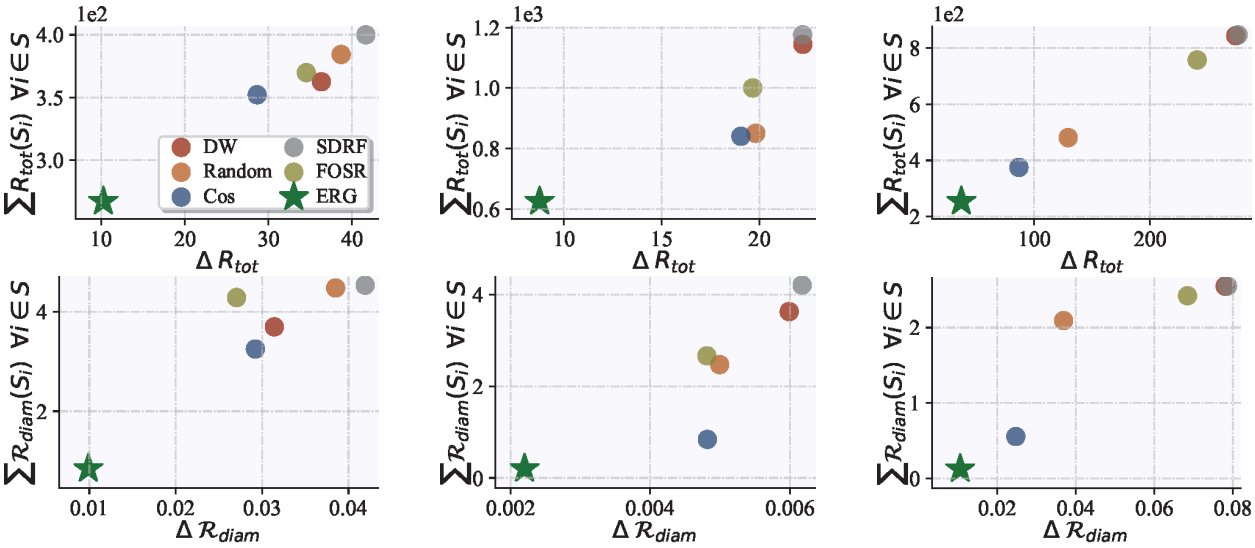
	$R_{tot}\downarrow$	$\mathcal{R}_{diam}\downarrow$	B_{R}
G	287.7	1.24	2.321
Random	175.7	1.03	2.293
SDRF	285.2	1.23	2.325
FOSR	258.1	1.17	2.315
DW	284.6	1.24	2.323
Cos	144.1	0.27	2.288
ERG	**108.5**	**0.06**	**2.185**

Table 8. Group Social Capital after Edge Augmentation. The best values are highlighted in bold. Note how edge augmentation via `ERG-Link` is able to not only increase the social capital of the disadvantaged group (females) but also of the rest of the groups (males).

In addition, the proposed system not only significantly mitigates the structural group unfairness ($\Delta\,\mathsf{R}_{\mathsf{tot}}$ and $\Delta\,\mathcal{R}_{\text{diam}}$); but also increases the social capital for all the groups in the graph, *i.e.*, the specific $\mathsf{R}_{\mathsf{tot}}(S_i)$ $\mathcal{R}_{\text{diam}}(S_i)$ or $\mathsf{B}_{\mathsf{R}}(S_i)$ for *all* the S_i groups in the graph, without harming any group. Table 8 depicts the group social capital for each group after each intervention, where `ERG-Link` consistently achieves the best results on each social capital measure for every of the groups.[11]

We also illustrate this result in Figure 20. For each dataset, the structural group unfairness metric ($\Delta\,\mathsf{R}_{\mathsf{tot}}$ or $\Delta\,\mathcal{R}_{\text{diam}}$) is shown on the x-axis of the graphs and the social capital metrics

[11]Note that, whereas for $\mathsf{R}_{\mathsf{tot}}(S_i)$ and $\mathcal{R}_{\text{diam}}(S_i)$ the lower the better (less isolated), the optimal value of B_{R} is to converge every group's control equal to $2 - 2/|\mathcal{V}|$.



(a) Facebook **(b)** UNC **(c)** Google

Figure 20. Pareto front of the structural group unfairness (X-axis) vs the sum of the group's isolation of all the groups (Y-axis) using R_{tot} and $\mathcal{R}_{\text{diam}}$ (denoted by X-axis' label). Best results correspond to the bottom left corner of the graph.

for all groups (sum of group isolation for all groups or sum of diameters for all groups) on the y-axis. In both axes, the lower the values, the better. Edge augmentation via `ERG-Link` clearly outperforms any other graph intervention strategy, providing evidence that it not only reduces inequalities in social capital between groups, but also improves the social capital of all groups.

3.4.6 Evolution of Structural Group Unfairness Throughout the Interventions

Figure 21 illustrates the evolution of the structural group unfairness metrics for the Google+ dataset as new edges are added to the network with a total budget of 5,000 edges. As seen in the Figure, edge augmentation via `ERG-Link` quickly mitigates the group isolation and diameter disparities, even after the addition of a small number of edges. Furthermore, edge augmentation via `ERG-Link` exhibits a smoother and more consistent reduction in control disparity ($\Delta\,\mathsf{B}_{\mathsf{R}}$), in contrast to the stair-step behavior observed when adding edges using the baseline methods.

Note that B_{R} is a finite resource to be allocated among the groups and cannot be globally maximized. Hence, the right-most graphs show how $\Delta\,\mathsf{B}_{\mathsf{R}}$ is improved by decreasing B_{R} of the group with the highest initial B_{R} and increasing it otherwise (top-right). The goal is to converge both to the optimal bound of $2 - 2/|\mathcal{V}|$ as indicated by the black line.

In addition, we conducted further experiments to analyse the evolution of the interventions. Specifically, we analyse the evolution of structural group unfairness and group social capital metrics with a small budget for adding edges ($B = 50$ in the Facebook dataset) or a large budget ($B = 5,000$ in the Facebook dataset). These results are included in Appendices B.3.1 and B.3.2, where (i) Figures 35 and 36 show the evolution of both the group social capital

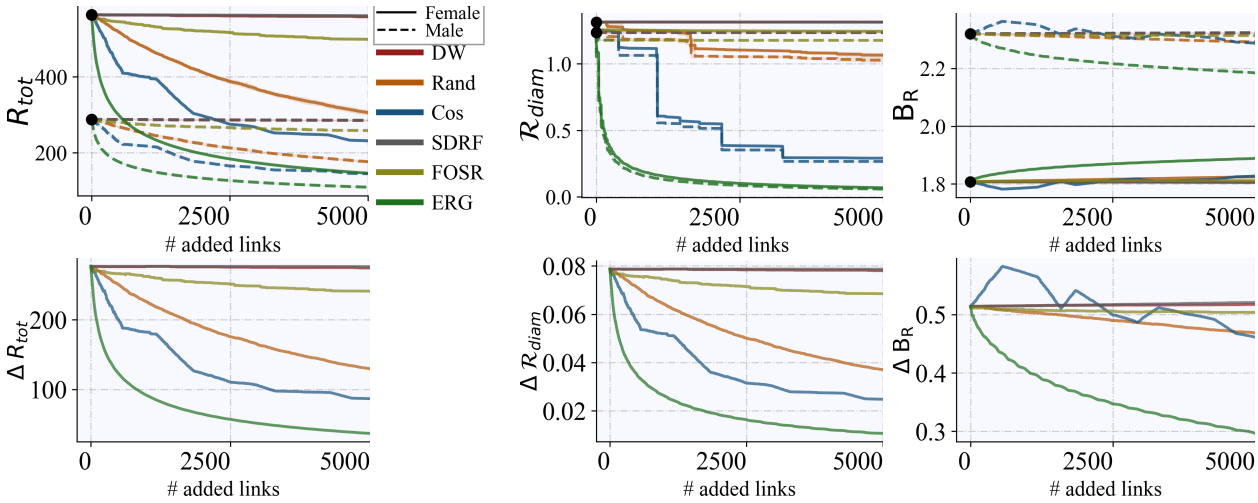


Figure 21. Evolution of the group metrics (top-row) and structural group unfairness metrics (bottom-row) as the number of added edges increases on the Google+ dataset with a total budget of 5,000 new edges.

and the structural group unfairness metrics, and (ii) Figure 34 shows the distribution of the individual nodes' social capital metrics in each of the groups in the original graph and the resulting graphs after edge augmentation. The results are consistent: for a fixed budget, edge augmentation via `ERG-Link` yields the best results both for the mitigation of the structural group unfairness and the increase in the overall social capital of all the groups in the graph. Additionally, for a fixed structural group unfairness mitigation goal, edge augmentation via `ERG-Link` achieves it with a significantly smaller budget of added edges than any of the baseline methods.

3.5 Discussion

The proposed `ERG-Link` algorithm may be seen as an *algorithmic reparative* social tool [120], *i.e.*, a computational approach designed to address and repair social inequalities, injustices or biases, in our case, in the context of social networks. Our method contributes to the use of social networks, seen as socio-technical systems [49], for social good purposes. While an extensive body of work has been dedicated to the identification and mitigation of algorithmic bias [42, 118], this chapter aims to trigger a deeper reflection and further discussion about the principles of algorithmic fairness and justice in relation social networks, their ultimate goals and their intricate business models of data trading among platform owners, data brokers, service and advertising companies [359]. There is not a universal interpretation of fairness [117], yet the distributive paradigm of fairness [331] dominates the political and philosophical discussions of social justice today. As a result, most discussions and implementations of fairness focus on how outcomes or benefits are distributed across different socially relevant groups. Nevertheless, there are alternative notions of fairness that remain unexplored in social networks. For example, a prioritarian conception of fairness [109] would give priority to the individuals belonging to a vulnerable social group. Moreover, prominent voices in social philosophy have long defended that fairness can be understood not only as the redistribution of individual social capital, but also as the social recognition of identity

groups [176]. In this chapter, we focus on group fairness and combine both a prioritarian and a Rawlsian approach since the proposed intervention aims to mitigate the social capital gap for the most disadvantaged group while increasing the social capital of all groups. Given the importance of social networks in the definition of the social fabric, the ultimate goal of our work is to spur a reflection on the potential use of social networks as *reparative* tools for social inequality [120].

While one could argue that defining groups in terms of protected attributes can be considered a form of discrimination, our approach does not aim to systematically increase the social capital of a pre-defined vulnerable social group, but to detect the social group or groups that is structurally disadvantaged in the network and implement a mechanism at scale that benefits an entire community. This type of mitigation of structural unfairness is particularly relevant since the potential disadvantages suffered by groups in social networks can add up to already existing social conditions that contribute to systemic injustice [416].

The implications of our work go beyond the mitigation of disparities in group social capital and information silos within social networks. The proposed metrics and graph intervention algorithm can also promote innovation and social mobility, since new connections between individuals across different groups are created while prioritizing the information flow gain to/from the most structurally disadvantaged group, and thus strengthening the influence of individuals that suffer from systemic injustice. From a practical perspective, we envision our proposal as a complement to existing graph interventions in a hybrid setup that combines the addition of edges connecting nodes with both small (strong ties) and large (weak ties) effective resistances. `ERG-Link` provides incentives to users in terms of increasing their social capital [44] and opportunities for innovation by connecting distant nodes [327]. Therefore the deployment of `ERG-Link` implies a more *socially responsible* way to manage connections. Note that the aim of this proposal is not to present a universal solution to group social capital disparities in social networks, but a contribution to spur further reflections regarding the societal impact of algorithms in relation to their purpose and use.

Finally, `ERG-Link`'s objectives are well aligned with the Art. 34 of the European Digital Services Act (EU DSA) [158], which aims to regulate online platforms and services, enhancing the transparency, accountability and safety of social platforms. The DSA imposes strict responsibilities to very large online platforms (those with more than 45 million EU users), given their significant societal impact. These platforms must conduct regular risk assessments, reduce systemic risks –such as the spread of harmful content or the manipulation of democratic processes– and submit to independent audits. In particular, Article 34 of the EU DSA states that "providers of very large online platforms need to assess their risks to society and adapt their algorithms if necessary" and Article 38 demands that "providers of very large online platforms and of very large online search engines that use recommender systems shall provide at least one option for each of their recommender systems which is not based on profiling". The solution offered by `ERG-Link`, where new edges are added with the goal of mitigating disparities in social capital across groups while increasing everyone's social capital, could be of great interest to providers of very large platforms in the context of this regulation.

3.6 Conclusion and Future Work

In this chapter, we have presented a novel method, based on the effective resistance, to measure and mitigate group social capital disparities within a social network, where the groups are defined according to the values of a protected attribute. Grounded in spectral graph theory, we have introduced three measures of group social capital based on the effective resistance and we have proposed to mitigate disparities in group social capital by means of `ERG-Link`, a budgeted edge augmentation algorithm that systematically increases the social capital of the most disadvantaged group in the network and hence reduces the disparities in group social capital. In extensive experiments with three benchmark graph datasets, we illustrate how `ERG-Link` is the most effective method to decrease disparities in group social capital when compared to five baselines.

In future work, we plan to explore alternative non-greedy edge augmentation algorithms to mitigate structural group unfairness using our effective resistance-based measures and we would like to incorporate additional node features corresponding to the characteristics of the individuals in the network. We also maintain ongoing discussions with several community organizations to define a project with participants that arrive in another country as refugees, for whom information access and connection with the local population of the hosting country are of paramount importance.

Chapter 4

Towards Human-AI Complementarity in Matching Tasks

Chapter summary and context

In this chapter, we focus on the **use** sphere, and propose `CoMatch`, a human-AI complementarity system for resource-allocation tasks that satisfies the harm-prevention requisite in Trustworthy AI while respecting the semi-automation constraint of Article 22(1) GDPR for socially consequential decisions. `CoMatch` makes the matching decisions it is most certain about, deferring the rest of the decisions to the human decision maker. Furthermore, `CoMatch` optimizes how many matching decisions it defers to the human to provably maximize the overall performance of both the algorithm and the human working together.

This chapter is based on the following working paper under review:

[25] **Adrian Arnaiz-Rodriguez**, Nina Corvelo, Suhas Thejaswi, Nuria Oliver, and Manuel Gomez Rodriguez. "Towards Human-AI Complementarity in Matching Tasks". In: *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases - The Third Workshop on Hybrid Human-Machine Learning and Decision Making*. Sept. 2025. URL: https://arxiv.org/abs/2508.13285

4.1 Introduction

As AI systems become more integrated into high-stakes decision-making processes, understanding how they interact with human users has become a growing concern. There is increasing consensus that the ultimate goal of machine learning systems for decision support is to achieve human-AI complementarity [15, 37, 208, 217, 364]. Human-AI complementarity aims for the decisions made by a human using a decision support system to be, in expectation, better than those made by either the human or the AI system alone.

In recent years, there have been significant advances in developing decision support systems that achieve human-AI complementarity by design [113, 123, 365, 367]. However, these advances have predominantly focused on classification tasks. In this work, we focus on decision support systems designed to achieve human-AI complementarity in matching tasks.

Matching tasks appear in a variety of high-stakes application domains, including matching refugees to resettlement locations [6, 7, 36, 178, 258], patients to appointments with clinicians [339], or blood and organ donors to recipients [30, 285]. In these cases, a decision maker needs to distribute a limited set of resources, such as locations, appointments, or donations, among a pool of individuals, such as refugees, patients, or recipients, through matching decisions so that individuals will use the resources effectively. However, at the time of assignment, there is uncertainty as to whether a given individual will actually make effective use of the resource.

In this context, data-driven algorithmic matching systems leverage machine learning models to predict how effectively each individual under consideration would use a given resource.[12] Then, they leverage these predictions to make matching decisions by solving a maximum weight bipartite matching problem [256, 376] where nodes represent individuals and resources, respectively, and edge weights represent the classifier's confidence that an individual would effectively use a given resource. The goal is to find an assignment between individuals and resources that maximizes the sum of the edge weights. While such algorithmic matching systems can optimize predicted outcomes, they still rely on human decision makers to achieve effective human-AI complementarity [7]. In practice, this means decision makers must be able to interpret and, when necessary, override the system's recommendations. However, there is growing evidence that decision makers often struggle to make these judgments appropriately [255, 372, 412, 422], which can undermine the potential benefits of the system. In this work, we propose a data-driven algorithmic matching system designed to achieve human-AI complementarity without requiring to understand when or how to override the system's recommendations. In detail, our contributions in this work are three-fold:

1. **A collaborative algorithmic matching framework:** We propose *Collaborative Matching* (`CoMatch`), a data-driven algorithmic matching framework that, unlike traditional matching frameworks, does not make all matching decisions. Instead, it matches only the individual–resource pairs that it is most certain about—by solving a maximum weight bipartite imperfect matching problem [329]—and defers the remaining decisions to a human decision maker. Along the way, our framework utilizes `UCB1`, a well-known multi-armed bandit algorithm [353], to optimize how many matching decisions to defer to the human decision maker to provably maximize the performance.

2. **Empirical validation through a large-scale human subject study:** We conduct a large-scale human subject study with 800 participants, who make 6400 matching decisions over 40 different instances of a stylized matching task. In this setup, human participants have access to more information than the classifier simulating a realistic human-AI collaboration scenario, allowing us to evaluate the benefits of deferring matching decisions to human decision-makers.

3. **Open data and implementation:** To facilitate future research and ensure reproducibility, we release the full dataset collected from our user study and an implementation of the proposed matching framework as open-source at `https://anonymous.4open.science/r/collaboration-matching-tasks`.

[12] In practice, one needs to measure whether an individual makes good use of a given resource using proxy variables which need to be chosen carefully to avoid perpetuating historical biases [61, 184, 375].

Further related work Our work builds upon related work on learning under algorithmic triage and multi-armed bandits.

The literature on algorithmic triage aims to develop classifiers that make predictions for a given fraction of the samples and leave the remaining to human experts, as instructed by a triage policy [100, 121, 122, 297, 298, 306, 326]. Here, the triage policy determines who predicts each sample independently of each other. In contrast, the proposed system determines who makes each matching decision for each individual in a pool *jointly* by solving a linear program. In this context, note that learning under algorithmic triage has been also extended to reinforcement learning settings [32, 180, 368, 386].

Furthermore, our research contributes to an extensive line of work that uses multi-armed bandits in real-world applications [5, 77, 131, 134, 293, 299, 365]. Within this area of research, our work is most closely related to that of Straitouri et al. [365], which has used multi-armed bandits to optimize the performance of decision support systems for classification tasks based on prediction sets.

The rest of the chapter is organized as follows. Section 4.2 formalizes the matching task and introduces a framework for human-AI collaboration in matching tasks. Section 4.3 describes an algorithm to find the optimal number of decisions to defer to a human decision maker. Section 4.4 presents an empirical evaluation of our framework using a large-scale human-subject study. Section 4.5 discusses limitations of our approach and outlines directions for future work. Finally, Section 4.6 offers concluding remarks.

4.2 A System for Human-AI Complementarity in Matching Tasks

We consider a matching task where, for each task instance, a decision maker needs to distribute a limited set $\mathcal{R}$ of k resources, each resource $r \in \mathcal{R}$ with a capacity c_r, among a pool $\mathcal{I}$ of n individuals using a given amount of information $\mathbf{z} = (z_i)_{i \in \mathcal{I}} \in \mathcal{Z}^n$ about them. Each individual $i \in \mathcal{I}$ can receive a single resource $r \in \mathcal{R}$. The pool size n and the capacities c_r may change across task instances. Further, each individual $i \in \mathcal{I}$ may ($y_i(r) = 1$) or may not ($y_i(r) = 0$) make an effective use of the resource r they receive. We say that effective use occurs when, upon receiving r, the individual i benefits from it in a way that fulfills the intended purpose of the allocation. The benefit depends on the application domain. For example, in job placement it might mean job retention or satisfaction; in healthcare, it could mean improved health outcomes, etc. Thus, the decision maker aims to make matching decisions that maximize the number of individuals who make effective use of the resources.

Let $f : \mathcal{X} \times \mathcal{R} \to [0, 1]^k$ be a pre-trained classifier that, for each resource $r \in \mathcal{R}$, maps an individual's feature vector $x = \phi(z) \in \mathcal{X}$, where $\phi(\cdot)$ is an arbitrary transformation[13], to a confidence score $f_r(x)$, which quantifies how much the classifier believes the individual will make good use of the resource r.[14] Given the individuals' confidence scores $\mathbf{f} = (f_r(x_i))_{i \in \mathcal{I}, r \in \mathcal{R}}$, our data-driven algorithmic matching system helps the decision maker by automatically making

[13]The transformation $\phi(\cdot)$ models the fact that, in most application domains of interest, the classifier may have access to less information about the individuals than the decision maker. Otherwise, one may argue that pursuing human-AI complementarity is not a worthy goal [15].

[14]The assumption that $f_r(x) \in [0, 1]$ is without loss of generality.

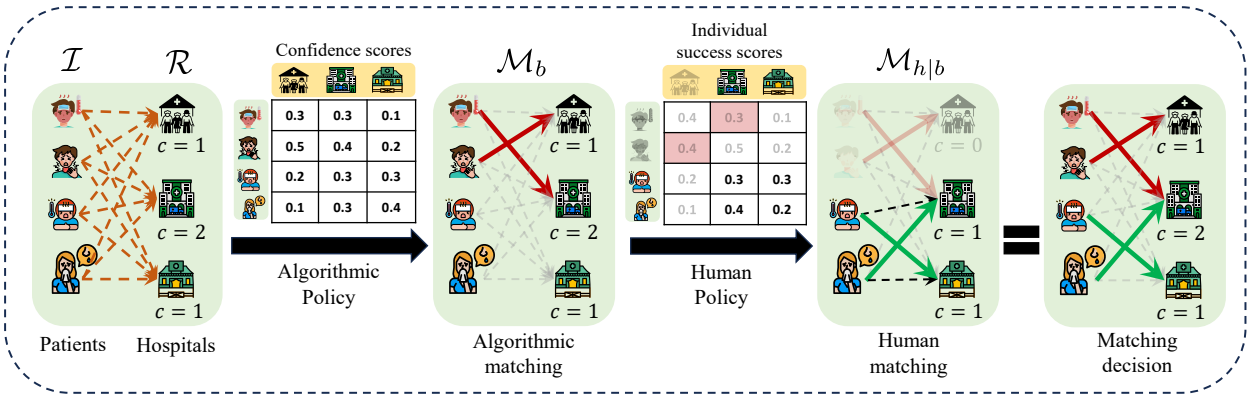


Figure 22. Our data-driven algorithmic matching framework, `CoMatch`, is illustrated in the context of matching patients to available appointment time slots in hospitals. Given a pool $\mathcal{I}$ of n patients and a set $\mathcal{R}$ of time slots, the framework assists the decision maker by automatically making (feasible) matching decisions $\mathcal{M}_b$ for b patients, using confidence scores from a pre-trained classifier. The remaining $n-b$ patients are deferred to a human decision maker h, who completes the matching decisions $\mathcal{M}_{h|b}$, using more accurate information—referred to as individual success score—on the likelihood of each patient attending an available appointment time slot. Algorithmic assignments are indicated by red arrows and human assignments are shown in green.

(feasible) matching decisions $\mathcal{M} \in 2^{\mathcal{I}\times\mathcal{R}}$ for a subset of the individuals $\mathcal{I}' \subseteq \mathcal{I}$ using an algorithmic policy $\pi(\mathbf{f}, \mathbf{c})$. Then, the decision maker needs to make (feasible) matching decisions $\mathcal{M}_h \in 2^{\mathcal{I}\setminus\mathcal{I}'\times\mathcal{R}}$ about the individuals that have been left unmatched by the system using an unknown policy $h(\overline{\mathbf{z}}, \overline{\mathbf{c}})$, where $\overline{\mathbf{c}}$ denotes the capacities left unused by the system and $\overline{\mathbf{z}}$ denotes the information about the individuals left unmatched by the system.[15] Figure 22 illustrates the proposed data-driven algorithmic matching system `CoMatch`.

The goal is to find the optimal algorithmic policy π^* that maximizes the average number of individuals who make an effective use of the resources across pools[16], *i.e.*,

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{(i,r)\in\mathcal{M}} Y_i(r) + \sum_{(i,r)\in\mathcal{M}_h} Y_i(r)\right]. \tag{20}$$

where the expectation is over the randomness in the pool generation process, the matching decisions $\mathcal{M}$ depend on the algorithmic policy π, and the matching decisions $\mathcal{M}_h$ depend both on the human policy h and indirectly on the algorithmic policy π. However, to solve the above maximization problem, we need to first specify the class of algorithmic policies we aim to maximize performance upon. Here, we consider algorithmic policies π_b that make $\mathcal{M}_b = n-b$ matching decisions by solving the following maximum weight bipartite imperfect

[15] Formally, $\overline{\mathbf{z}} = (z_i)_{i\in\mathcal{I}\setminus\mathcal{I}'}$ and $(\overline{\mathbf{c}})_r = c_r - |\{i \,|\, (i,r) \in \mathcal{M}\}|$.

[16] We denote random variables with capital letters and realizations of random variables with lower case letters.

matching problem [329]:

$$\begin{aligned} \underset{\mathcal{M}_b}{\text{maximize}} \quad & \sum_{(i,r)\in\mathcal{M}_b} f_r(x_i) \\ \text{subject to} \quad & |\{r \,|\, (i,r) \in \mathcal{M}_b\}| \leq 1 \quad \forall i \in \mathcal{I}, \\ & |\{i \,|\, (i,r) \in \mathcal{M}_b\}| \leq c_r \quad \forall r \in \mathcal{R}, \\ & |\mathcal{M}_b| = \max(n-b, 0) \end{aligned} \tag{21}$$

where $b \in \{0, \ldots, N\}$ is a parameter that controls the number of matching decisions that the system defers to the human decision maker and N is the maximum pool size.

Note that, by definition, the algorithmic policies π_b make the $n - b$ matching decisions that, according to the predictions made by the classifier, maximize the average number of individuals who would make an effective use of the resources. Moreover, finding the optimal policy that maximizes performance reduces to the problem of finding the optimal parameter value b^*, *i.e.*,

$$b^* = \arg\max_b \; \mathbb{E}\left[\sum_{(i,r)\in\mathcal{M}_b} Y_i(r) + \sum_{(i,r)\in\mathcal{M}_{h|b}} Y_i(r)\right], \tag{22}$$

where we denote the matching decisions made by the human decision maker using $\mathcal{M}_{h|b}$ to explicitly indicate that they indirectly depend on the parameter b. Importantly, the matching decisions made by a decision maker who uses the proposed system with the optimal parameter value b^* are guaranteed to be better or equal, in expectation, than the matching decisions made by either the decision maker or the system on their own, achieving human-AI complementarity. This is because, under $b = N$, the human decision maker makes all matching decisions and, under $b = 0$, the proposed data-driven algorithmic matching system makes all matching decisions.

4.3 Optimizing for Human-AI Complementarity

Similarly as in the (standard) maximum weight bipartite matching problem [256, 376], we can recover the solution to the imperfect matching problem defined by Equation (21), which may be non unique, from the solution to the following linear program:

$$\begin{aligned} \underset{\mathbf{v}}{\text{maximize}} \quad & \sum_{i\in\mathcal{I}, r\in\mathcal{R}} f_r(x_i)\, v_{ir} \\ \text{subject to} \quad & \sum_{r\in\mathcal{R}} v_{ir} \leq 1 \quad \forall i \in \mathcal{I}, \\ & \sum_{i\in\mathcal{I}} v_{ir} \leq c_r \quad \forall r \in \mathcal{R}, \\ & \sum_{i\in\mathcal{I}, r\in\mathcal{R}} v_{ir} = \max(n-b, 0) \\ & v_{ir} \geq 0 \quad \forall i \in \mathcal{I}, r \in \mathcal{R}. \end{aligned} \tag{23}$$

In particular, the above linear program is guaranteed to have an optimal integral solution $\mathbf{v}^* = (v^*_{ir})_{i\in\mathcal{I}, r\in\mathcal{R}} \in \{0,1\}^{n\cdot k}$, as shown in Appendix C.1, and thus it holds that $\mathcal{M}_b =$

Algorithm 3: UCB1

Input: T, N

$t \leftarrow 0$, $\boldsymbol{\gamma} \leftarrow \mathbf{0}$, $\boldsymbol{\nu} \leftarrow \mathbf{0}$

while $t < T$ **do**

  for $b \in \{0, \ldots, N\}$ **do**

    $\mu(b) = \gamma(b)/\nu(b)$

    $\epsilon(b) = \sqrt{2 \log T / \nu(b)}$

  $b_t \leftarrow \arg\max_b \mu(b) + \epsilon(b)$

  $\mathbf{z}_t, \mathbf{c}_t \sim P(\boldsymbol{Z}, \boldsymbol{C})$

  $\mathcal{M}_{b_t} \leftarrow \pi_{b_t}(\phi(\mathbf{z}_t), \mathbf{c}_t)$

  $\mathcal{M}_{h|b_t} \leftarrow h(\overline{\mathbf{z}}_t, \overline{\mathbf{c}}_t)$

  $\gamma(b_t) \leftarrow \sum_{(i,r) \in \mathcal{M}_{b_t}} y_i(r) + \sum_{(i,r) \in \mathcal{M}_{h|b_t}} y_i(r)$

  $\nu(b_t) \leftarrow \nu(b) + 1$

$\{(i, \arg\max_{r \in \mathcal{R}} v^*_{ir}) \,|\, i \in \mathcal{I} \wedge \exists r \in \mathcal{R}, v^*_{ir} > 0\}$. In this context, it is also worth noting that there exist specialized algorithms to find an optimal integral solution in polynomial time [329].

Now, since we know how to find the algorithmic policy $\pi_b(\mathbf{f}, \mathbf{c})$ for any parameter value $b \in \{0, \ldots, N\}$, we look at the problem of finding the optimal parameter value b^* defined in Equation (22) from the perspective of multi-armed bandits from online learning [353]. In our problem, each arm corresponds to a different parameter value b and, at each round t, our system makes matching decisions $\mathcal{M}_{b_t}$ about $n - b_t$ individuals, a (potentially different) decision maker makes matching decisions $\mathcal{M}_{h|b_t}$ about the remaining b_t individuals, and the decision maker obtains a reward $\sum_{(i,r) \in \mathcal{M}_{b_t}} y_i(r) + \sum_{(i,r) \in \mathcal{M}_{h|b_t}} y_i(r)$. Then, the goal is to find a sequence of parameter values $\{b_t\}_{t=1}^T$ with desirable properties in terms of total regret $R(T)$, which is given by:

$$\begin{aligned} R(T) = T \cdot \mathbb{E}\left[\sum_{(i,r) \in \mathcal{M}_{b^*}} Y_i(r) + \sum_{(i,r) \in \mathcal{M}_{h|b^*}} Y_i(r)\right] \\ - \sum_{t=1}^{T} \mathbb{E}\left[\sum_{(i,r) \in \mathcal{M}_{b_t}} Y_i(r) + \sum_{(i,r) \in \mathcal{M}_{h|b_t}} Y_i(r)\right], \end{aligned} \tag{24}$$

where the expectation is over the randomness in the pool generation process.

To this end, we resort to UCB1 (Algorithm 3), a well-known multi-armed bandit algorithm, which is guaranteed to achieve an expected regret $\mathbb{E}[R(T)] \leq O(\sqrt{Nt \log T})$ for any $t \leq T$, where the expectation is over the randomness in the execution of the algorithm, as shown elsewhere.

4.4 Evaluation via a Human Subject Study

To empirically validate our approach, we conducted a large-scale human-subject study involving 800 participants recruited via the Prolific platform, who make 6400 matching decisions

Figure 23. User interface of our human subject study. The web-based user interface used on the Prolific platform is shown with an example matching task instance with $b = 11$ patients and 11 time slots, where a human participant has already assigned 4 patients. The bottom row displays the remaining availability for each appointment slot, which is dynamically updated as participants assign patients. Confidence scores are visualized using a color gradient ranging from white to green, where white indicates low confidence and darker shades of green indicates higher confidence. Additional task-relevant information—such as the number of pending assignments and the current score—were shown on the left and right sides of the interface, respectively.

over 40 different instances of a stylized matching task. This study was approved by the ethical review board of the faculty of mathematics and computer science at Saarland University, Germany.

In the first phase of the study, we designed a stylized setting in which participants were tasked with matching patients to appointment slots in a fictional hospital, in order to evaluate how effectively human decision-makers perform under a bounded time frame. This phase established a performance baseline for human decision-making. In the second phase, we used the data collected from this study in `CoMatch`, which strategically partitions the matching task between the algorithm and the human expert. By design, the resulting system aims to achieve human-AI complementarity: it produces matching outcomes whose expected utility meets or exceeds that of the human or the algorithm acting alone.

4.4.1 Human Subject Study Setup

The task involves matching patients on a web interface with available appointment times in a fictional hospital, with the objective of maximizing the likelihood that patients will make use of their assigned appointments. To support this, we assume access to a predictive classifier that estimates, for each patient-time-slot pair, a confidence score indicating the likelihood

of attendance.

Our study was conducted using a web-based user interface and participants were recruited via the Prolific platform. The interface, shown in Figure 23, visualizes individual success scores as a heat map grid, with patients as rows and appointment slots as columns, using a color gradient ranging from white to dark green—where white indicating very low values and progressively darker shades correspond to higher individual success scores. The interface displays a score indicating the utility of the current matching to provide participants with an opportunity to revise their assignments to improve their score. To further encourage them to find optimal assignments, additional monetary incentives were offered to the top 4% of the participants who consistently achieved high scores across all matching tasks, thereby motivating them to perform at their best.

Each matching task involves assigning a pool of $n = 20$ and $k = 10$ time slots, each slot with a capacity of up to 2 patients. For a given matching instance, the algorithmic policy first matches $n - b$ patients by solving the maximum weight imperfect bipartite matching problem defined in Equation (23), where the deferral parameter $b \in \{5, \ldots, 20\}$ determines the number of patients that are left unmatched. By design, the classifiers' confidence scores available to the algorithm are inaccurate. Conversely, the human participant has access to more accurate individual success scores—modeling a setting in which human decision-makers have richer and more reliable information about the patients. The remaining b patients are then assigned to a human participant recruited via the Prolific platform, who should complete the matching under a time limit of 2 minutes for each deferred assignment instance. Although submitting before the 2-minute limit was permitted, participants could submit only after all assignments were completed. If the time limit was exceeded, the current assignment was automatically submitted, and the system proceeded to the next task. Each participant completed 8 matching task instances, with the deferral parameter $b \in \{5, \ldots, 20\}$ varying across tasks.[17] To ensure coverage and diversity, each participant was assigned either all even or all odd values of b, sampled in a uniformly random order—thereby systematically varying the number of decisions deferred to the participant across the tasks. Further details on the generation of the matching task is provided in Appendix C.2.

The study begins with a detailed explanation of the matching task, followed by an introductory assignment that allowed participants to familiarize with the interface before starting the actual tasks. To ensure data quality, two attention checks were interleaved at the second and seventh positions to detect automated bots and verify that participants were paying adequate attention throughout the study. These attention checks were approved by experts from the Prolific platform. Failing either of these checks led to disqualification of 56 participants out of an initial cohort of 856, resulting in the 800 valid participants in our study.

Participants who successfully completed all tasks received a base compensation of £2 (£8.5 per hour). The top 4% of participants, based on the highest average expected utility across all tasks, received an additional bonus of £8.

4.4.2 Finding the Optimal Number of Decisions to Defer

Using data from the previously described stylized human-subject study, our goal is to learn the optimal number of patients b^* to defer to human decision-makers so that, by design, it

[17] In our experiments, we observed that small instances of matching problems were trivial for human participants to solve for $b \leq 4$. Therefore, the study considered instances with larger values of $b \geq 5$.

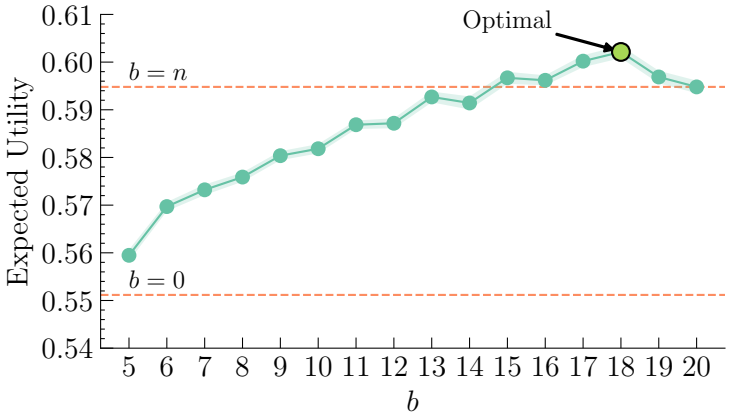


Figure 24. Empirical expected utility per arm achieved by `UCB1` across 100 independent realizations, each with a time horizon of $T = 2000$. The two orange dotted lines, correspond to the average utility of the matching decisions made solely by algorithmic ($b = 0$) and human ($b = n$) policy, respectively. Shaded bands denote the 95% confidence intervals across the 100 runs for each value of b.

achieves human-AI complementarity. As presented in Section 4.3, we cast the problem of finding b^* as a multi-armed bandit problem which we solve using `UCB1` algorithm. At each time step t, we randomly sample a pool with $n = 20$ patients, and `UCB1` selects an arm b_t, which corresponds to the number of patients deferred to the human decision maker. The reward associated with b_t is computed as the sum of the utility of algorithmic matching for $n - b_t$ patients and the utility of human matching for the remaining b_t patients.[18]

Figure 24 depicts the empirical expected utility for each value of b recovered by `UCB1`. Note that the algorithmic policy relies on inaccurate confidence scores to produce an assignment (or matching) that maximizes utility with respect to those scores. However, the resulting utility may fall short of the maximum achievable under accurate individual success scores, and can therefore be lower than the utility achieved by a human decision-maker. This is because the human decision-maker operates with different—and more accurate—confidence scores, reflecting a more informed understanding of the patients' conditions. This can be clearly observed in the figure the orange dotted horizontal lines depicting $b = 0$ and $b = n$.

The expected utility increases with the value of b—the number of patients deferred to the human decision-maker—as human participants, equipped with the individual success scores, are able to take more informed decisions. This trend supports our hypothesis that human judgment can complement algorithmic recommendations: combining human and algorithmic matching decisions yields higher expected utility than relying solely on the algorithmic policy. The utility peaks at $b^* = 18$, beyond which the expected utility begins to decline, which suggests that deferring more than 18 decisions to the human may surpass their cognitive capacity, leading to reduced performance.

[18]When multiple human decision makers have completed the matching for a given value of b for the same matching task instance, we randomly choose one of their solutions to compute the reward. If the human assignments were incomplete, the remaining unassigned patients were assigned to the available slots at random (see Section 4.4.3 for more details).

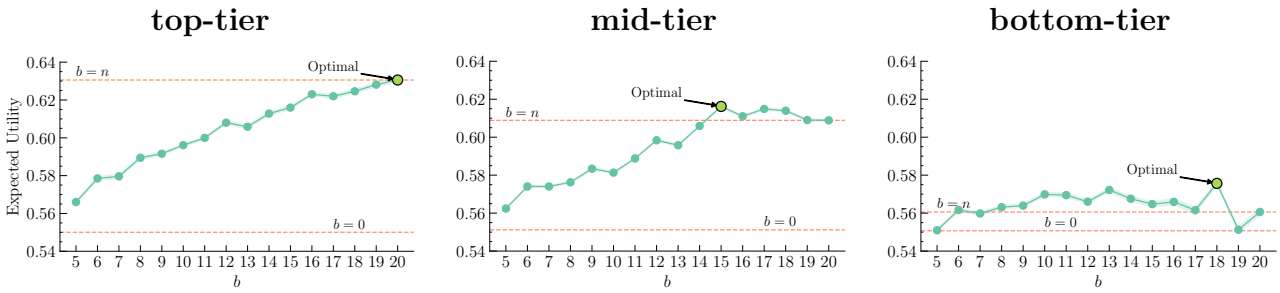


Figure 25. Empirical expected utility per arm achieved by `UCB1` across 100 runs (with time horizon $T = 1000$) for top-, mid- and bottom-tier participants. Red dotted lines mark $b = 0$ (algorithm-only) and $b = n$ (human-only); shaded bands show 95% confidence intervals.

Influence of human policy on matching outcomes A key factor that impacts the expected utility is the performance of the human decision-makers. Although the algorithm consistently returns a matching decision that maximizes the expected utility with respect to its (noisy) confidence scores, the total utility achieved by the combined human–algorithm policy depends largely on the quality of the human decisions. There is considerable variability across individual human decision makers, stemming from differences in cognitive ability, decision-making strategy, and subjective judgment. To better understand this effect, we conducted a more fine-grained analysis by categorizing human participants into three groups, top-tier, mid-tier, and bottom-tier, based on the quality of the matching decisions that they produced in the human subject study. The grouping of the participants was performed as follows. We computed the distribution of expected utility achieved by each participant, divided the distribution into deciles, and grouped participants accordingly: the bottom 40% were classified as bottom-tier, the middle 20% as mid-tier, and the top 40% as top-tier decision-makers. This resulted in 320, 160, and 320 participants in each group, respectively.

Figure 25 depicts the empirical expected utility for each value of $b \in \{5, \ldots, 20\}$, stratified by participant tier: top (left), mid (center), and bottom (right). As expected, in the case of top-tier participants, the human policy outperforms both the algorithmic and combined policies. This improvement is due to the effectiveness of the human participants as evidenced by the high utility of the human-only policy at $b = n$, where `CoMatch` achieves the peak utility ($b^* = 20$), suggesting that when human decision-makers are highly reliable deferring the decisions to them is beneficial. For mid-tier participants, the combined human–algorithm policy outperforms both standalone policies, with slightly lower average expected utility compared to top-tier participants. In this group, the peak utility is observed at $b^* = 15$, indicating a more balanced strategy—sharing the decision-making load between the algorithm and the human—could be better.

Regarding bottom-tier participants, one might naturally expect that deferring more decisions to the algorithm would lead to better outcomes, given the limitations of relying on low-performing human decision-makers. However, in our study setup, the average utility achieved by human-only and algorithm-only policies in this group is relatively close, with the algorithm-only policy performing slightly worse. As a consequence, the combined human-algorithm policy achieves its peak expected utility at $b^* = 18$, a slightly larger value than that obtained for the mid-tier participants. While this may appear counterintuitive, since one might expect that as the gap between human-only and algorithm-only policies narrows,

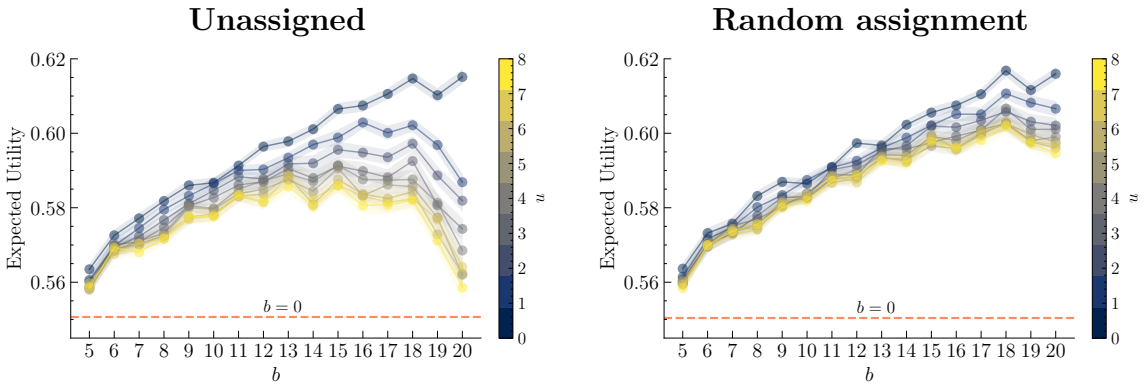


Figure 26. Empirical expected utility per arm achieved by `UCB1` across 50 independent realizations, each with a time horizon of $T = 2000$. Shaded regions indicate the 95% confidence interval. We progressively remove human participants—along with all their matching assignments—who left more than one patient unassigned in at least $u \in \{0, \ldots, 8\}$ of the matching assignment tasks, and report the empirical (average) expected utility for each u. In the left plot, the expected utility is computed by considering the confidence scores of the partial matchings completed by human participants, leaving the unassigned patients excluded while computing the expected utility. In the right plot, these unassigned patients are randomly assigned to available time slots, and their corresponding confidence scores are included when computing the expected utility of the combined matching.

the optimal value b^* should decrease, favoring fewer assignments to humans. However, we observe that the performance of participants in this tier is consistently poor across all values of $b > 10$. Thus, the benefit of assigning more individuals to participants in this tier is not as pronounced. As a consequence, `UCB1` converges to a relatively larger value of b^* compared to mid-tier participants, not because deferring more is better, but because no value of b yields reliably better human performance.

4.4.3 Handling Partial Assignments in the Human Subject Study

To recap, when computing the expected utility of the combined matching outcome, the algorithm first assigned $n-b$ patients by solving a maximum weight imperfect matching using (noisy) confidence scores, and the remaining b assignments were deferred to a human. The utilities from both components were then combined to obtain the expected utility of the full matching. If a participant was unable to complete all assignments within the 2-minute time limit, their partial assignment was saved as is. For the experiments in Section 4.4, to compute the expected utility of a partial human assignment, we *randomly* assigned the unmatched patients to the remaining available time slots. While this random assignment is reasonable, to further investigate the impact of different strategies for handling unassigned patients, we explore alternatives to random assignment. Specifically, we examine how the overall utility of the matching changes under these alternative assignment strategies. Additionally, we analyze the effect of filtering out participants who left some instances incomplete, to better understand how partial human assignments influence the expected utility of the combined matching.

Assessing empirical expected utility under partial human assignments

To understand how partial human assignments affect the empirical expected utility, we conducted an ablation study by progressively excluding participants—and all of their matching assignments—who left more than one patient unassigned in at least u tasks,[19] varying $u \in \{0, \ldots, 8\}$. Here, $u = 0$ corresponds to including only participants who completed all assignments in every matching task, while $u = 8$ includes all participants regardless of whether they left any assignments incomplete. For the unassigned patients in partial assignments, we considered two strategies. In the first, unassigned patients were left as-is and excluded while computing the expected utility. In the second, they were randomly assigned to available time slots, and the corresponding confidence scores were included to compute the expected utility. In Figure 26, we report the empirical expected utility of `UCB1` for each $b \in \{5, \ldots, 20\}$—the number of decisions deferred to human participants—averaged over 50 independent realizations. The left plot shows results using partial matchings as-is, while the right plot uses random assignments for unassigned patients. In the second strategy, random assignments were independently resampled for each realization when computing the expected utility. The expected utility of the algorithm-only policy ($b = 0$) is shown using an orange dotted line. However, the expected utility of the human-only policy varies for each u, depending on the number of participants retained after filtering. It corresponds to the utility at $b = 20$ for each u, *i.e.*, the last point in each curve.

As the value of u decreases, we retain only those participants who consistently completed all assignments. Under these stricter filtering conditions, our framework—splitting the task between humans and the algorithm—achieves higher expected utility across all values of b. Notably, when $u = 0$, *i.e.*, considering only the most reliable participants who completed all assignment tasks, the expected utility is close to optimal, even when the matching task with $b = 20$ is assigned to these human participants. This reinforces our claim: when human decision-makers are efficient at solving the matching task, deferring decisions to them is more beneficial than relying on the algorithm. On the other hand, as u increases being less restrictive and we include matchings from under-performing participants, the expected utility consistently decreases—both under partial matchings (with unassigned patients excluded) and under random assignment. This trend is particularly pronounced for higher values of b, where a greater portion of the matching task is delegated to human participants.

A fine-grained analysis of partial human assignments

We further examine the case, where all participants—and all of their matching assignments—are included, regardless of whether any patients were left unassigned. Figure 27 compares the expected utility of `UCB1`, with the left plot showing results where unassigned patients are left as-is, and the right plot showing results where they are randomly assigned to available time slots. An immediate observation is that as b increases, deferring more assignments, participants tend to leave more patients unassigned. This reinforces our hypothesis that, when faced with larger matching tasks under limited time, human cognitive limits hinder their ability to complete all assignments. Further, we observed that the gap between the expected utility of algorithm-only ($b = 0$) and human-only ($b = n$) policies is smaller when unassigned patients are left as-is, compared to when they are assigned randomly. This

[19] If exactly one patient remains unmatched, there is a single available time slot left. In this case, we assign the patient to the available slot, as the choice is unambiguous.

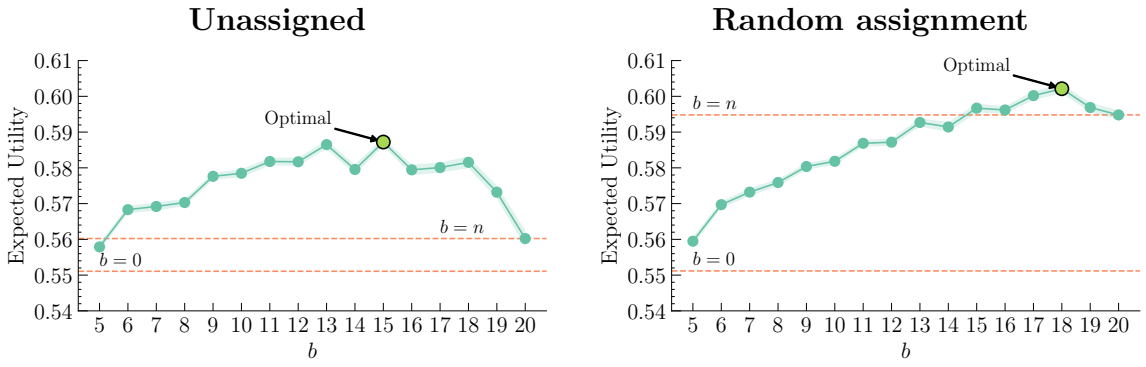


Figure 27. Empirical expected utility per arm achieved by `UCB1` across 100 independent realizations, each with a time horizon of $T = 2000$. In the left plot, unassigned patients are left as is and excluded while computing the expected utility. In the right plot, unassigned patients from human partial matchings are randomly assigned to available time slots, and the expected utility of the resulting combined matching is computed. Orange dashed lines indicate the expected utility of the algorithm-only ($b = 0$) and human-only ($b = n$) assignments. Shaded bands represent 95% confidence intervals.

suggests that, as expected, even randomly assigning patients improves the average expected utility, making it a preferable strategy compared to leaving patients unassigned entirely.

We further conducted a more fine-grained analysis by stratifying human participants into three groups—top-, mid-, and bottom-tier—based on the quality of the matchings they produced in the human subject study.[20] In Figure 28, we present the empirical expected utility achieved by `UCB1` for different values of b, by leaving the unassigned patients as-is while computing the expected utility. As expected, the gap between algorithm-only ($b = 0$) and human-only ($b = n$) assignments is largest for top-tier participants, who outperform the algorithm with a significant margin. This gap narrows for mid-tier participants, but reverses for bottom-tier participants, where human performance is substantially worse than the algorithm-only assignment—as clearly illustrated by the dotted orange lines in the plots. These findings contrast with the results in Figure 25, where—especially among bottom-tier participants who performed poorly—randomly assigning the unassigned patients improved the overall expected utility, underscoring the impact of how unassigned patients are handled.

We further observed that when human participants perform well, deferring more assignments to them is advantageous. This is most evident in top-tier, where the highest expected utility is achieved at $b = 20$, indicating that they can successfully solve the entire matching task without any algorithmic assistance. For mid-tier participants, the peak utility was observed at $b = 15$, while for bottom-tier participants, it drops to $b = 10$. These results reinforce our broader recommendation: as human performance decreases, it becomes less effective to defer a larger share of the task to human decision-makers.

[20]We computed the distribution of expected utility achieved by each participant, divided the full distribution into deciles, and grouped participants accordingly: the bottom 40% were classified as bottom-tier, the middle 20% as mid-tier, and the top 40% as top-tier decision-makers. This resulted in 320, 160, and 320 participants in each group, respectively.

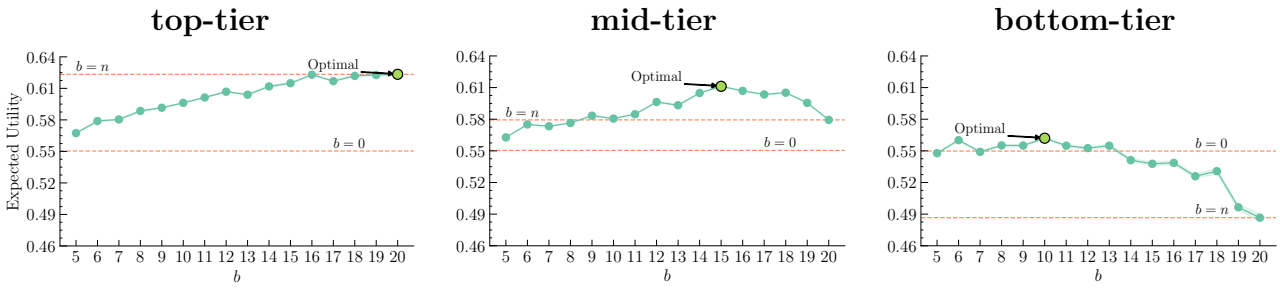


Figure 28. Empirical expected utility per arm achieved by `UCB1` across 100 different realizations, each of them with an horizon $T = 1000$, for a different subset of humans divided by their performance, leaving unassigned patients as is and excluded from the computation of the expected utility. Shaded bands show 95% confidence intervals.

4.5 Discussion and Limitations

In this section, we discuss several assumptions and limitations of our work, pointing out avenues for future research.

Methodology `CoMatch` contributes with a promising strategy to achieve human-AI complementarity in matching tasks by learning to defer a subset of matching decisions—depending on the size of the task and the ability of the human—to human decision-makers to achieve optimal expected utility. Our experimental results reveal that the expected utility decreases as the number of decisions are deferred to the human, suggesting that deferring based on problem size is indeed beneficial to achieve complementarity. However, task size alone does not always reflect the true complexity of the matching task. In many cases, the distribution of confidence scores, their proximity, and the perceptual difficulty for a human to distinguish between them may better indicate task difficulty than size alone. This highlights the need for more principled strategies that determine what to defer to humans, rather than merely how much to defer. A small matching instance may still be difficult for human decision-makers to solve, if the confidence scores are close, making them difficult to distinguish under limited time. Conversely, poorly calibrated confidence scores misrepresent the true likelihood of an individual utilizing a resource. In both cases, deferring decisions solely based on task size can be suboptimal.

Human subject study During the human subject study, an important design choice was the visualization of the confidence scores: while we used a heatmap interface, alternative visual representations could influence human interpretability and decision accuracy, potentially influencing the overall utility. Additionally, we designed the data generation process to balance realism with experimental control, aiming to create real-world scenarios while also retaining the flexibility to simulate conditions relevant to human-AI complementarity in matching tasks. To simulate an algorithmic matching policy that is less efficient in terms of average expected utility, we restricted the access to certain features while generating the algorithm's confidence scores, allowing us to tune—to some extent—the relative performance of human and algorithmic decision-makers. However, the setup is not without limitations. For example, one might expect the algorithm to outperform bottom-tier human participants, thereby creating clear conditions where deferring more decisions to the algorithm would yield

better utility. This pattern did not always emerge in our experimental setting. Thus, carefully engineering the synthetic generator in future work—*e.g.*, by varying access to features or task difficulty—could help isolate when and for whom deferral to humans or algorithms is most beneficial in matching tasks.

4.6 Conclusions

In this work, we have introduced `CoMatch`, a hybrid decision-support framework aimed at achieving human-AI complementarity in matching tasks, with a particular focus on identifying how many decisions to defer in large-scale matching tasks. Through a large-scale human subject study, we have empirically demonstrated that deferring a subset of decisions to human decision-makers can improve expected utility—especially when those decision makers are effective and capable. However, our findings also reveal important limitations: the utility gains from human input depends on the decision maker's performance, and diminishing returns emerge as cognitive load increases beyond a certain point. These insights suggest that future decision-support systems should be designed to account not only for algorithmic uncertainty, but also for the variability and cognitive limitations of human decision makers using these systems. A promising direction for future work is the development of deferral strategies that account for which decisions to defer, rather than simply how many.

Chapter 5

Effective AI Regulation

Chapter summary and context

In this chapter, we focus on the **governance** sphere, acknowledging the sociotechnical nature of implementing Trustworthy AI.
We analyze the overlap between existing Spanish regulations, European AI regulation, and what is technically feasible to develop an *effective AI regulation*. Specifically, this chapter explores the intersection between AI-based decision systems, the EU AI Act, and **Spanish labor law**, focusing specifically on how automated and semi-automated decisions in worker management intersect with both the requirements of Trustworthy AI and the legal obligations concerning *non-discrimination* and *principle of sufficient reason* in employment contexts.
Parts of this chapter are based on the following articles published in law journals:

[18] Adrian Arnaiz Rodriguez and Julio Losada Carreño. "La intersección de la IA fiable y el Derecho del Trabajo. Un estudio jurídico y técnico desde una taxonomía tripartita". Spanish. In: *Revista General de Derecho del Trabajo y de la Seguridad Social* 69 (2024). EN: The Intersection of Trustworthy AI and Labour Law. A Legal and Technical Study from a Tripartite Taxonomy, p. 2. ISSN: 1969-9626. URL: https://www.iustel.com/v2/revistas/detalle_revista.asp?id_noticia=427491

[17] Adrian Arnaiz Rodriguez and Julio Losada Carreño. "Estudio de la causalidad en la toma de decisiones algorítmicas: el impacto de la IA en el ámbito empresarial." Spanish. In: *Revista Internacional y Comparada de Relaciones Laborales y Derecho del Empleo* 12.3 (Dec. 2024). EN: Studying Causality in Algorithmic Decision Making: the Impact of IA in the Business Domain. ISSN: 2282-2313. URL: https://ejcls.adapt.it/index.php/rlde_adapt/issue/view/105

5.1 Introduction

The adaptation of AI systems to legal and ethical requisites is necessarily *context-dependent*. While the requirements of Trustworthy AI (TAI) offer a normative foundation for aligning

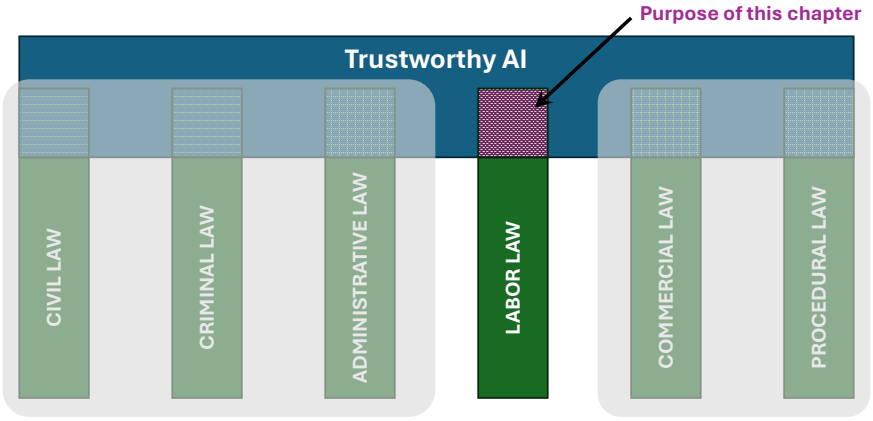


Figure 29. Trustworthy AI as a traversal framework intersecting with sectoral obligations

AI with fundamental rights [209], their concrete implementation must be tailored to the characteristics of each domain. In each scenario, tensions between the existing regulation, the technical nature of AI systems, and the AI regulation might emerge, which, when solved, will lead to a desirable *effective AI regulation* (See Figure 30).

In this chapter, we frame this study in decisions regarding employee-employer relationships regulated by the Spanish labor law (See Figure 29). This use case is an example of a domain already regulated by national labor law [72], a high-risk use case of AI as per the EU AI Act [162]. Structural power imbalances between employers and workers characterize this domain. In this context, AI systems are increasingly used to automate or assist in labor decisions, including hiring, task assignment, performance evaluation, promotions, and dismissals. Therefore, these decisions:

- must adhere to the principles and requirements of Trustworthy AI to minimize harm,
- and they are subject to legal constraints.

Contributions The main contributions of this chapter are twofold:

1. We analyze the intersection and alignment of Trustworthy AI requirements, the EU AI Act, and the Spanish labor law; and
2. We perform an in-depth study of a particular friction point: the *correlation-causation dilemma*. While data-driven AI systems are typically based on learning correlations from data, many employment-related decisions require existence and sufficient causes (*principle of sufficient reason*) and must avoid discriminatory reasoning. This dilemma exemplifies the broader misalignment between the algorithmic inner workings and legal norms in high-risk sociotechnical systems.

Figure 30 illustrates this chapter's contributions to the effective AI regulation landscape.

Interdisciplinary approach The methodology adopted in this chapter is interdisciplinary, bringing together experts from machine learning and labor law [21] In this chapter, we adopt

[21]Part of this work has been carried out under a Collaboration Agreement C039/23OT between Red.es and UCLM, which is focused on implementing the Digital Rights Charter in the context of digital rights in the workplace: https://www.derechosdigitales.gob.es/

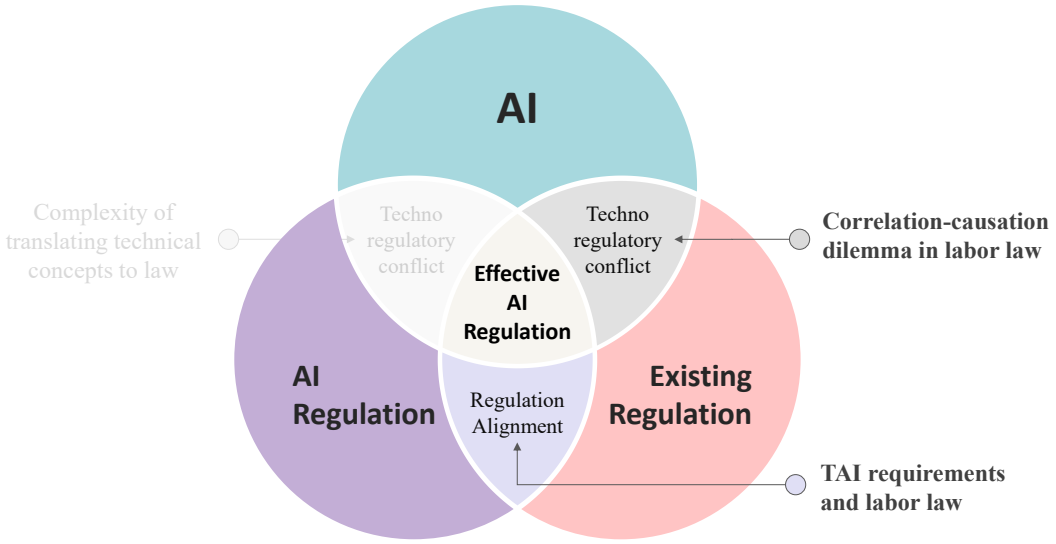


Figure 30. Contributions of this chapter at the overlap between AI systems, existing regulation, and the EU AI Act.

the following TAI HLEG guidelines [209]: (a) Consider alignment between technical and non-technical methods; (b) communicate clearly and proactively the capabilities and limitations of AI systems; (c) involve stakeholders throughout the AI system's life cycle; and (d) pay special attention to asymmetries of power and information, which is the main characteristic of the labor market.

5.1.1 Trustworthy AI as a Regulatory Anchor

The concept of *Trustworthy AI* is not merely an ethical aspiration; it holds explicit normative value within the European regulatory landscape. As articulated in multiple recitals (*e.g.*, 2, 3, 7, 20, 27, 89, 165) and provisions (*e.g.*, Art. 95.2(a)) of the EU AI Act, the requirements of TAI form a foundational pillar of the EU's approach to AI governance. Notably, Recital 27 of the EU AI Act directly references the centrality of TAI requirements in guiding expectations for the design, deployment, and oversight of AI systems.

Although the TAI framework identifies legality as one of three essential core components, it stresses that it is insufficient. As the group asserts:

> "Achieving Trustworthy AI requires not only compliance with the law, which is but one of its three components. Laws are not always up to speed with technological developments, can at times be out of step with ethical norms, or may not be well suited to addressing certain issues." [209]

This statement highlights a crucial point: legal frameworks often fail to keep up with rapid technological advances, resulting in regulatory gaps [172]. The TAI framework, rooted in ethical principles and human-centric values, helps to bridge these gaps by guiding responsible innovation in areas where legal norms may be absent or outdated. Rather than conflicting with legal compliance, TAI reinforces it by encouraging organizations to exceed minimum legal standards and proactively integrate ethical and societal considerations throughout the AI system life cycle.

5.2 The (Semi-)Automation of Cognitive Tasks

Over the last three centuries, productivity and labor efficiency have increased dramatically, driven by successive technological revolutions. While the first three industrial revolutions focused on the automation or semi-automation of physical tasks, the fourth industrial revolution that we are immersed in centers on the **automation and augmentation of cognitive functions** [307]. As labor law historically evolved in parallel with these technological shifts to protect the structurally weaker party in the employment relationship, examining the distinctive characteristics and implications that AI systems introduce in employer-employee relations becomes essential.

In the labor context, AI systems can be used for two types of tasks:

i. **Personnel/human resource management tasks** are performed by the employer. They commonly materialize the functions of management/organization, supervision/-control, and reward/sanction, which affect the employees.

ii. **Production cycle tasks** are typically assigned by an employee's manager and carried out by workers. These may involve physical activities – such as manual labor on a construction site – or cognitive functions — such as granting credit — providing legal advocacy in litigation, or making medical diagnoses and treatments.

This chapter focuses on **personnel management** tasks. These AI systems are called **AI-based worker management** (AIWM) [146, 239], characterized principally by the fact that a machine performs the core element generating the power asymmetry in labor relations in favor of the company, instead of a human [333]. Although the asymmetry of power is natural in the labor market, these technical systems reinforce it through the automation of relevant decisions, diminished transparency and accountability, with direct effects on workers' psychosocial factors [147].

Such dynamics pose significant risks to employee rights that current regulations cannot adequately address. Accordingly, it is imperative to exercise caution when designing, developing, and deploying AI systems to (semi-)automate labor decision-making.

Types of AIWM Labor Decisions AIWM systems can support or execute various decisions that human managers traditionally perform. These include actions taken during the *pre-contractual* and *contractual* phases of the employment relationship, and are currently regulated under the labor law.

During the *pre-contractual* phase, AI systems can assist with tasks such as recruitment, candidate selection, and decisions at the end of the trial period.

During the *contractual* phase, AI systems can be used to formalize contracts, monitor compliance with trial periods, adjust working conditions, implement disciplinary measures, handle temporary suspensions and reduced working hours, decide promotions, and determine contract termination. AI systems can also support decisions relating to geographic mobility, task assignment, or substantively altering employment conditions.

Finally, AI systems can guide voluntary and involuntary *terminations*, including individual or collective dismissals. Although these decisions vary significantly regarding their legal treatment and requirements, they all have one thing in common: AI systems can now influence or directly influence them.

5.2.1 The Case of Spain: Labor Law Regulations

The use of AI systems in employment contexts must comply with multiple layers of Spanish labor law, which encompasses the body of legal norms that apply to workers and govern the employment relationship.[22]. The main regulations that apply to the employee-employer relationship are as follows:

- At the core lies the Workers' Statute (*Texto Refundido de la Ley del Estatuto de los Trabajadores*, **TRET**)[72], which governs the individual and collective rights of workers, including hiring, dismissal, remuneration, and working conditions.
- The Employment Law (*Ley de Empleo*, **LE**)[63] regulates employment intermediation and public-private labor services, which may encompass algorithmic platforms like LinkedIn when used for matching or training purposes.
- The Law on the Prevention of Occupational Risks (*Ley de Prevención de Riesgos Laborales*, **LPRL**)[64] ensures workers' health and safety, applying to AI systems that affect physical or psychosocial well-being.
- Finally, the Law on Infractions and Sanctions in the Social Order (*Texto Refundido de la Ley sobre Infracciones y Sanciones en el Orden Social*, **TRLISOS**)[73] defines infractions and sanctions, providing the legal basis for penalizing employers whose use of machines violates labor protections, potentially applied to AI systems.

While these laws provide fundamental employment protections, they do not fully address the diverse risks introduced by using machines and AI systems in the labor domain. In practice, AI deployment in the workplace intersects with various regulatory regimes, including occupational safety, data protection, product safety standards, and collective bargaining mechanisms. Identifying and analyzing this intricate web of regulations is crucial to operationalize the seven Trustworthy AI requirements in labor contexts, as worker protections often stem from regulations beyond TRET, LE, LPRL, and TRLISOS.

Therefore, we identify which regulations apply to AI systems in the labor market, considering the phases of an AI system's life cycle, and we present it in Table 9 categorized by component. This holistic perspective allows for a structured classification of relevant normative instruments across three key domains: (1) *Input*, encompassing all data used during system development and operation; (2) The *AI system* itself, subject to both preventive (*ex ante*) and reactive (*ex post*) regulatory layers based on risk materialization; and (3) *Output*, focusing on the system's decisions or actions, must comply with the automated task's legal framework.

This analysis pays particular attention to how these regulations intersect with labor law, especially given the inherent power asymmetry between employer and employee and the robust legal protections granted to workers under Spanish and EU labor norms.

Importantly, when the management of employment relationships use AI systems (*e.g.*, hiring, task assignment, or wage determination) **the relevant labor law provisions remain fully applicable, regardless of whether the decisions are made directly by human**

[22] https://employment-social-affairs.ec.europa.eu/policies-and-activities/rights-work/labor-law_en Note that in this chapter, we use the American English spelling for consistency in the document.

Table 9. Regulations applicable to AI systems in the Labor Context

Component	Legal Instruments	Labor Law Relevance
Input	EU AI Act [162], EU DSA [158], EU Digital Markets Act [157], EU Data Governance Act [159], EU Data Act [154], EU GDPR [153], EU Free Flow of Non-Personal Data Regulation [156]	Platforms such as LinkedIn may act as employment intermediaries (Art. 40 LE [63]). Their data curation and recommendation influence labor-related AI systems, making data provenance legally relevant.
AI system (Ex Ante)	EU AI Act, EU DSA, EU Machinery Regulation [160], EU General Product Safety Regulation [161]	CE marking and EU conformity declarations affect occupational risk and equipment safety. Employers must inform workers about risks and safety measures (Art. 3.25 Machinery Regulation).
AI system (Ex Post)	EU Proposed AI Liability Directive [150], EU Proposed EU Product Liability Directive [151]	Covers fault-based and product-based civil liability for damages to workers in outsourced or multi-employer settings (Art. 24 LPRL [64]).
Standardization	Standardization bodies: ISO, IEEE, CEN-CENELEC, ETSI, AENOR (UNE in Spain)	Standards define safety requirements. Non-compliance can lead to civil, administrative, or criminal liability in workplace harm scenarios [71].
Output	TRET [72], additional national regulations about equal salary compensation [70] or working hours [67, 68], Collective Bargaining Agreements	AI decisions on shifts, wages, or benefits must follow labor law. Employers remain accountable regardless of whether decisions are human- or algorithm-driven (Arts. 28, 3438 TRET).

managers or indirectly via automated systems, *i.e.*, the use of an AI system does not bypass the regulation.

For the sake of illustration, we also present in Table 11 a list of the identified regulations applicable to AI systems, categorized by their origin — European or Spanish. Additionally, for a more extensive and detailed identification of the relevant regulations and analysis of their applicability, we refer the reader to the full manuscript, where the cases and legal provisions are thoroughly defined; see Arnaiz Rodriguez and Losada Carreño [18, Sec. IV].

In conclusion, deploying AI systems in employment contexts in Spain implicates a complex web of European and national regulations that span the AI life cycle. Consequently, understanding the full scope of applicable legal instruments is essential for legal compliance and safeguarding workers' rights in an increasingly digitized labor market.

5.3 Legal Definitions of AI

A key challenge in AI is the lack of a common understanding across disciplines [168, 312], defining key terms is essential. Therefore, before analyzing the concrete connections between

Trustworthy AI requirements, its regulation, and labor law, it is essential to clarify how AI is legally defined, as this determines the scope of the applicable regulations.

AI system

It is legally important to differentiate between an AI model and an AI system:

- **AI Model**: According to the OECD [310], this is "a central component of an AI system used to make inferences from inputs to produce outputs." In addition, ISO standards define an AI model as a "physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, process or data" [218], whereas a machine learning model is a "mathematical construct that generates an inference or prediction based on input data".

- **AI system**: Defined as "a machine-based system that is designed to operate with varying levels of autonomy and that can show adaptability after deployment, and that, for explicit or implicit objectives, infers from the input information it receives how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (Art. 3.1 EU AI Act).

Note that the TAI requirements and the EU AI Act usually refer to the risks and obligations of *AI systems*, in contrast to the AI model alone.

AI life cycle

From a practical perspective, AI systems are software programs integrated into physical hardware. According to European and national regulations, software is considered a machine, whether integrated into hardware or not. Like any machine, AI systems are subject to a life cycle [309], which can be divided into distinct phases. This can be graphically represented as in Figure 31.

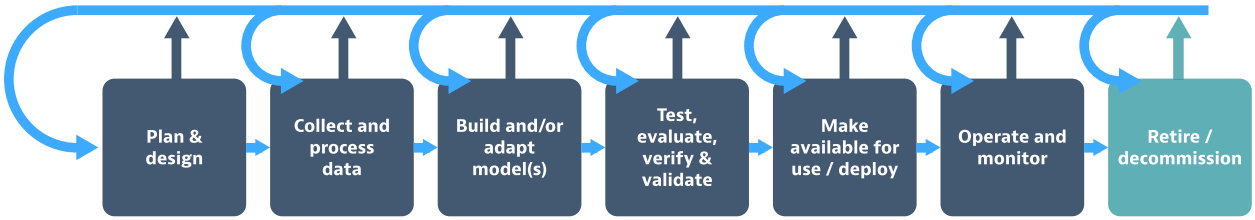


Figure 31. AI system life cycle. Source: OECD [309].

However, for legal purposes, it is more useful to distinguish only between two phases within an AI system's life cycle: the pre-implantation phase (construction phase) and the post-implantation phase (inference phase). The operations of an AI system across these two phases can be visually summarized in Figure 32.

Next, we present the different definitions related to the elements of an AI system as per the EU AI Act (Art. 3) [162]. This explanation is presented in simple technical terms, allowing us to draw an easier connection with the legal definitions.

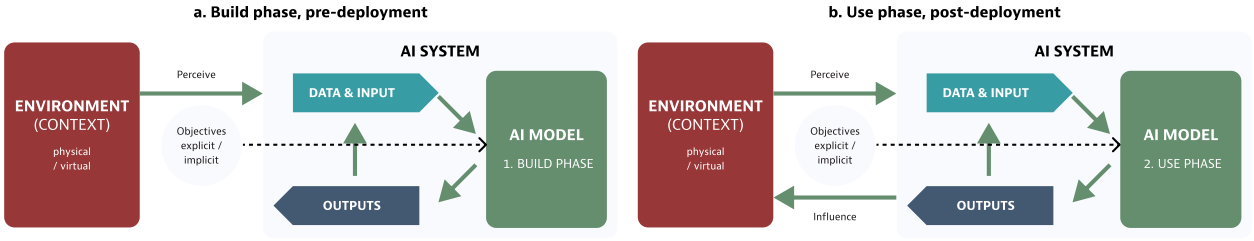


Figure 32. AI system pre and post-deployment stages. Source: OECD [309].

Pre-implantation The pre-implantation phase for AI models involves crucial training, validation, and testing stages. Training is where the AI model learns from provided data, adjusting its internal parameters to identify patterns and relationships, often in an iterative process of prediction, error calculation, and parameter adjustment (Art. 3.29 EU AI Act). Validation evaluates the model's performance using a separate dataset to detect overfitting, adjust hyperparameters, and compare configurations (Art 3.30 EU AI Act). Finally, testing assesses the model's performance on an entirely new, independent dataset to provide an unbiased, realistic estimate of its real-world behavior and ensure it generalizes well without overfitting or underfitting (Arts. 3.31-32 EU AI Act). After these phases, the AI model is integrated with additional software components to become an AI system, which then undergoes conformity assessment, receives CE marking, and is accompanied by a declaration of conformity and instructions for use before being commercialized (Arts 3.20, 3.24, and 3.15 EU AI Act).

Post-implantation The post-implantation phase involves deploying the AI system into various environments, such as workplaces, to automate or semi-automate tasks. During this phase, the system receives "input data" (Art. 3.33 EU AI Act). It then processes this data through the "inference" process and provides "output data" (Arts. 3.31-32 EU AI Act) through content, predictions, recommendations, or decisions. AI systems require significant computational power for complex tasks, measured in FLOPs (Floating Point Operations Per Second) (Art. 3.67 EU AI Act). FLOPS quantify their mathematical operations per second and shed light on the needed resources to deploy the systems. The post-implantation phase concludes when the provider recalls the AI system (Art 3.16 EU AI Act) or withdraws from the market at the end of its operational life (Art. 3.17 EU AI Act).

5.4 Trustworthy AI Requirements, EU AI Act and Labor Law

To grasp the practical implications of Trustworthy AI, it is crucial to examine how its core requirements, which were originally developed as high-level ethical guidelines, interact with specific legal domains.

In this section, we focus on the seven requirements proposed by the High-Level Expert Group on AI (HLEG) and adopted in the EU AI Act, examining their intersection with Spanish labor law. In the employment context, this requires addressing structural power imbalances between employers and employees and the intricate obligations stemming from labor-specific legal protections.

This section's main goal is, therefore, to provide a schematic yet illustrative analysis of how each TAI requirement interacts with Spanish labor law.

Human Agency and Oversight TAI requires that AI systems be designed to allow appropriate human intervention and control (recital 27 EU AI Act).[23] In labor law, this implies that workers must be empowered to supervise and, when needed, override AI systems deployed in their environment. Employers are responsible for ensuring human oversight aligned with Art. 14 of the EU AI Act, which includes clear usage instructions and references to human control measures (recital 72; Annex IV EU AI Act). The supervisor must have the necessary competence, training, and authority (recital 73; Art. 26.2 EU AI Act). Training becomes a new regulatory requirement – "AI literacy" – that must be tailored to the type of AI system, its effects, and associated risks (recitals 20, 91, 3.1.56 and 4 EU AI Act). Specific authorizations may be needed, particularly in contexts involving significant risks [69]. Improper role assignment could trigger employer liability under the principle of *culpa in eligendo*.

Technical Robustness and Safety AI systems must resist internal and external failures (recital 27 EU AI Act).[24] Internally, feedback loops in learning systems may cause self-reinforcing biases (Art. 15.4 EU AI Act); externally, they may be vulnerable to cyber-attacks, including those initiated by other AI systems. In labor law, such systems qualify as "work equipment" [65] and are thus subject to occupational safety obligations under (Art. 17.1 Spanish LPRL). Both the employer, labor representatives, and the Labor and Social Security Inspectorate, or Labor Inspectorate (*Inspección de Trabajo y Seguridad Social*, ITSS) are entitled to halt system operation in case of grave and imminent risk (Art. 44 Spanish LPRL, Art. 19.5 Spanish TRET). Employers must establish protocols for safe shutdown and align them with evolving EU machinery and product safety regulations.

Privacy and Data Governance Systems must be developed in line with privacy and data protection regulations.[25] For labor, this includes complete application of the GDPR: workers are entitled to protection against automated decisions (Art. 22, EU AI Act), data protection by design (Art. 25, EU AI Act), and data impact assessments (Art. 35, EU AI Act). Compliance must extend to the Data Governance Act's requirements for data quality and integrity [159]. Current Spanish enforcement is fragmented in different supervisory agencies, such as the ITSS, AESIA, and AEPD, which share responsibilities (recitals 157, 168; Art. 99.7.c EU AI Act). This fragmentation complicates protection.[26]

[23] This requirement is referenced in recitals 27, 66, 72, 73, 91, 96 and articles 13.3.d, 14, paragraphs 2 and 3 of article 26.2, 27.1.e, and paragraphs 2.e and 3 of Annex IV (relating to compliance with the technical documentation obligations set out in Art. 11.1 of the EU AI Act).

[24] This requirement is referred to in recitals 27, 58.2.i, 66, 74, 75, 77, 122, articles 13.3.b.ii and 15, and section 2.g of Annex IV of the EU AI Act.

[25] This requirement is referenced in recitals 9, 10, 27, 28, 45, 48, 67, 69, 141, and articles 19, 26, 59, and section 5 of section C of Annex VIII of the EU AI Act.

[26] AESIA is the Spanish Agency for the Supervision of Artificial Intelligence (*Agencia Española de Supervisión de la Inteligencia Artificial*); AEPD is the Spanish Data Protection Agency (*Agencia Española de Protección de Datos*).

Transparency TAI systems must be traceable, explainable, and must disclose interaction with users (Recital 27, EU AI Act).[27] In the workplace:

- *Traceability* implies maintaining detailed activity logs (*e.g.*, logs of decisions on task allocation or performance), analogous to mandatory salary transparency tools (Art. 28.2-28.3 Spanish TRET or salary regulations [70]).
- *Awareness* requires that workers know when they are interacting with an AI system (*e.g.*, during recruitment filtering).
- *Capability disclosure* involves informing employers and workers of system limitations. Education on those capabilities must support this, so that workers can exercise their rights meaningfully. Transparency must also balance with intellectual property protections, especially for closed-source systems.

Diversity, Non-discrimination, and Fairness AI systems must prevent both direct and indirect discrimination, and promote inclusion (recital 27 EU AI Act).[28] In labor law, this principle connects to the constitutional principle of equality in the Art. 14 of the Spanish Constitution (*Constitución Españ ola*, **CE**) [62] and is challenged by the technical opacity of many AI systems. Systems may replicate or worsen inequalities if unexamined. At the same time, AI systems can support accessibility for workers with physical or cognitive disabilities, improving residual capacities. Social dialogue is crucial: collective bargaining and agreements should accompany AI deployment. Spain is already integrating AI clauses into labor agreements [82, 310].

Societal and Environmental Well-being This requirement promotes long-term monitoring of AI's social impact.[29] In labor law, it focuses on the impact of AI on employment structures, such as displacement, task redefinition, and job loss. It implies the need for policies such as unemployment benefits, retraining programs, and active labor market measures to guarantee fair transitions.

Accountability Accountability includes the auditability of systems and the presence of internal complaint mechanisms.[30] Worker representatives may request system audits if irregularities are suspected, under Art. 64.4.d) of the Spanish TRET. Whistleblower protections are grounded in recitals 150 and 165, and Art. 87 of the EU AI Act, and relate to European Directives [155]. Public-sector deployments must include stakeholder consultation, possibly including trade unions. Accountability mechanisms are essential for protecting labor rights in algorithmic environments.

For the sake of illustration, Table 10 presents a summary of the preceding analysis of the intersection of TAI requirements, EU AI Act, and Spanish labor law.

[27] This requirement is referenced in recitals 9, 26, 27, 53, 59, 66, 72, 101, 102, 104, 107, 132, 134, 137, 173, 174, and arts. 1.2.d), 13, 50, 53, and Annex XII of the EU AI Act.

[28] This requirement is referenced in recitals 27, 28, 31, 45, 48, 56, 57, 58, 80, 165 of the EU AI Act. Recitals 80 and 165 refer to persons with disabilities.

[29] This requirement is referenced in recitals 6 and 27 of the EU AI Act.

[30] This requirement is referenced in recitals 79, 85, 89, 91, 101, 125, 141, Art. 17.1(m) and 25 of the EU AI Act.

Table 10. Mapping Trustworthy AI Principles to Labor Law Implications

TAI Principle (HLEG)	**Labor Law Implication**	**EU AI Act References**
Human Agency & Oversight	Employer must ensure worker supervision, training, and override authority. Training duties ("AI literacy") and safety analogies apply.	Rec. 27, 66, 72, 73, 91, 96 EU AI Act; Art. 13.3.d, 14, 26.2, 27.1.e; Annex IV (2.e, 3)
Technical Robustness & Safety	AI systems are "work equipment"; malfunction protocols must exist. Employers/workers may halt use in case of risk.	Rec. 27, 58.2.i, 66, 74, 75, 77, 122; Art. 13.3.b.ii, 15.4; Annex IV (2.g)
Privacy & Data Governance	GDPR rights entirely apply. Oversight is fragmented (ITSS, AESIA, AEPD), reducing clarity and enforcement.	Rec. 9, 10, 27, 28, 45, 48, 67, 69, 141; Arts. 19, 26, 59; Annex VIII (C.5)
Transparency	Logging, system awareness, and capability disclosure are required. Worker training is critical.	Rec. 9, 26, 27, 53, 59, 66, 72, 101, 102, 104, 107, 132, 134, 137, 173, 174; Arts. 1.2.d, 13, 50, 53; Annex XII
Diversity, Non-discrimination & Fairness	Systems may entrench discrimination. Must include union participation and support inclusivity (*e.g.*, disability adaptation).	Rec. 27, 28, 31, 45, 48, 56–58, 80, 165
Social & Environmental Well-being	AI job impact requires fair transitions, reskilling, and labor integration policies.	Rec. 6, 27, 113
Accountability	Worker reps may audit AI. Whistleblower protections and internal grievance systems are required.	Rec. 79, 85, 89, 91, 96, 125, 150, 165; Art. 17.1(m), 25

5.5 Correlation vs Causation in Labor Decisions

As shown in the previous section, aligning AI systems with labor law entails adapting general regulatory requirements to the characteristics of each legal domain. Even within the Spanish labor market, such alignment involves navigating multiple overlapping legal instruments with different scopes and degrees of applicability.

To illustrate the concrete normative tensions that arise from this process, this section focuses on a specific use case: the impact of the *correlation-causation dilemma* in decisions delegated to AIWM systems. In particular, we examine how this technical limitation interacts with the legal requirement of sufficient reason for the labor decision and the broader prohibition of discriminatory treatment. This problem arises when delegating labor decisions that require the existence and sufficiency of legal causes, instead of correlations to AI systems that base their operation on correlations instead of causal relations.

Specifically, our aim is to understand how algorithmic management challenges two central legal safeguards in labor law: the need for existence and sufficient cause (positive dimension) and the non-existence of a cause or prohibition of discrimination (negative dimension).

To study these risks, we will first summarize how most AI systems function in decision-making and how their architecture limits causal reasoning. We will then outline the types of workplace decisions that legally require causality, distinguishing between their positive and negative dimension. Finally, we propose both technical and legal pathways to bridge this causality gap and mitigate its potential harms.

5.5.1 Understanding the Causality Gap in Algorithmic Systems

Most AI systems currently used in employment contexts are based on machine or deep learning models, which operate by detecting correlations and patterns that maximize predictive accuracy, rather than establishing causal relationships [257]. As a result, they may generate predictions, recommendations, or decisions that are not grounded in valid causal reasoning.

The distinction between correlation and causation is crucial in labor law, where many decisions, such as disciplinary dismissals or hiring, require a lawful and sufficient or non-discriminatory cause. Otherwise, the actions may be declared null and void [72]. These concerns are further exacerbated by other effects, such as the *black-box* nature of today's deep learning models, hampering the ability to trace or contest decisions [312].

For example, if an algorithm correlates social media use during working hours with poor performance and recommends disciplinary action, it could fail to recognize that both behaviors are symptoms of a third factor, such as workplace stress. The misidentification of the causes would make the decision legally unjustified.

Correlations might come from at least three different relationships between the variables:

1. **Direct causal relationship,** where one variable causes another ($A \rightarrow Y$), and such causality can be substantiated through domain knowledge and robust design. In this case, there would be an equivalence between correlation and causation.

2. **Correlations due to external factors.** These may arise from at least two reasons: (1) a shared confounding variable ($A \leftarrow C \rightarrow Y$), where a common cause (C) explains both the input (A) and the outcome (Y). The previously mentioned example about stress in the workplace, would be depicted as `use of social media` $\leftarrow$ `stress` $\rightarrow$ `low performance`; and (b) through a mediating variable or proxy ($A \rightarrow M \rightarrow Y$), where the proxy variable (M) transmits the effect of A on Y. For instance, a candidate's postal code may act as a proxy for socioeconomic status, education level, or ethnicity, indirectly leading to discrimination if used in hiring models [318]. An AI system that selects candidates based on education may inadvertently reproduce structural inequality without directly referencing the protected attribute.

3. **Spurious correlations:** Coincidental associations lacking any causal mechanism ($A \rightarrow C \leftarrow Y$), where conditioning on a shared effect (C) induces spurious dependence between unrelated variables. For example, a correlation between Christmas tree sales in the U.S. and the number of labor relations technicians in Hawaii [394]. Similarly, if both age and performance influence assignment to a prestigious task, and the AI model is trained only on data from those already assigned to that task, it may learn an artificial correlation between age and performance that does not generalize to the population.

Technical challenges to identify causality in AI models Beyond their reliance on statistical associations, other intrinsic technical properties of data-driven AI systems exacerbate the correlation-causation dilemma. These features, recognized in both the EU AI Act and national labor regulations [162], make it difficult to determine whether algorithmic decisions are based on a causal justification or are merely correlation-driven.

- **Complexity:** AI systems often comprise highly interconnected components that process complex datasets. This multi-element nature makes it difficult to identify which input features and data transformations influenced a decision, and whether this influence is legally relevant or coincidental.
- **Opacity:** Often referred to as the "black box" effect: many AI systems –especially those based on deep learning – produce outputs through internal processes that are not interpretable even to their designers. This lack of transparency obstructs accountability and legal scrutiny of the causal reasoning behind decisions [52]. It is worth mentioning that this technical property also overlaps with other requirements of Trustworthy AI. For instance, transparency is established as an essential requirement for reliable AI and is made effective in practice through the algorithmic information contemplated in Art. 64.4.d of the Spanish TRET. This information must be provided in cases where business decisions about the worker have been delegated to AI systems. In this case, the worker is subjected to decisions or proposed decisions generated by a machine whose operation is opaque, which leaves the recipient of these decisions defenseless, as they are unaware of the internal processes of the AI system. However, in practice, this requirement is difficult to fulfill meaningfully unless the system has been designed with interpretability in mind.
- **Adaptability and Probabilistic Nature:** Machine learning models are probabilistic, so that identical inputs may yield different outputs. This behavior might be exacerbated as the models are updated over time with new data. This behavior makes it difficult to verify whether a decision was based on a consistent and lawful rationale or simply shifting correlations [310].
- **Autonomous behavior:** When managerial decisions such as hiring, supervision, or dismissal are delegated to AI systems operating with limited human oversight, it becomes challenging to ensure that such decisions are supported by legally valid causal grounds [144].
- **Data dependence:** AI systems rely on large volumes of personal and non-personal data, which includes proxy variables that may inadvertently encode indirect correlations with protected attributes, potentially leading to discriminatory decisions even in the absence of explicit intent [148, 243].

In summary, the reliance on today's machine learning models on correlations, the consequent inability to guarantee causality, and their black-box nature raises substantial legal concerns. These issues are especially pressing in the labor domain, where invalid or discriminatory decisions can infringe fundamental rights and lead to legal nullification. Addressing this gap between technical practice and legal obligations requires interdisciplinary strategies, which we discuss in the remainder of the chapter.

5.5.2 Legal Dimensions of Causality in Algorithmic Labor Decisions

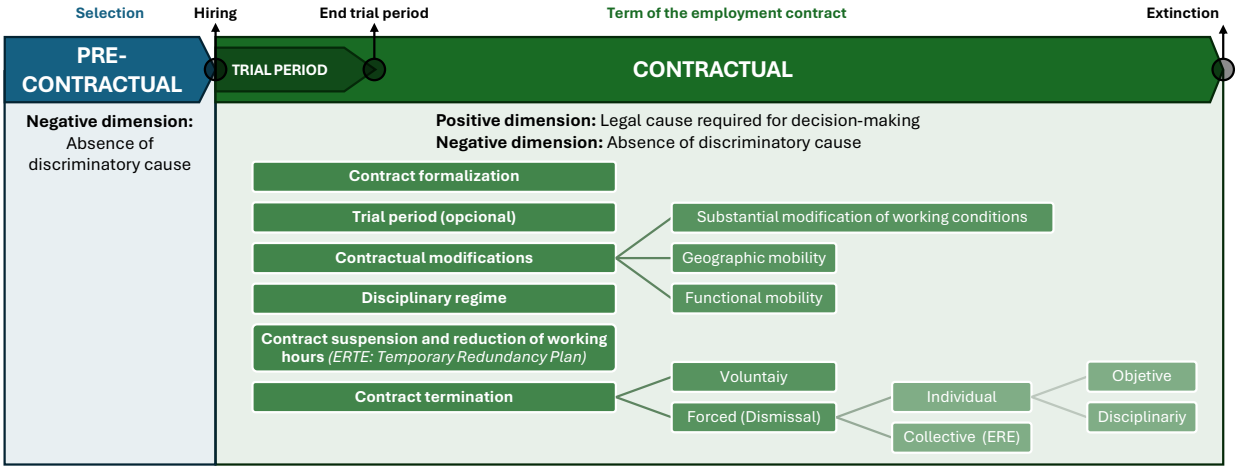


Figure 33. AIWM labor decisions through the life cycle of employment relationship.

In this section, we explore how two key dimensions –positive and negative, explained below– in labor law manifest in decisions related to employment, both in pre-contractual (*e.g.*selection and hiring), and post-contractual (*e.g.*employment management, modification, and termination) decisions. Figure 33 illustrates these decisions in the life cycle of the employment relationship. For clarity, we also provide illustrative examples for each case.

The two dimensions that we refer to are as follows:

- The **positive dimension** refers to the *principle of sufficient reason*. In the Spanish labor law, several managerial decisions –including but not limited to substantial modifications of working conditions, disciplinary sanctions, layoffs, or mobility decisions – must be justified by a demonstrable causal relationship between legally relevant circumstances and the decision taken. For instance, if condition X (*e.g.*, a proven organizational restructuring) occurs, then the employer is legally permitted to adopt decision Y (*e.g.*, terminate contracts or reassign tasks). Delegating such decisions to AI does not exempt the employer from this requirement.

- The **negative dimension** addresses the *prohibition of discriminatory motivations*. For instance, article 17.1 of the Spanish TRET [72] explicitly prohibits employment decisions based on protected attributes, such as gender, race, religion, or political beliefs. It also extends this prohibition to indirect discrimination and instructions to discriminate. This legal protection must apply to decisions made by algorithmic systems, even if the discrimination is indirect, occurring through proxy variables or learned biases in training data.

5.5.3 Pre-contractual Phase: Hiring and Selection

At the pre-contractual stage, employers in Spain can exercise their constitutional freedom to contract (Art. 38 Spanish Constitution (CE) [62]). However, this discretion is circumscribed

by constitutional and statutory commitments to non-discrimination (Art. 14 CE). Consequently, while there is no obligation to justify hiring or rejection through a lawful cause, any decision must still respect fundamental rights.

Positive Dimension Since labor law does not require causal justification for hiring decisions, algorithmic tools in this phase do not typically conflict with the positive dimension. The employer has discretion if decisions do not violate other legal principles.

Negative Dimension The legal concern lies in discriminatory outcomes arising from algorithmic profiling. AI systems may:

- Infer sensitive characteristics such as gender, religion, or ethnicity based on indirect markers like handwriting, zip codes, or lexical features (violating Art. 9 GDPR [153]).
- Make decisions correlated with seemingly neutral variables and protected attributes, thereby leading to discrimination (Art. 17.1 Spanish TRET).

Examples include Amazon's hiring tool, which penalized resumes that contained indicators of female identity, due to historical biases in training data [119]; and the HireVue system, which prioritized candidates with slower speech patterns – a characteristic found to lead to indirect discriminatory effects [351]. These violations can constitute severe administrative infractions under the Arts. 16.1.c and 8.12 of the Spanish TRLISOS [73], and may also lead to civil liability under the Spanish Civil Code.

5.5.4 Contractual Phase: Employment Decisions

The contractual phase involves a wide range of tightly regulated employer decisions. Many of these decisions –ranging from disciplinary actions to contractual modifications and dismissals– are contingent upon the existence of a legally sufficient cause and must not be motivated by discriminatory criteria.

Formalization of the Contract This refers to the act of executing the employment agreement. While the decision to hire falls under the pre-contractual phase, the contract formalizes the agreed-upon working conditions. This is not a purely managerial decision, yet the choice of contract type is not entirely discretionary. Most contracts are causal and must reflect actual productive needs.

Trial Period Spanish labor law includes a trial period. During this period, the employer can terminate the contract without justification. However, termination must not be based on discriminatory motives, creating a scenario where only the negative dimension is legally actionable.

Contractual Modifications Changes in employment conditions must comply with specific legal requirements.

- *Functional mobility* refers to changes in tasks assigned to the worker. When mobility exceeds the limits of the job classification, a technical or organizational cause is needed (Art. 39.2 Spanish TRET).

- *Geographical mobility* and *substantial modification of the working conditions* require justification under Arts. 40 and 41 Spanish TRET. These include economic, technical, organizational, or production-related reasons.

Decisions of this kind are subject to judicial review and may be challenged in court.

Disciplinary Regime Employer disciplinary powers (Art. 58 Spanish TRET) must be exercised within the boundaries set by legal and contractual norms. Misconduct must be typified, and sanctions must be proportionate. While the law does not explicitly use the term "cause", it implicitly requires factual justification.

Suspension or Reduction (*ERTE*, for its initials in Spanish) Suspensions of contracts or reductions in working hours must be justified by causes such as economic downturn, *force majeure*, or organizational change. The involvement of AI in selecting affected employees poses risks under both the negative and the positive dimensions. The Spanish ITSS must verify no discrimination in the selection process (Art. 22 Spanish Royal Decree on collective dismissal procedures, contract suspension and working time reduction [66]).

Dismissal and Termination Terminating the employment relationship is Spain's most sensitive and regulated decision.

- *Objective and disciplinary dismissals* must be causally justified. The burden of proof lies with the employer, and the decision must be documented following the Arts. 53 and 54 Spanish TRET.

- *Collective dismissals* (ERE, for its initials in Spanish) require cause under Art. 51.1 Spanish TRET. The Spanish Labor Authority and ITSS must evaluate both the sufficiency of the cause and the presence of any discriminatory selection criteria.

Judges can declare a dismissal to be *fair*, *unfair*, or *null and void* depending on whether the legal cause exists and whether any rights have been violated. Again, in this, the contractual phase, we differentiate between the two dimensions:

- *Positive Dimension*: Many of the previously described employer decisions require a legal cause. The employer must articulate the cause and prove its existence and proportionality. Delegating such assessments to AI systems creates significant evidentiary challenges. Moreover, the opacity and lack of explainability of many AI models hinder judicial oversight. Illustrative examples include Xsolla's mass layoffs conducted based solely on an AI system's productivity scores [8, 139], and Amazon's reported use of AI to dismiss workers with low performance metrics automatically [351].

- *Negative Dimension*: Discrimination remains a persistent risk in algorithmic management. Systems may implicitly learn biases present in historical data or introduce new discriminatory patterns through proxies. For instance, the use by Uber of customer rating scores to trigger disciplinary actions has been found to lead to disparate impacts on racialized and marginalized workers [335].

In sum, the dual obligations of lawful justification and non-discrimination are foundational to Spanish labor law. Delegating sensitive decisions to AI systems disrupts this legal architecture. While the promise of AI includes efficiency and objectivity, the correlation-based nature and lack of transparency of today's deep learning models create regulatory blind spots. These challenges threaten not only workers' rights but also legal certainty for employers and institutional capacity for oversight.

5.5.5 Proposed Solutions

This section proposes technical and legal solutions to the previously described challenges.

1. Technical Guidelines

AIWM systems can leverage various technical tools and methods to mitigate bias and opacity. These tools facilitate greater transparency and understanding about how decisions are made, enabling better human oversight, and eliminating or at least reducing uncertainty about the non-existence or existence and sufficiency of legal cause in algorithmic management. Among others, we highlight the following technical approaches:

Algorithmic Fairness. To address the challenges related to the negative dimension, algorithmic fairness methods help audit and mitigate discrimination when using algorithms to make decisions [42].

Automated decision support systems (ADSS). These systems consider effective human-machine collaboration, avoiding full automation of the decision [92].

Causal inference. Causal inference methods mathematically model and quantify cause-and-effect relationships [318]. They could play a crucial role in understanding the existence of causal relationships in automated decisions [57]. However, their application requires deep domain knowledge and the ability to interpret results, among other limitations.

Interpretable AI and explainable AI (XAI). To address the opacity of AI models and facilitate their explainability, there are two main approaches:

- *Interpretable AI*, which aims to create AI models designed with an internal decision-making process that is understandable to humans [295]. Interpretable AI models are often less effective at performing their tasks than non-interpretable models;
- *XAI* aims to make black-box AI models more understandable by humans by providing explanations about their decision-making process, without imposing restrictions on the model itself. XAI methods can provide global (how the models make decisions at a general level) or local (how the models make decisions for a specific case) explanations about how the AI system works [295]. However, it often has significant limitations due to the low reliability and robustness of the explanations, as it is challenging to know which variables influenced the decision. Furthermore, explanations are always imperfect approximations of the internal decision-making processes of black-box models, and there is currently no clear or agreed-upon way to evaluate them.

2. Legal Guidelines

From a legal perspective, we propose three potential solutions:

Avoid full automation Only allow the automation of non-causal decisions with semi-automation of causal decisions. Note that AI systems can automate tasks (by making decisions) or semi-automate tasks (by assisting humans who make the decisions). Given the previously explained limitations of today's AI systems, it is logical to consider that labor decisions that require a concurrent cause should not be automated under any circumstances. However, AI systems could assist the human controller, who must ensure the existence and sufficiency of a legal cause and the absence of a discriminatory cause in decision-making.

AI literacy (ex. Art. 3.56, EU AI Act) Another necessary solution consists of AI literacy efforts for workers, employers, and officials in the judiciary and Spanish ITSS. Regarding the latter, we have given examples where labor officials can rule on the positive and negative dimensions of the correlation-causality dilemma. This requires adequate technical and legal knowledge; otherwise, risks to legally recognized rights and interests could arise. Strategic Axis 5 of the Spanish National Artificial Intelligence Strategy provides this technical training in new technologies.

Legislative reform. Finally, although this study has sought to present concrete, practical, and logical solutions to the problems raised, it is also essential to emphasize that in Labor Law, the most prudent and effective long-term solution is constantly developing an understanding between the social partners –the true protagonists of the entire construction of labor law– especially when it is embodied in a legal norm. Therefore, we highlight the importance of future legal regulation that addresses the issues raised here.

5.6 Conclusion and Future Work

In this chapter, first we analyzed the regulations that apply to AI for labor systems when used in the context of labor decisions under Spanish labor law. For simplicity, we summary them in Table 11 identifying the Spanish or European scope. Second, we have examined how they intersect with the seven requirements of Trustworthy AI, and we have described the correlation vs causation dilemma, which creates a misalignment between the technical properties of AI models and legal requirements. Finally, we examined which employment life cycle decisions require lawful cause or absence of discriminatory motivation, proposing technical and regulatory strategies to address these tensions.

More generally, this chapter demonstrates that effective AI regulation requires sociotechnical alignment. As emphasized throughout this dissertation, lawful AI systems must bridge the gap between technical design and legal norms. Although the legal and technical domains operate according to different logics, a joint understanding is necessary to ensure compliance. Robust safeguards, such as documentation, transparency mechanisms, and clear legal responsibilities, are critical for the lawful deployment of AI.

Additionally, applying the TAI requirements and derived regulations must be tailored to specific sectors' characteristics and legal demands, such as labor (other socially relevant examples include health and autonomous vehicles).

However, further research is needed to explore additional forms of misalignment between technical architectures and normative legal standards in high-risk domains.

A logical next step would be to investigate how the requirements of Trustworthy AI reflected in the EU AI Act could be implemented through standardization efforts, with some initiatives already moving in this direction, such as AI-related standards [219, 220, 355] and the Code of Practice of the General Purpose AI Models [152]. This would also involve developing sector-specific compliance mechanisms and indicators translating ethical and legal requirements into technical actions.

Finally, we recommend further studying institutional safeguards, such as the role of unions and collective worker representation, to monitor and challenge the use of AIWM systems. Ensuring the correct application of labor law in hybrid or AI decision-making systems will require technological transparency and participatory governance structures.

Table 11. Legal instruments applicable to AI in labor contexts categorized by origin

	Legal Instruments	Labor Law Relevance
EU	*AI Act* (Regulation (EU) 2024/1689) [162]	Establishes obligations and transparency requirements for high-risk AI systems, including those used in employment.
	Digital Services Act (Regulation (EU) 2022/2065) [158]	Imposes transparency and audit obligations for online platforms, relevant when used in hiring or workplace monitoring.
	GDPR (Regulation (EU) 2016/679) [153]	Ensures data protection rights, especially regarding automated decisions (Art. 22) and impact assessments (Art. 35).
	Data Governance Act (Regulation (EU) 2022/868) [159]	Regulates the sharing and reuse of data, including sensitive employment-related datasets.
	Data Act (Regulation (EU) 2023/2854) [154]	Establishes fair access and use of data generated by users and devices, which may include workplace data.
	Machinery Regulation (Regulation (EU) 2023/1230) [160]	Applies to AI-enabled machinery and safety systems used in workplaces.
	Product Safety Regulation (Regulation (EU) 2023/988) [161]	Covers consumer-facing AI systems and general safety provisions in products affecting workers.
	Proposed AI and Product Liability Directives [150, 151]	Extend civil liability to providers and users of AI systems when harm occurs, including in employment settings.
Spain	TRET (Real Decreto Legislativo 2/2015) [72]	Governs employment relationships and ensures rights to information, oversight, and non-discrimination in AI-based decisions.
	Ley de Empleo (Ley 3/2023) [63]	Regulates employment intermediation, including algorithmic platforms.
	LPRL (Ley 31/1995) [64]	Establishes obligations to prevent occupational risks, applicable to AI systems that affect workers' physical or psychosocial health.
	TRLISOS (Real Decreto Legislativo 5/2000) [73]	Defines labor law infractions and sanctions, including employer misuse of automated systems.
	RD 902/2020 [70]	Regulates pay transparency and audits; relevant for algorithmic compensation systems.
	RD 1561/1995 [67]	Covers special work schedules, applicable to AI systems assigning shifts or managing hours.
	RD 2001/1983 [68]	Regulates general and special work time arrangements, which may be affected by algorithmic scheduling.
	RD 1215/1997 [65]	Classifies AI systems as "work equipment"; imposes safety obligations.
	RD 614/2001 [69]	Applies to AI systems presenting electrical hazards; relevant in automated industrial settings.
	RDL 1/2007 [71]	Covers consumer and user protection, potentially applicable to dual-use AI systems in labor.
	RD 1483/2012 [66]	Regulates collective dismissals, contract suspensions, working time reductions, algorithmic productivity evaluations, or risk-based scoring.

Part III

Conclusion

Chapter 6

Conclusion and Open Questions

6.1 Overview of Findings

This thesis proposes a sociotechnical framework based on the principles of Trustworthy AI and considering three interconnected spheres: *technical* design, human *use*, and *governance*. Motivated by the growing use of AI systems in high-risk domains –such as education, healthcare, credit scoring, hiring, and the spread of information – our goal has been to systematically identify, measure, and mitigate the harms induced by these systems. The thesis argues that the effective implementation of TAI cannot rely solely on technical solutions; it must also be aligned with human behavior and institutional constraints.

Within the **technical sphere**, while fairness metrics are often applied to evaluate models, they offer limited insight into the origin of disparities. To address this limitation, we introduced `FairShap` (Chapter 2), a novel data valuation method that quantifies the influence of individual training examples on group-level fairness metrics. This contribution provides interpretable and actionable tools for mitigating discrimination in high-risk decision-making systems, and it is aligned with the EU AI Act, which mandates fairness auditing and non-discrimination in high-risk systems.

In Chapter 3, we shifted our focus to fairness in social graphs, where risks include marginalization, polarization, and information exposure disparities. Motivated by the lack of consensus around operational definitions for these harms, we proposed the notion of *structural group unfairness*, measured by a family of graph-based metrics that capture disparities in access and visibility in social networks. From a sociotechnical perspective, the relevance of this contribution lies in its alignment with the European Digital Services Act, which imposes legal obligations on platforms to assess and reduce systemic risks.

With respect to the **use sphere**, we proposed `CHAIM` in Chapter 4, a novel hybrid decision making system that combine algorithmic predictions with human judgment. While full automation poses risks of opacity and over-reliance, human-only systems often suffer from inefficiency and inconsistency. `CHAIM` addresses these challenges of human-AI complementarity system in resource allocation tasks, by leveraging bandit-based strategies to optimize the allocation of tasks between the algorithm and the human. We ran a user study with 800 participants to collect data about how humans perform matching tasks, which was later used to evaluate `CHAIM`'s performance. Our experimental results highlight the risks of over- or under-reliance on algorithmic support and demonstrate the potential of hybrid approaches that allocate decision authority dynamically. `CHAIM` aligns with technical robustness and hu-

man oversight requirements in semi-automated systems, as encouraged by European GDPR provisions and TAI principles.

Finally, in the **governance sphere**, Chapter 5 examined the intersection between the technical and regulatory dimensions of TAI with Spanish labor law in the context of AI systems that are used for worker management. First, it identified the relevant regulations that apply to the use of AI systems across the AI System life-cycle. Next, it provided an analysis of the alignment between the seven TAI requirements and the Spanish labor law. Finally, potential normative tensions between AI models and regulations were highlighted, and particularly the correlation-causation dilemma: while AI systems often learn from correlations, legal standards require causality, justification and non-discrimination in certain decision-making processes. The chapter also included a set of technical and legal guidelines to develop solutions to the identified challenges.

Together, the contributions to this thesis demonstrate that Trustworthy AI is not a unidimensional problem, but rather a sociotechnical challenge spanning algorithmic performance, human factors, and institutional design. Each chapter addresses a different aspect of this issue, yet their coherence lies in their **shared commitment to uncovering how harms arise and can be mitigated at various stages of the AI life-cycle**.

Ultimately, we argue that the TAI principles must be translated into measurable practices, manifested in technical design, grounded in the law, and embedded in institutional governance. This work lays the groundwork for such a goal.

6.2 Conclusion and Open Questions

Algorithmic Fairness AI systems are increasingly robust in terms of non-discrimination in decision-making, with state-of-the-art methods proving surprisingly competitive [114]. Thus, future research should focus not only on improving accuracy, but on understanding and shaping the interactions between data, models, and societal outcomes, as shown in Chapter 2.

One promising direction is to explore methods that enhance fairness *without explicitly using demographic attributes*, particularly relevant in scenarios where access to sensitive data is limited, but technical safe-guards are required [254]. In addition, we encourage further study of fairness under alternative tensions and synergies, such as privacy, explainability or efficiency [57, 115, 116]. Such an analysis could improve technical trade-offs and clarify the practical and normative boundaries of fair AI deployment.

Systemic Risks and Complex Harms While fairness in decision-making has well-established metrics and benchmarks [220, 288], other socially-defined harms derived from the use of AI lack common mathematical formalizations, as addressed in Chapter 3.

This lack of consensus around the mathematical formalization of certain harms may stem from multiple sources. In some cases, the underlying social problem is not yet clearly defined or agreed upon, requiring further interdisciplinary dialogue and conceptual development. In others, the harm is understood, but its formalization involves significant mathematical or computational complexity that challenges current modeling techniques. For example, in the context of social networks, widely used architectures such as GNNs are built upon theoretical foundations that have recently come under critical scrutiny [20]. Clarifying these conceptual ambiguities is an essential step toward developing more expressive models capable of capturing and mitigating systemic harms.

The development of consensus metrics and mitigation methods for these harms remains an open area of research – particularly as these challenges extend to General Purpose AI (GPAI) models [52, 152], where identifying and operationalizing systemic risks is even more complex.

Pluralism in Fairness and Risk Notions It is important to recognize that fairness and harm are social constructs whose technical formulations are always simplifications. The commonly used mathematical definitions of fairness from the machine learning literature, such as demographic parity or equal opportunity, are not neutral. Rather, they are formal proxies designed to reflect normative goals in specific application contexts [94].

First, operationalizing such fairness notions inevitably requires compromises, such as trade-offs between expressiveness and tractability or context-sensitivity and generality. As a result, technical fairness metrics can only approximate the broader, often contested, societal meanings of justice, equity, or harm.

Second, as these definitions are socially rooted, their interpretation varies across legal jurisdictions, cultural contexts, and affected communities. What constitutes discrimination or harm may differ between regulatory environments or stakeholder groups.

Consequently, a key area for future research is the development of systems capable of accommodating multiple, potentially conflicting conceptions of fairness or harm. This could entail designing configurable objectives, participatory modeling interfaces or multi-objective optimization frameworks that allow for legal, cultural or organizational variation. Supporting pluralism in technical design is ultimately a necessary step towards aligning AI systems with the diverse societies in which they operate.

Regulatory and Institutional Alignment Developing *efficient AI regulation* in the labor market and other high-risk domains requires translating abstract principles into enforceable standards. Some promising efforts are already underway, including the emergence of AI-specific standardization bodies and codes of practice. For example, the CEN-CENELEC Joint Technical Committee 21 (JTC 21) plays a key role in implementing the EU AI Act through technical standardization.[31]

However, standardization must go beyond general guidelines. Future work should explore sector-specific compliance mechanisms, context-sensitive indicators, and cross-disciplinary tools to translate ethical and legal requirements into practical safeguards.

Moreover, legal scholars should engage with the technical realities of AI systems to propose regulatory modifications, resolve inter-normative tensions, and adapt existing laws to new risks — as exemplified in the correlation-causation dilemma discussed in Chapter 5.

Finally, institutional safeguards are critical. As algorithmic management becomes more prevalent, the role of unions and worker representation in overseeing AI systems must be rethought. Ensuring labor protections in hybrid or automated systems requires both technical transparency and participatory governance structures.

Automating Alignment with TAI Principles and the AI Act An important open question is whether the alignment of AI systems with TAI principles, EU AI Act requirements, and technical standards can be fully or partially automated. Such automation would

[31] https://jtc21.eu/

enhance auditability for internal reviews, conformity processes, and oversight by national authorities.

From a technical perspective, this would require the formalization of regulatory criteria into machine-readable formats, such as the encoding of fairness constraints, or the generation of standardized documentation, such as model cards and datasheets [152]. Although full automation is unlikely due to the context-sensitive and often vague nature of legal norms, semi-automated pipelines could identify compliance risks and streamline assessments.

This line of research could bridge the gap between regulation and implementation, reducing compliance burdens while enabling more consistent and transparent audits.
Taken together, these open questions point toward a rich research agenda at the intersection of AI design, institutional governance, and regulatory practice.

In conclusion, this thesis demonstrates that bridging the gap between the principles of Trustworthy AI and its implementation is not merely a technical challenge, but rather a multi-layered sociotechnical task. Although each contribution focused on a particular area, they are all aligned and collectively demonstrate that Trustworthy AI requires more than just correctness; it necessitates collaboration.

Part IV

Appendix of Core Publications

Appendix A

Appendix of Chapter 2

A.1 Notation

Symbol	Description
$\mathcal{D} = \{(x_i, y_i)\}_{i=0}^{n}$	Training dataset
$\mathcal{T} = \{(x_j, y_j)\}_{i=0}^{m}$	Reference dataset
$S \subseteq \mathcal{D}$	Subset of a dataset $\mathcal{D}$
A	Set of variables that are protected attributes.
$\text{TPR}_{A=a}$	True positive rate for test points with values in the protected attribute equal to a. Also TPR_a if the protected attribute is known. The same logic applies to FPR, TNR, and FNR.
$p(y\|x, \mathcal{D})$	Predictive distribution of data point x when trained with $\mathcal{D}$.
$p(y = y_j\|x_j, \mathcal{D})$	Likelihood of correct classification of data point x when trained with $\mathcal{D}$.
$\phi_i(\mathcal{D}, v)$	Shapley Value for data point i in the training dataset $\mathcal{D}$ according to the performance function v
$\boldsymbol{\phi}(\mathcal{D}, v)$	Vector with all the SVs of the entire dataset $\in \mathbb{R}^{\lvert\mathcal{D}\rvert}$.
$v(S, T)$	Value of dataset S w.r.t a reference dataset $\mathcal{T}$. *e.g.*, the accuracy of a model trained with S tested on $\mathcal{T}$ ($v = \text{Acc}$) or the value of Equal Opportunity of of a model trained with S tested on $\mathcal{T}$ ($v = \text{EOp}$)
$\boldsymbol{\Phi} \in \mathbb{R}^{\lvert\mathcal{D}\rvert \times \lvert\mathcal{T}\rvert}$	Matrix where $\Phi_{i,j}$ is the contribution of the training point $i \in \mathcal{D}$ to the correct classification of $j \in \mathcal{T}$ according to Appendix A.3.1
$\overline{\boldsymbol{\Phi}}_{i,:}$	Mean of row i
$\overline{\boldsymbol{\Phi}}_{i,:\|A=a}$	Mean of row i conditioned to columns where $A = a$
$\mathbf{1}$	Vector of ones $:= [1, 1, ..., 1]$

Table 12. Table of Notation

A.2 Shapley Values Proposed in FairShap

FairShap proposes $\phi(\text{EOp})$ and $\phi(\text{EOdds})$ as the data valuation functions for fairness. These functions are computed from group-specific $\phi(\text{TPR}), \phi(\text{FPR}), \phi(\text{TNR})$ and $\phi(\text{FNR})$ functions, leveraging the Efficiency axiom of the SVs, and the decomposability properties of fairness metrics.

First, using our proposed formalization of the pairwise contribution of a training data point to the correct classification of a test datapoint defined as

$$\Phi_{i,j} = \mathbb{E}_{S\sim\mathcal{P}(\mathcal{D}\setminus\{i\})}[\underbrace{p(y=y_j|x_j, S\cup\{i\}) - p(y=y_j|x_j,S)}_{\text{LOO}(i,j,S)}],$$

the Shapley value for accuracy [223] can be redefined as

$$\phi_i(\text{Acc}) := \frac{1}{|\mathcal{T}|}\sum_{j\in\mathcal{T}}\Phi_{i,j} = \mathbb{E}_{j\sim p(\mathcal{T})}[\Phi_{i,j}].$$

In the following, we summarize all proposed valuation functions presented in this work:

True/False Positive/Negative rates:

$$\begin{aligned}\phi_i(\text{TPR}) &:= \mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=1)}\left[\mathop{\mathbb{E}}_{S\sim\mathcal{P}(\mathcal{D}\setminus\{i\})}[p(y=1|x_j,S\cup\{i\}) - p(y=1|x_j,S)]\right]\\ &= \mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=1)}[\Phi_{i,j}] = \frac{\sum_{j\in\mathcal{T}}\Phi_{i,j}\mathbb{I}[y_j=1]}{|\{x: x\in\mathcal{T}\,|y=1\}|}\end{aligned}$$

$$\begin{aligned}\phi_i(\text{TNR}) &:= \mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=0)}\left[\mathop{\mathbb{E}}_{S\sim\mathcal{P}(\mathcal{D}\setminus\{i\})}[p(y=0|x_j,S\cup\{i\}) - p(y=0|x_j,S)]\right]\\ &= \mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=0)}[\Phi_{i,j}] = \frac{\sum_{j\in\mathcal{T}}\Phi_{i,j}\mathbb{I}[y_j=0]}{|\{x: x\in\mathcal{T}\,|y=0\}|}\end{aligned}$$

$$\begin{aligned}\phi_i(\text{FNR}) &:= \mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=1)}\left[\mathop{\mathbb{E}}_{S\sim\mathcal{P}(\mathcal{D}\setminus\{i\})}[p(y=0|x_j,S\cup\{i\}) - p(y=0|x_j,S)]\right]\\ &= \frac{1}{|\mathcal{D}|} - \phi_i(\text{TPR})\end{aligned}$$

$$\begin{aligned}\phi_i(\text{FPR}) &:= \mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=0)}\left[\mathop{\mathbb{E}}_{S\sim\mathcal{P}(\mathcal{D}\setminus\{i\})}[p(y=1|x_j,S\cup\{i\}) - p(y=1|x_j,S)]\right]\\ &= \frac{1}{|\mathcal{D}|} - \phi_i(\text{TNR})\end{aligned}$$

where to compute $\phi_i(\text{FNR})$ we use the following equality:

$$\text{TPR} = 1 - \text{FNR} \rightarrow \sum \phi_i(\text{TPR}) = 1 - \sum \phi_i(\text{FNR}) \rightarrow \sum \phi_i(\text{TPR}) = \sum 1/n - \phi_i(\text{FNR})$$
$$\rightarrow \phi_i(\text{TPR}) = 1/n - \phi_i(\text{FNR}) \rightarrow \phi_i(\text{FNR}) = 1/n - \phi_i(\text{TPR}).$$

Conditioned True/False Positive/Negative rates:

$\phi_i(\text{TPR}_a) := \mathbb{E}_{j\sim p(\mathcal{T}|Y=1,A=a)}[\Phi_{i,j}] = \frac{\sum_{j\in\mathcal{T}} \Phi_{i,j}\mathbb{I}[y_j=1,A_j=a]}{|\{x:x\in\mathcal{T}|y=1,A=a\}|} = \overline{\mathbf{\Phi}}_{i,:|Y=1,A=a}$

$\phi_i(\text{TNR}_a) := \mathbb{E}_{j\sim p(\mathcal{T}|Y=0,A=a)}[\Phi_{i,j}] = \frac{\sum_{j\in\mathcal{T}} \Phi_{i,j}\mathbb{I}[y_j=0,A_j=a]}{|\{x:x\in\mathcal{T}|y=0,A=a\}|} = \overline{\mathbf{\Phi}}_{i,:|Y=0,A=a}$

$\phi_i(\text{FPR}_a) := \frac{1}{|\mathcal{D}|} - \phi_i(\text{TNR}_a)$

$\phi_i(\text{FNR}_a) := \frac{1}{|\mathcal{D}|} - \phi_i(\text{TPR}_a)$

When A≠Y:

$$\begin{aligned}\phi_i(\text{EOp}) :=& \phi_i(\text{TPR}_a) - \phi_i(\text{TPR}_b) \\ =& \mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=1,A=a)}[\Phi_{i,j}] - \mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=1,A=b)}[\Phi_{i,j}] \qquad (25)\\ =& \mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=1,A=a)}\left[\mathop{\mathbb{E}}_{S\sim\mathcal{P}(\mathcal{D}\setminus\{i\})}[p(y=1|x_j,\mathcal{D}\cup\{i\}) - p(y=1|x_j,\mathcal{D})]\right] \\ &- \mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=1,A=b)}\left[\mathop{\mathbb{E}}_{S\sim\mathcal{P}(\mathcal{D}\setminus\{i\})}[p(y=1|x_j,\mathcal{D}\cup\{i\}) - p(y=1|x_j,\mathcal{D})]\right]\end{aligned}$$

$$\phi_i(\text{EOdds}) := \frac{1}{2}\left((\phi_i(\text{FPR}_a) - \phi_i(\text{FPR}_b)) + (\phi_i(\text{TPR}_a) - \phi_i(\text{TPR}_b))\right) \qquad (26)$$

$$\begin{aligned}=& \frac{1}{2}\left(\left(\left(\frac{1}{|\mathcal{D}|} - \phi_i(\text{TNR}_a)\right) - \left(\frac{1}{|\mathcal{D}|} - \phi_i(\text{TNR}_b)\right)\right)\right. \qquad (27)\\ &\left. + \left(\mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=1,A=a)}[\Phi_{i,j}] - \mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=1,A=b)}[\Phi_{i,j}]\right)\right) \\ =& \frac{1}{2}\left(\left(\left(\frac{1}{|\mathcal{D}|} - \mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=0,A=a)}[\Phi_{i,j}]\right) - \left(\frac{1}{|\mathcal{D}|} - \mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=0,A=b)}[\Phi_{i,j}]\right)\right)\right. \\ &\left. + \left(\mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=1,A=a)}[\Phi_{i,j}] - \mathop{\mathbb{E}}_{j\sim p(\mathcal{T}|Y=1,A=b)}[\Phi_{i,j}]\right)\right)\end{aligned}$$

We refer to the reader to Appendix A.3.2 for more details on how to obtain these formulas from the algorithmic fairness definitions.

When A=Y:

$\phi_i(\text{EOp}) := \phi_i(\text{EOp}) = \phi_i(\text{TPR}) + \phi_i(\text{TNR}) - \frac{1}{|\mathcal{D}|}$

or its bounded version $\phi_i(\text{EOp}) = \frac{\phi_i(\text{TPR})+\phi_i(\text{TNR})}{2}$.

See Appendix A.3.3 for more details on how to derive these formulas.

A.3 Methodology

A.3.1 Efficient k-NN Shapley Value

Jia et al. [223] propose an efficient, exact calculation of the Shapley Values by means of a recursive k-NN algorithm with complexity $O(N \log N)$. The proposed method yields a

matrix $\mathbf{\Phi} \in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{T}|}$ with the contribution of each training point to the accuracy of each point in the reference data set $\mathcal{T}$. Therefore, $\Phi_{i,j}$ defines how much data point i in the training set contributes to the probability of correct classification of data point j in $\mathcal{T}$. The intuition behind is that $\Phi_{i,j}$ quantifies to which degree a training point i helps in the correct classification of j. The k-NN-based recursive calculation is at follows.

For each j in $\mathcal{T}$:

- Order $i \in \mathcal{D}$ according to the distance to $j \in \mathcal{T} \rightarrow (x_1, x_2, ..., x_N)$
- Calculate $\Phi_{i,j}$ recursively, starting from the furthest point:

$$\Phi_{N,j} = \frac{I[y_{x_N} = y_j]}{N}$$

$$\Phi_{i,j} = \Phi_{i+1,j} + \frac{I[y_i = y_j] - I[y_{i+1} = y_j]}{\max K, i}$$

- $\mathbf{\Phi}$ is a $|\mathcal{D}| \times |\mathcal{T}|$ matrix given by:

$$\mathbf{\Phi} = \begin{bmatrix} \Phi_{00} & \cdots & \Phi_{0|\mathcal{T}|} \\ \vdots & \ddots & \vdots \\ \Phi_{|\mathcal{D}|0} & \cdots & \Phi_{|\mathcal{D}||\mathcal{T}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{T}|}$$

where $\Phi_{i,j}$ is the contribution of training point i to the accuracy of the model on point j in $\mathcal{T}$. Thus, the overall SV of a training point i with respect to $\mathcal{T}$ is the average of all the values of row i in the SV matrix:

$$\phi_i(\text{Acc}) = \frac{1}{m}\sum_{j=0}^{m} \Phi_{i,j} = \overline{\mathbf{\Phi}}_{i,:} \in \mathbb{R}$$

Note that the mean of a column j in $\mathbf{\Phi}$ is the accuracy of the model on that test point. The vector with the SV of every training data point is computed as:

$$\boldsymbol{\phi}(\text{Acc}) = [\phi_0, \cdots, \phi_n] \in \mathbb{R}^{|\mathcal{D}|}$$

In addition, given the efficiency axiom of the Shapley Value, the sum of $\boldsymbol{\phi}$ is the accuracy of the model on the training set.

$$V(\mathcal{D}) = \sum_{i=0}^{n} \phi_i = \sum_{i=0}^{n} \frac{1}{m} \sum_{j=0}^{m} \Phi_{i,j} = \text{Acc}$$

Technically speaking, the process may be parallelized over all points in $\mathcal{T}$ (columns of the matrix) since the computation is independent, reducing the practical complexity from $O(N \log N)$ to $O(N)$.

A.3.2 $\phi_i(\textbf{EOp})$ and $\phi_i(\textbf{EOdds})$ derivation when $A \neq Y$

We derive $\boldsymbol{\phi}(\text{EOp})$ and $\boldsymbol{\phi}(\text{EOdds})$ when $A \neq Y$ using the definitions for EOdds and EOp given by:

$$\text{EOp} = \text{TPR}_{A=a} - \text{TPR}_{A=b}$$
$$\text{EOdds} = \frac{1}{2}((\text{TPR}_{A=a} - \text{TPR}_{A=b}) + (\text{FPR}_{A=a} - \text{FPR}_{A=b}))$$

Leveraging the Efficiency property of SVs, $\boldsymbol{\phi}(\text{EOp})$ is computed as:

$$\text{EOp} = \sum_{i\in\mathcal{D}} \phi_i(\text{TPR}_{A=a}) - \sum_{i\in\mathcal{D}} \phi_i(\text{TPR}_{A=b})$$
$$\text{EOp} = \sum_{i\in\mathcal{D}} (\phi_i(\text{TPR}_{A=a}) - \phi_i(\text{TPR}_{A=b})) \rightarrow \phi_i(\text{EOp}) = \phi_i(\text{TPR}_{A=a}) - \phi_i(\text{TPR}_{A=b})$$

Similarly, $\boldsymbol{\phi}(\text{EOdds})$ can be obtained as follows:

$$\begin{aligned}
\text{EOdds} &= \frac{1}{2}\Big((\text{TPR}_{A=a} - \text{TPR}_{A=b}) + (\text{FPR}_{A=a} - \text{FPR}_{A=b})\Big) \\
&= \frac{\left(\sum_{i\in\mathcal{D}} \phi_i(\text{TPR}_{A=a}) - \sum_{i\in\mathcal{D}} \phi_i(\text{TPR}_{A=b})\right) + \left(\sum_{i\in\mathcal{D}} \phi_i(\text{FPR}_{A=a}) - \sum_{i\in\mathcal{D}} \phi_i(\text{FPR}_{A=b})\right)}{2} \\
&= \frac{\sum_{i\in\mathcal{D}}\Big(\phi_i(\text{TPR}_{A=a}) - \phi_i(\text{TPR}_{A=b})\Big) + \sum_{i\in\mathcal{D}}\Big(\phi_i(\text{FPR}_{A=a}) - \phi_i(\text{FPR}_{A=b})\Big)}{2} \\
&= \frac{\sum_{i\in\mathcal{D}}\Big(\big(\phi_i(\text{TPR}_{A=a}) - \phi_i(\text{TPR}_{A=b})\big) + \big(\phi_i(\text{FPR}_{A=a}) - \phi_i(\text{FPR}_{A=b})\big)\Big)}{2} \\
&\rightarrow \phi_i(\text{EOdds}) = \frac{\big(\phi_i(\text{TPR}_{A=a}) - \phi_i(\text{TPR}_{A=b})\big) + \big(\phi_i(\text{FPR}_{A=a}) - \phi_i(\text{FPR}_{A=b})\big)}{2}
\end{aligned}$$

A.3.3 $\phi_i(\textbf{EOp})$ derivation when $A = Y$

When $A = Y$ in a binary classification task, TPR and TNR are the accuracies for each protected group, respectively. In this case, DP collapses to $\mathbb{P}(\hat{Y} = 1|A = a) \rightarrow \mathbb{P}(\hat{Y} = 1|Y = a)$. In this case, EOp measures the similarity of TPRs between groups.

As a result, when $A = Y$ in a binary classification scenario, the group fairness metrics measure the relationship between TPR, TNR, FPR and FNR not conditioned on the protected attribute A, since these metrics already depend on Y and $A = Y$. As an example, Equal opportunity is defined in this case as $(\text{TPR} + \text{TNR})/2 \in [0, 1]$ [206]:

$$\text{EOp} = \frac{\text{TPR} - \text{FPR} + 1}{2} = \frac{\text{TPR} - (1 - \text{FNR}) + 1}{2} = \frac{\text{TPR} + \text{TNR}}{2} \in [0, 1]$$

Consequently, $\phi_i(\text{EOp}) \in [0, 1]$ when $A = Y$ can be obtained as follows:

$$\text{EOp} = \frac{\sum_{i\in\mathcal{D}} \phi_i(\text{TPR}) + \sum_{i\in\mathcal{D}} \phi_i(\text{TNR})}{2} = \sum_{i\in\mathcal{D}} \frac{\phi_i(\text{TPR})}{2} + \sum_{i\in\mathcal{D}} \frac{\phi_i(\text{TNR})}{2}$$

$$\phi_i(\text{EOp}) = \frac{\phi_i(\text{TPR}) + \phi_i(\text{TNR})}{2}$$

A.3.4 Extension to multi-label and categorical sensitive attribute

As in the binary setting, the group fairness metrics are computed from TPR, TNR, FPR and FNR. Taking as an example TPR, the main change consists of replacing $y = 1$ or $y = 0$ for y_j=y:

$$\phi_i(\text{TPR}_{|Y=y}) = \mathbb{E}_{j\sim p(\mathcal{T}|Y=y)}[\Phi_{i,j}] = \frac{\sum_{j\in\mathcal{T}} \Phi_{i,j}\mathbb{I}[y_j = y]}{|\{x : x \in \mathcal{T}\,|y = y\}|} \tag{28}$$

The conditioned version $\phi_i(\text{TPR}_a)$ may be obtained as:

$$\phi_i(\text{TPR}_{|Y=y,A=a}) = \mathbb{E}_{j\sim p(\mathcal{T}|Y=y,A=a)}[\Phi_{i,j}] = \frac{\sum_{j\in\mathcal{T}} \Phi_{i,j}\mathbb{I}[y_j = y, A_j = a]}{|\{x : x \in \mathcal{T}\,|y = y, A = a\}|} \tag{29}$$

where y and a can be categorical variables. In the scenario where a is not a binary protected attribute, EOp is calculated as $\text{EOp}_a = |\text{TPR} - \text{TPR}_{A=a}|\ \forall a \in A$, and then the maximum difference is selected as the unique EOp for the model $\text{EOp} = \max_{\forall a\in A} \text{EOp}_a$, *i.e.*, the EOp for the group that most differs from the TPR of the entire dataset. Therefore, $\phi_i(\text{EOp})$ for each data point is computed as $\phi_i(\text{EOp}) = \phi_i(\text{TPR}_a) - \phi_i(\text{TPR})$ being a the value of the protected attribute with maximum EOp. The same procedure applies to EOdds. In other words, $\phi_i(\text{EOp}) = \mathbb{E}_{j\sim p(\mathcal{T}|Y=1,A=a)}[\Phi_{i,j}] - \mathbb{E}_{j\sim p(\mathcal{T}|Y=1)}[\Phi_{i,j}]$.

A.4 Dataset Statistics

Image Datasets Total number of images and male/female distribution from the CelebA, LFWA and FairFace datasets are shown on Table 13.

Dataset	Train	Validation	Test
CelebA	94,509|68,261	11,409|8,458	12,247|7,715
LFWA	7,439|2,086	2,832|876	–
FairFace	45,986|40,758	9,197|8,152	5,792|5,162

Table 13. Face Datasets Statistics. Rows stand for `#male|#female`.

Fairness Benchmark Datasets Table 14 below summarizes the statistics of the German, Adult, and COMPAS datasets regarding the distribution of labels and protected groups. Note that all the nomenclature regarding the protected attribute names and values is borrowed from the official documentation of the datasets.

Specific details about the features of each dataset and additional information can be found in the original papers: German [230], COMPAS [16], Adult [249], and ACSIncome [16]. All datasets are pre-processed using AIF360 [51], which uses the same pre-processing as in Calmon et al. [90].

Table 14 (a) shows the distributions of sex and label for the German dataset [230]. It contains 1,000 examples with the target binary variable the individual's *credit risk* and protected groups *age* and *sex*. We use 'Good Credit' as the favorable label (1) and 'Bad

Credit‘ as the unfavorable one (0). Regarding *age* as a protected attribute, ‘Age>25’ and ’Age<25’ are considered the favorable and unfavorable groups, respectively. When using *sex* as a protected attribute, *male* and *female* are considered the privileged and unprivileged groups, respectively. Features used are the one-hot encoded credit history (delay, paid, other), one-hot encoded savings (>500, <500, unknown), and one-hot encoded years of employment (1-4y, >4y, unemployed).

Table 14 (b) depicts the data statistics for the Adult Income dataset [249]. This dataset contains 48,842 examples where the task is to predict if the *income* of a person is more than 50k per year, being >50k considered as the favorable label (1) and <50k as the unfavorable label (0). The protected attributes are *race* and *sex*. When *race* is the protected attribute, *white* refers to the privileged group and *non-white* to the unprivileged group. With *sex* as a protected attribute, *male* is considered the privileged group and *female* the disadvantaged group. The features are the one-hot encoded age decade (10, 20, 30, 40, 50, 60, >70) and education years (<6, 6, 7, 8, 9, 10, 11, 12, >12).

Table 14 (c) contains the statistics about the COMPAS [16] dataset. This dataset has 5,278 examples with the target binary variable *recidivism*. We use *Did recid* as the unfavorable label (0) and *No recid* as the favorable label (1). When *sex* is the protected attribute, *male* is the disadvantaged group and *female* is the privileged one. When using *race* as a protected attribute, *caucasian* is the privileged group and *non-caucasian* the disadvantaged one. Regarding the features, we use one-hot encoded age (<25, 25-45, >45), one-hot prior criminal records of defendants (0, 1-3, >3), and one-hot encoded charge degree of defendants (Felony or Misdemeanor).

(a) German

A\Y	Bad	Good	Total
Male	191	499	690 (69%)
Female	109	201	310 (31%)
Age>25	220	590	810 (81%)
Age<25	80	110	190 (19%)
Total	300 (30%)	700 (70%)	1,000

(b) ADULT

A\Y	<50k	>50k	Total
White	31,155	10,607	41,762 (86%)
non-White	6,000	1,080	7,080 (14%)
Male	22,732	9,918	32,650 (67%)
Female	14,423	1,769	16,192 (33%)
Total	37,155 (76%)	11,687 (24%)	48,842

(c) COMPAS

A\Y	Recid	No Recid	Total
Male	2,110	2,137	4,247 (80%)
Female	373	658	1,031 (20%)
Caucasian	822	1,281	2,103 (40%)
non-Cauc.	1,661	1,514	3,175 (60%)
Total	2,483 (47%)	2,795 (53%)	5,278

Table 14. Tabular datasets statistics. (a) German Credit, (b) Adult Income and (c) COMPAS.

Appendix B

Appendix of Chapter 3

B.1 Effective Resistance and Information Flow

B.1.1 Graph Diffusion Measures

Discrete Information Propagation Information propagation in networks has been widely studied [240], prominently by means of graph diffusion and Random Walks methods. A Random Walk (RW) on a graph is a Markov chain that starts at a given node i, and moves randomly to another node from its neighborhood with probability $1/D_{i,i}$. The RW transition probability matrix is given by $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$ and defines the discrete probability of a random walker to move from node u to node v. $\mathbf{P}^k$ is the k-th power of the transition matrix P: the entry $(P^k)_{ij}$ denotes the probability of transitioning from node i to node j in exactly k steps. The graph's diffusion matrix is defined as $\mathbf{T} = \sum_{k=0}^{\infty} \theta_k \mathbf{P}^k$, and it represents the cumulative effect of multiple steps of a random walk on the graph. Each entry T_{ij} of $\mathbf{T}$ corresponds to the probability of transitioning from node i to node j over an infinite number of steps. θ_k is known as the teleport probability at step k in the random walk. It quantifies the likelihood that, at each step, the random walker will teleport to a random node instead of following an edge. Thus, the sequence $\{\theta_k\}$ is a series of teleport probabilities over the steps. The resulting T captures the cumulative probabilities of transitioning between nodes over an infinite number of steps in the random walk, such that the probability of co-occurrence of two nodes on a random walk corresponds to the probability of information flowing between these two nodes.

However, this approach to assess information flow between nodes in a graph has several limitations. First, it requires considering all the potential paths in a graph, which might not be computationally feasible for large graphs. To overcome this issue, a value of k is typically chosen, which limits the power of the method. Second, the teleport probabilities, θ_k, need to be defined for each k-hop. Several methods have studied how to approximate it, such as Independent Cascade [240], Katz [236], SIR or PageRank [311]. Independent Cascade or SIR methods [240] are based on infection models, where they sample guided random walks and, thus, usually rely on expensive Monte Carlo simulations leading to a sub-optimal probability of transition, unable to consider the topology of the entire graph.

Graph Continuous Diffusion Metrics Graph continuous diffusion metrics —such as the Heat kernel distance, [110], effective resistance (or *commute times* distance) [174, 246]

or the bi-harmonic distance [268]— arise as a generalization of random walk metrics. Their mathematical foundations allow for a better characterization of the information flow and an intuitive interpretation of the diffusion processes in a network.

Diffusion metrics define distances based on fine-grained nuances of the topology of the graph that are not captured by simple geodesic distances. When two nodes can be reached by many paths, they should be *closer* than when they can be reached only by few paths of equal length. When two nodes can be reached by a set of edge-independent paths, they are *closer* than when they are reached by redundant paths. Similarly, when two nodes are separated by a shorter path, they are *closer* than when they are separated by a longer path [78].

In addition, these metrics provide a node embedding, *i.e.*, a numerical representation of each node in the graph that reflects its importance in the process of information diffusion. These embeddings capture the global structure of the network because they incorporate both the local and global geometry of the graph.

The continuous diffusion metrics can be computed using the pseudo-inverse (or Green's function) of the combinatorial graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$, or the normalized Laplacian $\mathcal{L} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$ [174]. The pseudo-inverse, denoted as $\mathbf{L}^{+}$ is computed using the spectral decomposition: $\mathbf{L}^{+} = \sum_{i \geq 0} \lambda_i^{-1} \phi \phi^{\top}$ where $\mathbf{L} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^{\top}$, where λ_i is the i-smallest eigenvalue of the Laplacian corresponding to the ϕ_i eigenvector.

In this chapter, we use the effective resistance, which is a continuous diffusion metric.

B.1.2 Effective Resistance and Commute Times

The Commute Time (CT) [274], $\mathsf{CT}(u, v)$, is the expected number of steps that a random walker needs to go from node u to v and come back to u. The Effective Resistance, R_{uv}, is the Commute Time divided by the volume of the graph [246]. In addition to providing a distance for all pairs of nodes —whether connected or not— R_{uv} may be viewed as an indicator of the criticality or importance of the edges in the flow of information throughout the network [380].

Intuitively, this distance captures how structurally similar and connected are two nodes in a graph. If two nodes are structurally similar to each other, then the effective resistance between them will be small. Conversely, if two nodes are weakly or not connected, then their effective resistance will be large. In addition, we can define a commute time embedding (CTE, $\mathbf{Z} = \sqrt{vol(G)}\mathbf{\Lambda}^{\mathbf{-1/2}}\mathbf{\Phi}^{\top}$) of the nodes in the graph —similar to the idea of the *node's access signature* in Bashardoust et al. [44]—, where the Euclidean distance in such an embedding corresponds exactly to the commute times $\mathsf{CT}(u, v) = ||Z_{u,:} - Z_{v,:}||^2 = \mathbb{E}_u[v] + \mathbb{E}_v[u] = 2|\mathcal{E}|R_{uv}$. This distance is upper bounded by the geodesic distance, with equality in the case of the graph being a tree.

Note that the effective resistance does not rely on any parameter and it is an accurate metric to measure the graph's information flow [23, 99, 188], as explained next. Finally, note that R_{uv} can be computed in a spectral manner:

$$R_{uv} = \sum_{i>0} \frac{1}{\lambda_i} \left(\phi_i(u) - \phi_i(v)\right)^2 \tag{30}$$

where it become explicit that R_{uv} depends on all the eigenvectors of the Laplacian, leading to a better characterization of the information flow.

Information Flow in a Graph The graph's information flow is given by the graph's conductance, which is measured leveraging the Cheeger Constant, h_G, of a graph [108]:

$$h_G = \min_{H \subseteq V} \frac{|\{e = (u, v) : u \in S, v \in \bar{H}\}|}{\min(vol(H), vol(\bar{H}))} \tag{31}$$

The larger h_G, the harder it is to disconnect the graph into separate communities. Therefore, to increase the information flow in the network, one could add edges to the original graph G, creating a new graph G', such that $h_{G'} > h_G$. In addition, by virtue of the Cheeger Inequality, h_G is bounded by the smallest non-zero eigenvalue of $\mathbf{L}$ defined as λ_2:

$$2h_G \leq \lambda_2 < \frac{h_G^2}{2} \tag{32}$$

Finally, the CT is bounded by λ_2 as per the Lovász Bound [274]

$$\left| \frac{\mathsf{CT}(u, v)}{\text{vol}(G)} - \left(\frac{1}{d_u} + \frac{1}{d_v} \right) \right| \leq \frac{1}{\lambda_2} \frac{2}{d_{min}} \tag{33}$$

where $\text{vol}(G)$ is the volume of G, *i.e.*, the sum of the degrees of the all nodes in the graph; d_u, d_v are the degrees of nodes u and v, respectively; and d_{min} is the minimum degree in the graph.

Therefore, a graph's information flow is bounded by λ_2 which is bounded by h_G. The intuition is that graphs with large $\lambda_2 \propto h_G$ have short CT distances and thus they have better information flow. Edge augmentation in a graph would lead to a new graph G' where $h_{G'} > h_G$, with smaller CT distances and therefore better information flow.

The effective resistance is also related to other ways of computing the information flow between two nodes in a graph, such as the Jacobian [59, 127, 383].

B.1.3 Measures Derived from Effective Resistance

The group social capital measures that we propose in this chapter (Sections 3.3.2 and 3.3.3) are grounded in previous metrics from the literature.

Total Effective Resistance $\mathsf{R}_{\mathsf{tot}}$ [140] is the sum of all effective distances in the graph. A lower value of $\mathsf{R}_{\mathsf{tot}}$ indicates ease of signal propagation across the entire network and hence larger information flow. $\mathsf{R}_{\mathsf{tot}}$ is given by:

$$\mathsf{R}_{\mathsf{tot}} = \mathsf{R}_{\mathsf{tot}}(\mathcal{V}) = \frac{1}{2} \mathbf{1}^\top \mathbf{R} \mathbf{1} = \frac{1}{2} \sum_{(v,u) \in V} R_{uv} \tag{34}$$

$$= n \sum_{2}^{n} \frac{1}{\lambda_n} = n \, \text{Tr}(\mathbf{L}^\dagger) \tag{35}$$

The minimum $\mathsf{R}_{\mathsf{tot}} = |V| - 1 = n - 1$ is achieved in a fully connected graph. Conversely, the maximum $\mathsf{R}_{\mathsf{tot}}$ is achieved on a path graph (or linear graph) and $\mathsf{R}_{\mathsf{tot}} = \sum_i^{n-1} i = \frac{1}{2}(n(n-1))$. Therefore, $\mathsf{R}_{\mathsf{tot}}$ is —for connected graphs— in the range $[n - 1, \frac{n(n-1)}{2}]$

Additionally, since the distance between u and v is the Euclidean distance in the embedding $\mathbf{Z}$, $\mathsf{R}_{\mathsf{tot}}$ can be obtained as follows [188]:

$$\mathsf{R}_{\mathsf{tot}} = \sum_{(v,u)\in V} ||Z_{u:} - Z_{v:}||^2 = n \sum_{u\in V} ||Z_{u:}||^2 \tag{36}$$

Similarly to R_{uv}, $\mathsf{R}_{\mathsf{tot}}$ is theoretically related to the connectivity of the graph defined by it smallest non-zero eigenvalue. Ellens et al. [140] demonstrated the relation between $\mathsf{R}_{\mathsf{tot}}$ and λ_2:

$$\frac{n}{\lambda_2} < \mathsf{R}_{\mathsf{tot}} \leq \frac{n(n-1)}{\lambda_2}. \tag{37}$$

Resistance Diameter The proposed group resistance diameter is based on the resistance diameter of a graph $\mathcal{R}_{\text{diam}}$, which is the maximum effective resistance on the graph [99, 325]:

$$\mathcal{R}_{\text{diam}} = \max_{u,v\in V} R_{uv} \tag{38}$$

$\mathcal{R}_{\text{diam}} \propto \lambda_2$ [99, 108], since

$$\frac{1}{n\lambda_2} \leq \mathcal{R}_{\text{diam}} \leq \frac{2}{\lambda_2} \tag{39}$$

and specifically [23, 325]:

$$h_G \leq \frac{\alpha^\epsilon}{\sqrt{\mathcal{R}_{\text{diam}} \cdot \epsilon}} vol(S)^{\epsilon - 1/2}, \tag{40}$$

By [325] we know that

$$\mathcal{R}_{\text{diam}} \leq \frac{1}{\lambda_2} \text{ and } \mathcal{R}_{\text{diam}} \leq \frac{1}{h_G^2} \tag{41}$$

In addition, $\mathcal{R}_{\text{diam}}$ is related to the *cover time* of the graph, which is the expected time required for a random walk to visit every node at least once, *i.e.*, the expected time for a piece of information to reach the entire network. $\mathcal{R}_{\text{diam}}$ can be used to estimate the cover time of the graph, as per [99]:

$$m\,\mathcal{R}_{\text{diam}} \leq \text{cover time} \leq O(m\,\mathcal{R}_{\text{diam}} \log n) \tag{42}$$

Resistance Betweenness and Curvature As the effective resistance is an information distance, this metric can be used to propose alternative betweenness or criticality metrics to the shortest path betweenness [303]. In the literature, several effective resistance-based measures have been proposed to determine a node's criticality, such as: the current flow betweenness [78, 79, 303, 380, 381], the resistance curvature of a node [125, 383], and the information bottleneck property of a node [23, 59, 234]. Note that the last two definitions are mathematically connected [125].

In this work, we focus on the information bottleneck which has a close connection with the resistance curvature of a node. The resistance curvature of a node [125] is expressed as

$$p_u = 1 - \frac{1}{2} \sum_{v\in\mathcal{N}(u)} R_{uv}. \tag{43}$$

Therefore, it fulfills the following equality:

$$p_u = 1 - \frac{1}{2}\sum_{v\in\mathcal{N}(u)} R_{uv} = 1 - \frac{1}{2}\,\mathsf{B}_\mathsf{R}(u)$$
$$\rightarrow \mathsf{B}_\mathsf{R}(u) = -2(p_u - 1). \tag{44}$$

Although the definition of a node's resistance curvature involves computing the sum of the effective resistances between the node and its neighboring nodes, the overall structure of the graph affects all R_{uv}'s and, consequently, the value of p_i and $\mathsf{B}_\mathsf{R}(u)$. The curvature of a node is bounded by $1 - d_u/2 \leq p_u \leq 1/2$.

In addition to the node resistance curvature, the link resistance curvature is defined in Devriendt and Lambiotte [125] as

$$\kappa_{uv} = \frac{2(p_u + p_j)}{R_{uv}} \tag{45}$$

and writing it in terms of the proposed metrics in this work it will translate in:

$$\begin{aligned}\kappa_{uv} &= \frac{2(p_u + p_j)}{R_{uv}} \\ &= \frac{2\left((1 - \frac{1}{2}\,\mathsf{B}_\mathsf{R}(u)) + (1 - \frac{1}{2}\,\mathsf{B}_\mathsf{R}(v))\right)}{R_{uv}} \\ &= \frac{4 - \mathsf{B}_\mathsf{R}(u) - \mathsf{B}_\mathsf{R}(v)}{R_{uv}}\end{aligned} \tag{46}$$

Additionally, this is the value that the *SDRF* algorithm [383] will use to identify the link to add links around. SDRF identifies the link with *lowest* κ_{uv} and adds an edge between a pair the neighbors of the endpoints of that edge that mostly improves the information bottleneck.

B.2 Group Social Capital Metrics and Edge Augmentation Algorithm

B.2.1 Group Social Capital Metrics

Group Isolation and Isolation Disparity

Group Isolation is based on the previously explained notion of total effective resistance of a graph and its close connection with the current flow closeness centrality [303].

We propose to define the isolation of a node as its total effective resistance $\mathsf{R}_\mathsf{tot}(u) = \sum_{v\in\mathcal{V}} R_{uv}$. This centrality metric based on all R_{uv} considers all the eigenvector for its computation, since R_{uv} can be defined using the whole spectrum of the graph. We proceed to derive $\mathsf{R}_\mathsf{tot}(u)$ in terms of the pseudo-inverse of the graph:

$$\begin{aligned}\mathsf{R}_\mathsf{tot}(i) =& \sum_{v}^{N} R_{uv} = \sum_{u}^{N} L^+_{uu} + L^+_{vv} - 2L^+_{uv} \\ =& NL^+_{uu} + \mathrm{Tr}(\mathbf{L}^+) - 2\sum_{v}^{N} L^+_{uv} \\ =& NL^+_{uu} + \mathrm{Tr}(\mathbf{L}^+)\end{aligned} \tag{47}$$

using Ellens et al. [140, Theorem 4.2] and Bozzo and Franceschet [78, Eq. 15]. Here, L^+_{uu} indicates how close node u is on expectancy to all the nodes, and a global component, $\mathrm{Tr}(\mathbf{L}^+)$, proportional to the average pairwise distance in the network, which is denotes of how large is the network overall, independently of the particular node u.

Based on the introduced $\mathsf{R}_{\mathsf{tot}}(u)$, we propose group isolation which is obtained as the expectation in $\mathsf{R}_{\mathsf{tot}}(u)$ for all the nodes in the group:

$$\begin{aligned}\mathsf{R}_{\mathsf{tot}}(S_i) &= \mathbb{E}_{u\sim S_i}\left[\mathsf{R}_{\mathsf{tot}}(u)\right] = |\mathcal{V}|\,\mathbb{E}_{u\sim S_i}\left[\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} R_{uv}\right] \\ &= |\mathcal{V}|\mathbb{E}_{u\sim S_i}\left[\mathbb{E}_{v\sim\mathcal{V}}\left[R_{uv}\right]\right] \\ &= |\mathcal{V}|\mathbb{E}_{u\sim S_i, v\sim\mathcal{V}}\left[R_{uv}\right] \end{aligned} \tag{48}$$

$\mathsf{R}_{\mathsf{tot}}(S_i)$ is computed as the expectation of $\mathsf{R}_{\mathsf{tot}}(u)$ for all nodes in group S_i:

$$\begin{aligned}\mathsf{R}_{\mathsf{tot}}(S_i) &= |\mathcal{V}|\mathbb{E}_{u\sim S_i, v\sim\mathcal{V}}\left[R_{uv}\right] \\ &= |\mathcal{V}|\frac{1}{|S_i|}\sum_{u\in S_i}\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} R_{uv} \\ &= \frac{1}{|S_i|}\sum_{u\in S_i}\sum_{v\in\mathcal{V}} R_{uv} = \frac{1}{|S_i|}\sum_{u\in S_i}\mathsf{R}_{\mathsf{tot}}(u) \\ &= \mathbb{E}_{u\sim S_i}\left[\mathsf{R}_{\mathsf{tot}}(u)\right] \end{aligned} \tag{49}$$

Finally, the cumulative group isolation across all groups in the graph fulfills the following equality with the total effective resistance of the graph.

$$\begin{aligned}\frac{1}{2}\sum_{i\in SA}|S_i|\,\mathsf{R}_{\mathsf{tot}}(S_i) &= \frac{1}{2}\sum_{i\in SA}|S_i|\,\frac{1}{|S_i|}\sum_{u\in S_i}\sum_{v\in\mathcal{V}} R_{uv} \\ &= \frac{1}{2}\sum_{v\in\mathcal{V}}\sum_{v\in\mathcal{V}} R_{uv} = \mathsf{R}_{\mathsf{tot}} = n\,\mathrm{Tr}(\mathbf{L}\dagger) \end{aligned} \tag{50}$$

Therefore, since $\mathsf{R}_{\mathsf{tot}}$ is related to the information flow, $\mathsf{R}_{\mathsf{tot}}(S_i)$ measures the ease of information flow through group S_i in the graph.

The optimal — yet extreme — scenario of maximum information flow in a graph is such where all nodes in the graph are connected. Hence, G will be a fully connected graph. In this scenario, the total effective resistance of the graph reaches its minimum $n-1$, where $n = |\mathcal{V}|$. The total effective resistance of all existing edges in the graph is $n-1$, and given that we have all connections, $\mathsf{R}_{\mathsf{tot}}$ also sums up to $n-1$. Therefore, every pair of nodes will be separated by $R_{uv} = 2/n$. In this scenario, $\mathsf{R}_{\mathsf{tot}}(u) = (2/n)(n-1) = 2-2/n$ for all nodes in the graph. It leads to an group isolation of $\mathsf{R}_{\mathsf{tot}}(S_i) = 2-2/n$ for all the different groups in the graph and therefore a Isolation disparity $\Delta\,\mathsf{R}_{\mathsf{tot}} = 0$.

Regarding group isolation disparity, Equation (16) can be redefined using Equation (49). It is equivalent to the equality for all groups of the mean R_{uv} between of the nodes in the group and all nodes in the graph:

$$\begin{aligned}\mathsf{R}_{\mathsf{tot}}(S_i) &= \mathsf{R}_{\mathsf{tot}}(S_j), \forall i,j \in SA \\ \mathbb{E}_{u\sim S_i, v\sim\mathcal{V}}\left[R_{uv}\right] &= \mathbb{E}_{u\sim S_j, v\sim\mathcal{V}}\left[R_{uv}\right], \forall i,j \in SA\times SA \end{aligned} \tag{51}$$

Group Control and Control Disparity

We provide the proof for the bounds associated with the proposed control metrics.

Bounds of $\mathsf{B}_\mathsf{R}(u)$ and $\mathsf{B}_\mathsf{R}(S_i)$ We show the proof of the bounds of the node and group control.

Theorem B.1. *The control of a node is bounded by $1 \leq \mathsf{B}_\mathsf{R}(u) \leq d_u$, being d_u the degree of node u. Equality on the upper bound holds when all the edges are cut edges,* i.e., *edges that if removed the graph would become disconnected.*

Proof. The resistance curvature of a node is known to be bounded by $1 - d_u/2 \leq p_u \leq 1/2$ [125], and the relation between the curvature and group control is given by the equality $p_u = 1 - \frac{1}{2}\mathsf{B}_\mathsf{R}(u)$. Therefore, we obtain the bounds as

$$\begin{aligned} 1 - \frac{d_u}{2} \leq p_u \leq \frac{1}{2} \rightarrow \\ 1 - \frac{d_u}{2} \leq 1 - \frac{1}{2}\mathsf{B}_\mathsf{R}(u) \leq -\frac{1}{2} \rightarrow \\ -d_u \leq -\mathsf{B}_\mathsf{R}(u) \leq -1 \rightarrow \\ 1 \leq \mathsf{B}_\mathsf{R}(u) \leq d_u \end{aligned}$$

□

Theorem B.2. *The control of a group is bounded by $1 \leq \mathsf{B}_\mathsf{R}(S_i) \leq \frac{vol(S_i)}{|S_i|}$, being $vol(S_i)$ the sum of the degrees of node u. Thus, $vol(S_i)/|S_i|$ is the average degree of all the nodes in S_i. Equality on the upper bound holds when the subgraph with all nodes of S_i and their neighbors is a tree graph,* i.e., *a connected acyclic undirected graph.*

Proof. The control of a group is defined as $\mathsf{B}_\mathsf{R}(S_i) = \mathbb{E}_{u\sim S_i}[\mathsf{B}_\mathsf{R}(u)]$ and using Theorem B.1, we derive the bounds of $\mathsf{B}_\mathsf{R}(S_i)$ as follows:

$$\begin{aligned} 1 \leq \mathsf{B}_\mathsf{R}(u) \leq d_u \rightarrow \\ 1 \leq \mathbb{E}_{u\sim S_i}[\mathsf{B}_\mathsf{R}(u)] \leq \mathbb{E}_{u\sim S_i}[d_u] \\ 1 \leq \mathsf{B}_\mathsf{R}(S_i) \leq \frac{\mathrm{vol}(S_i)}{|S_i|} \end{aligned} \tag{52}$$

□

Control as an Allocation Problem Group control is ($\mathsf{B}_\mathsf{R}(S_i)$) a limited resource to be distributed in the network. The sum of R_{uv} for every edge always equals to $|\mathcal{V}| - 1$ [140]:

$$\sum_{(u,v)\in\mathcal{E}} R_{uv} = |V| - 1.$$

Therefore, the sum of the node controls in the graph is defined as

$$\sum_{u\in\mathcal{V}} \mathsf{B}_\mathsf{R}(u) = \sum_{u\in\mathcal{V}} \sum_{v\in\mathcal{N}(u)} R_{uv} = 2 \times \sum_{(u,v)\in\mathcal{E}} R_{uv} = 2|V| - 2,$$

and the expectation as:

$$\mathbb{E}_{u\sim V}[\mathsf{B}_{\mathsf{R}}(u)] = 2 - \frac{2}{|V|}$$

independently of the number of edges (density) of the graph.

As a consequence of the definition of group control (Equation (15)), the weighted sum of group control for all groups in the graph and the weighted mean also remain constant at:

$$\sum_{S_i \in V} |S_i| \times \mathsf{B}_{\mathsf{R}}(S_i) = 2|V| - 2$$

and

$$\frac{1}{|V|} \sum_{S_i \in V} |S_i| \times \mathsf{B}_{\mathsf{R}}(S_i) = 2 - \frac{2}{|V|}.$$

B.2.2 Group Isolation vs Group Information Share

The group isolation $\mathsf{R}_{\mathsf{tot}}(S_i)$ is given by Equation (13) and it is proportional to the expected information distance when sampling one node from group S_i and another node at random. Thus, $\mathsf{R}_{\mathsf{tot}}(S_i)$ quantifies the information access of a group of nodes in the network.

Conversely, the amount of information that a group of nodes S_i *shares* with the rest of the network is given by the expected information distance between a random node v from the network and a node in group S_i. We define this metric as the *Group Information Share* ($R_{\text{From-}S_i}(V)$).

Mathematically, both concepts are are equivalent as $\mathsf{R}_{\mathsf{tot}}(S_i)$ not only quantifies how isolated a group is (in terms of information access), but also how much it contributes to the information flow that reaches other nodes in the network. Therefore, the disparity in information access between groups and the disparity in information sharing between groups on a randomly selected node in the network are also equivalent.

Mathematical Equivalence The Group Isolation is defined as the expected total effective resistance (or information distance, R_{uv}) between a randomly chosen node in group S_i to all other nodes in the network:

$$\mathsf{R}_{\mathsf{tot}}(S_i) = \mathbb{E}_{u\sim S_i}\left[R_{\text{tot}}(u)\right] = |\mathcal{V}|\mathbb{E}_{u\sim S_i}[\mathbb{E}_{v\sim\mathcal{V}}[R_{uv}]], \tag{53}$$

where $R_{\text{tot}}(u)$ represents the total effective resistance from node u to all other nodes.

Similarly, the Group Information Share measures the expected information distance from a randomly selected node in the network to a group S_i:

$$R_{\text{From-}S_i}(V) = \mathbb{E}_{v\sim V}[R_{\text{From-}S_i}(v)] = |\mathcal{V}|\mathbb{E}_{v\sim\mathcal{V}}[\mathbb{E}_{u\sim S_i}[R_{uv}]] \tag{54}$$

Thus, the expected information distance when a random node is sampled from group S_i is the same as when a random node is sampled from the network and connected to group S_i:

$$|\mathcal{V}|\mathbb{E}_{u\sim S_i}[\mathbb{E}_{v\sim\mathcal{V}}[R_{uv}]] = |\mathcal{V}|\mathbb{E}_{v\sim\mathcal{V}}[\mathbb{E}_{u\sim S_i}[R_{uv}]] \tag{55}$$

$$\mathsf{R}_{\mathsf{tot}}(S_i) = R_{\text{From-}S_i}(V) \tag{56}$$

Group Isolation Disparity In terms of disparity between groups, the Group Isolation Disparity is defined as the difference in isolation between two groups S_i and S_j:

$$\Delta \mathsf{R}_{\mathsf{tot}} = \mathsf{R}_{\mathsf{tot}}(S_i) - \mathsf{R}_{\mathsf{tot}}(S_j) \tag{57}$$

Similarly, the Group Information Share Disparity measures the disparity in information contribution between groups:

$$\Delta R_{\text{From-}S}(V) = R_{\text{From-}S_i}(V) - R_{\text{From-}S_j}(V) \tag{58}$$

Since $\mathsf{R}_{\mathsf{tot}}(S_i) = R_{\text{From-}S_i}(V)$, it follows that:

$$\Delta \mathsf{R}_{\mathsf{tot}} = \Delta R_{\text{From-}S_i}(V) \tag{59}$$

Thus, the disparities in Group Isolation and in Group Information Share are the same, highlighting the dual role of the proposed metric in capturing both access to information from the network and information sharing with the rest of the network.

B.2.3 Efficient Version of ERG

Algorithm 4 shows an efficient manner to update the pseudo-inverse of the Laplacian after adding one edge to the graph. Therefore, we avoid the computation of $\mathbf{L}^\dagger$ after every edge addition. $\mathbf{L}^\dagger$ is easily updated using the Woodbury's formula which is based on the values of $\mathbf{L}^\dagger$ (see Black et al. [59] for a proof).

Algorithm 4: ERG-Link

Data: Graph $G = (\mathcal{V}, \mathcal{E})$, a protected attribute SA, budget B of total number of edges to add

Result: Rewired Graph $G' = (\mathcal{V}', \mathcal{E}')$

$\mathbf{L} = \mathbf{D} - \mathbf{A}$;

$S_d = \text{argmax}_{S_i \forall i \in SA} \mathsf{R}_{\mathsf{tot}}(S_i)$;

$\mathbf{L}^\dagger = \sum_{i>0} \frac{1}{\lambda_i} \phi_i \phi_i^\top = \left(\mathbf{L} + \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)^{-1} - \frac{\mathbf{1}\mathbf{1}^\top}{n}$;

Repeat

- $\mathbf{R} = \mathbf{1}\,\text{diag}(\mathbf{L}^\dagger)^\top + \text{diag}(\mathbf{L}^\dagger)\mathbf{1}^\top - 2\mathbf{L}^\dagger$;
- $C = \{(u, v) \mid u \in S_d \text{ or } v \in S_d, (u, v) \notin E\}$; `// Select edge candidates`
- $\mathcal{E}' = \mathcal{E}' \cup \arg\max_{(u,v) \in C} R_{uv}$;
- `// Fast update of` $\mathbf{L}$ `and` $\mathbf{L}^\dagger$
- $\mathbf{L} = \mathbf{L} + (\mathbf{e}_u - \mathbf{e}_v)(\mathbf{e}_u - \mathbf{e}_v)^\top$;
- $\mathbf{L}^\dagger = \mathbf{L}^\dagger - \frac{1}{1+R_{uv}} \times (\mathbf{L}^\dagger_{u,:} - \mathbf{L}^\dagger_{v,:}) \otimes (\mathbf{L}^\dagger_{u,:} - \mathbf{L}^\dagger_{v,:})$; `// updated by Woodbury`

Until $|\mathcal{E}' \setminus \mathcal{E}| = B$;

return G';

B.3 Additional Experiments

All the experiments were done using a workstation with 16GB RAM; and processor 11th Gen Intel(R) Core(TM) i7, 3.00GHz, 2995 Mhz, 4 Core(s), 8 Logical Processor(s).

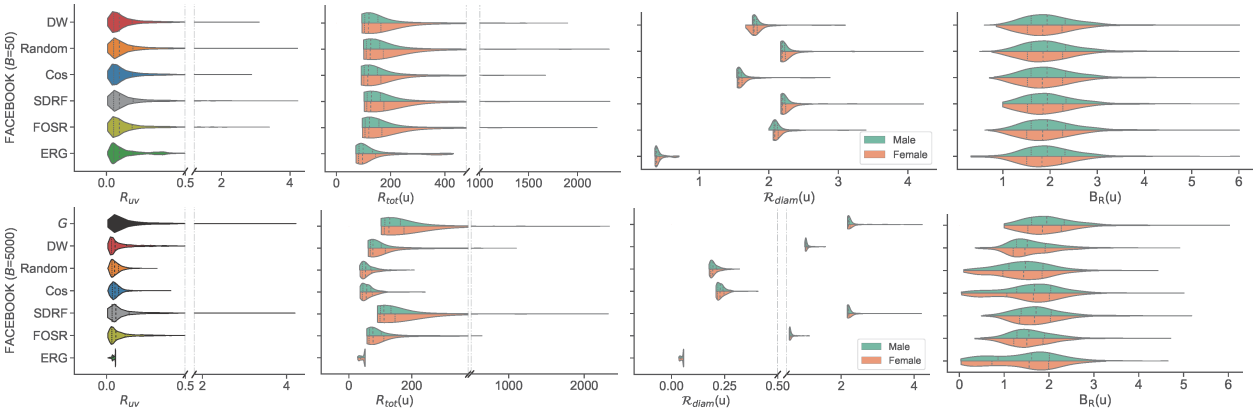


Figure 34. Distribution of R_{uv} and proposed social capital metrics on the Facebook dataset and after different graph interventions, 50 links in the top row and 5,000 in the bottom row. Columns show (left) distribution of all R_uv distances, (center-left) distribution of all $\mathsf{R}_{\mathsf{tot}}(u)$, (center-right) distribution of all $\mathcal{R}_{\text{diam}}(u)$, (right) distribution of all $\mathsf{B}_{\mathsf{R}}(u)$. The distributions for the node metrics are shown for the two groups according to the protected attribute (gender).

B.3.1 Distribution of Social Capital by Group

For completeness, Figure 34 illustrates the distributions of all effective resistances R_{uv} and the node's social capital metrics –$\mathsf{R}_{\mathsf{tot}}(u)$, $\mathcal{R}_{\text{diam}}(u)$ and $\mathsf{B}_{\mathsf{R}}(u)$– in the graph before and after the edge augmentation interventions for each of the groups.

The plots correspond to two different edge augmentation experiments on the Facebook dataset, with budgets of B=50 and B=5,000 new edges. Edge augmentation via `ERG-Link` is able to drastically reduce all effective resistances of the graph, unlike the other methods and even for the small budget. Note how there is still a long tail in the distribution of effective resistances after edge augmentation with the baseline methods, which illustrates that there are still nodes that struggle to exchange information even after the intervention.

Regarding the node social capital metrics, edge augmentation via `ERG-Link` reduces all $\mathsf{R}_{\mathsf{tot}}(u)$ and $\mathcal{R}_{\text{diam}}(u)$, which explains why $\mathsf{R}_{\mathsf{tot}}(S_i)$ and $\mathsf{R}_{\mathsf{tot}}(S_i)$ is improved for all groups in the graph, and particularly for the disadvantaged group. Thus, $\Delta\,\mathsf{R}_{\mathsf{tot}}$ and $\Delta\,\mathcal{R}_{\text{diam}}$ are significantly reduced.

B.3.2 Evolution of Group Social Capital During the Interventions

We provide additional results regarding the evolution of group social capital metrics during the graph's intervention, as initially shown in Figure 21. The goal is to analyze the effectiveness of the edge augmentation methods on both small ($B = 50$ edges) and large ($B = 5,000$ edges) budget scenarios. Figures 35 and 36 show the evolution of both the group social capital metrics for each group and the structural group unfairness metrics on the Facebook dataset.

We observe in Figure 35 how edge augmentation via `ERG-Link` is able to significantly improve the group social capital for all groups and reduce all structural unfairness metrics. In contrast, the baselines fail to do so.

Figure 36 depicts the evolution of group social capital and disparity metrics when adding 5,000 edges to the Facebook dataset. The more edges we add to a graph, the denser the

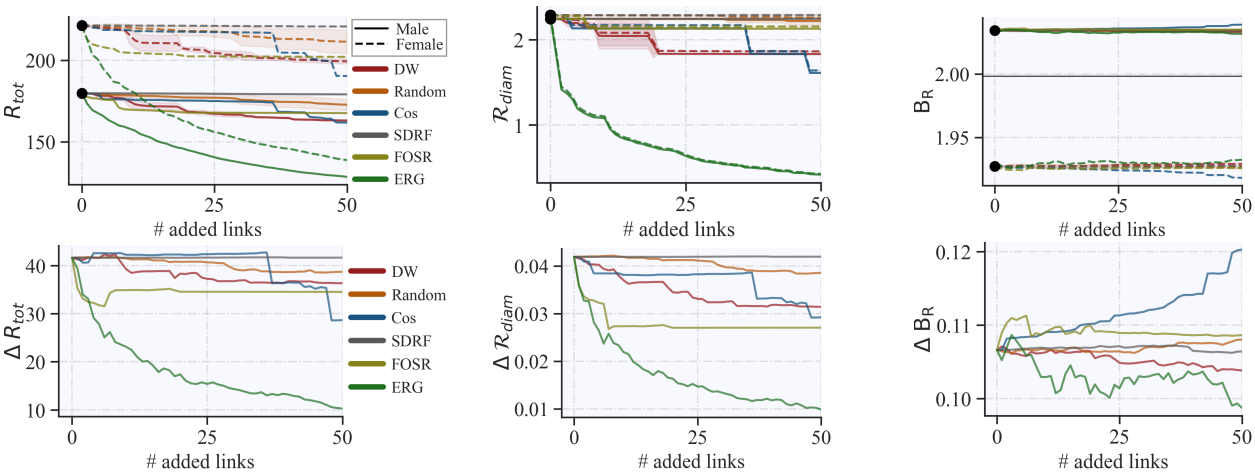


Figure 35. Evolution of group social capital and disparity metrics for both groups as the number of added edges increases on the Facebook dataset with a budget of 50 new edges.

graph becomes and therefore the better the information flow. Thus, one can expect that after a large number of added edges, all methods behave similarly.

However, we observe some differences. First, the convergence to minimal isolation and diameter disparities is significantly faster when adding edges via `ERG-Link` than any other method. Second, the decrease in $\mathsf{R}_{\mathsf{tot}}(S_i)$ is very significant for both groups (males and females) even after just adding a small number of edges by means of `ERG-Link`. Third, regarding group control and control disparity, edge augmentation via `ERG-Link` systematically reduces $\Delta\,\mathsf{B}_{\mathsf{R}}$ while converging each group control to the optimal $\mathsf{B}_{\mathsf{R}}(S_i) = 2 - 2/|\mathcal{V}|$.

We also show in Figure 37 the evolution of $\mathsf{B}_{\mathsf{R}}(S_i)$ and $\Delta\,\mathsf{B}_{\mathsf{R}}$ on the UNC28 dataset with 1,000 edge additions along with the distribution of node's control $\mathsf{B}_{\mathsf{R}}(u)$ after the intervention to showcase an scenario where `ERG-Link` is able to reach the optimal control disparity in the graph, allocating the same amount of control for each group, $\mathsf{B}_{\mathsf{R}}(S_i) \approx 2 - 2/|\mathcal{V}|\,\forall\, i \in SA \rightarrow \Delta\,\mathsf{B}_{\mathsf{R}} \approx 0$.

Last but not least, the different baselines do not reach a better behavior than the random method, neither on the group social capital metrics nor in terms of structural unfairness. After 5,000 edge additions, the random method significantly reduces the unfairness metrics $\Delta\,\mathsf{R}_{\mathsf{tot}}$ and $\Delta\,\mathcal{R}_{\mathrm{diam}}$ while also achieving decent rated of $\mathsf{R}_{\mathsf{tot}}(S_i)$ and $\mathcal{R}_{\mathrm{diam}}(S_i)$ since the budget is high enough to improve the information flow with no strategy. However, all baselines struggle to optimize $\Delta\,\mathsf{B}_{\mathsf{R}}(S_i)$ and $\mathsf{B}_{\mathsf{R}}(S_i)$. None of them is able to improve, and the cosine similarity approach even increased the Control Disparity.

B.3.3 Additional Edge Augmentation Examples

For completeness, we qualitatively illustrate on Figures 38 and 39 the behavior of four edge augmentation methods on five synthetic graphs. The Figures depict the synthetic graphs and the first two edges added by each algorithm. They also show quantitatively how the methods are able to improve the overall flow of information in the graph, defined as $\mathsf{R}_{\mathsf{tot}} = \sum_{(u,v)\in\mathcal{E}} R_{uv}$. This analysis not only provides insight into the performance of the edge augmentation methods, but also demonstrates their adaptability to different graph topologies.

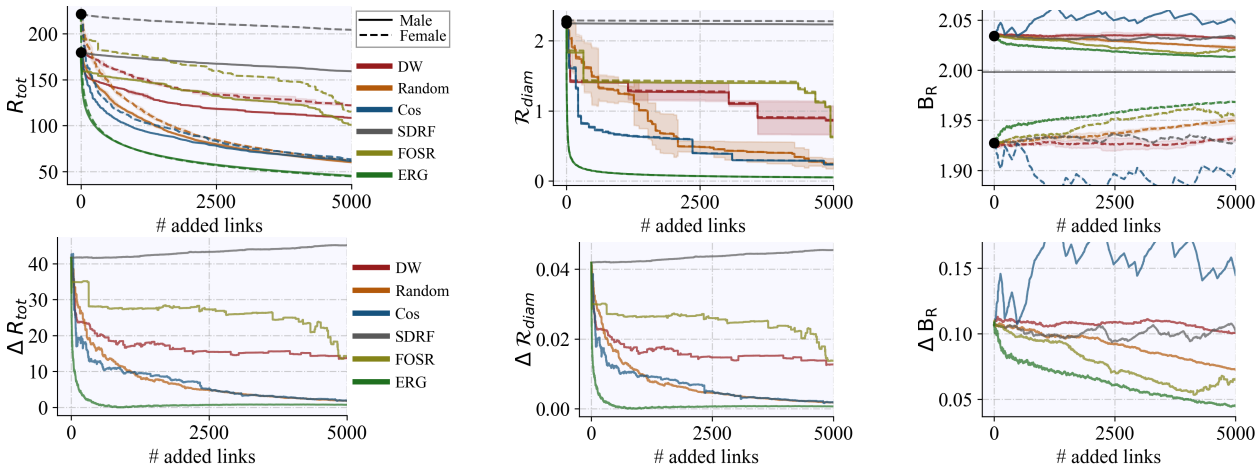


Figure 36. Evolution of group social capital metrics for both groups and fairness metrics as the number of added links increases, on Facebook dataset after adding 5,000 edges. Edge augmentation via `ERG-Link` exhibits a faster rate of convergence to the optimal scenario than the baselines.

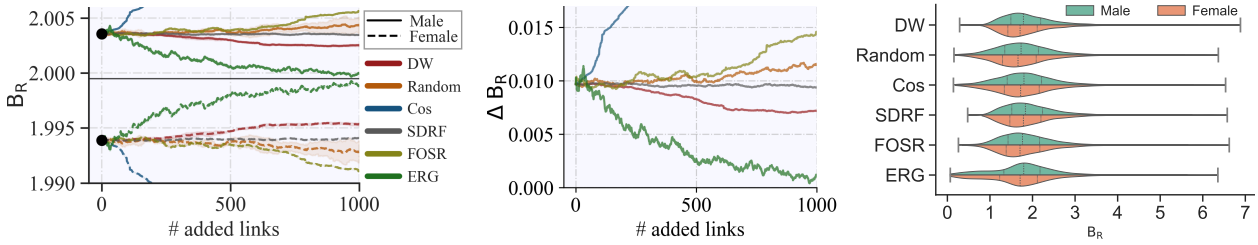


Figure 37. Illustration of the evolution of the group control (left, $\mathsf{B}_\mathsf{R}(S_i)$) and control disparity (middle, $\Delta\,\mathsf{B}_\mathsf{R}$) through an edge augmentation with $B = 1,000$ edges on the UNC dataset. Rightmost figure: distribution of the nodes' control ($\mathsf{B}_\mathsf{R}(u)$) for each group. Note how edge augmentation via `ERG-Link` yields a control for all groups approaching their optimal value of $2-2/|\mathcal{V}|$ by reducing the control of the privileged group (males) while increasing the control of the disadvantaged group (females).

The Figures correspond to graphs from five synthetic datasets, each designed to represent different graph characteristics: (1) a Barbell graph with Barbell size of 4 and path size of 5; (2) a power-law clustered graph with 50 nodes, 2 random new edges per each new node, and a 0.95 probability of closing a triangle; (3) the same graph, but with a 0.8 probability of closing a triangle, meaning that the bottleneck and clustered structures are weaker; (4) an SBM model with intra-cluster probability of 0.4 and inter-cluster probability of 0.01; and (5) a hand-crafted path graph with communities in the middle to simulate a different type of bottleneck. The Barbell graph is shown in Figure 38 and the rest of the graphs in Figure 39.

The Figures illustrate the results for all algorithms described in Section 3.4 except for DeepWalk because its behavior closely resembles randomness. This similarity arises from the impracticality of performing extensive training after each edge addition, given the computational limitations of the proposed task. Performing extensive training after each edge addition would be computationally infeasible within the scope of our study.

Observe how `ERG-Link` always reduces R_tot the most by adding the edges with the maximum distance. Other approaches, such as FOSR, do not exhibit a consistent performance across different graph topologies. In fact, graphs with complex structures and/or less defined

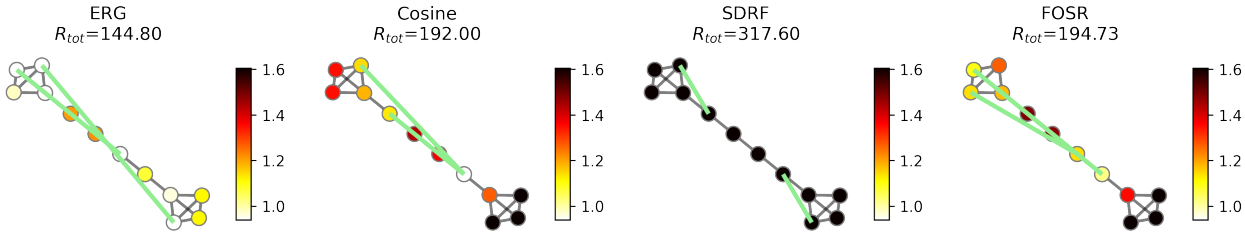


Figure 38. Example of the added links by the different algorithms on a Barbell synthetic graph. Nodes are colored by $\mathsf{R}_{\mathsf{tot}}(u)$. More examples in Figure 39.

bottlenecks exhibit a higher degree of difficulty for this method. Cosine similarity suffers from the same disadvantage. SDRF adds edges around the graph's bottleneck to avoid over-squashing, but this strategy does not necessarily translate into an improvement of the graph's information flow.

B.4 Computation Time

Table 15 depicts the number of edge additions per second that each algorithm is able to perform. For instance, in the conducted experiment on the Facebook dataset with $B = 5,000$ edges, edge augmentation via `ERG-Link` required approximately 3 minutes and 30 seconds; via cosine Similarity (Cos), approximately 6 minutes; and by means of DeepWalk (DW), over 3 hours. In the case of the Google+ dataset, edge augmentation via `ERG-Link` required 36 minutes; via cosine Similarity (Cos), 72 minutes; and by means of DeepWalk (DW), over 17 hours.

Dataset	Facebook	Google+	UNC28
SDRF	1.00	0.25	0.46
FOSR	166.6	125.0	83.3
DeepWalk	0.39	0.08	0.07
Cosine	13.9	1.15	1.39
ERG	23.8	2.31	1.66

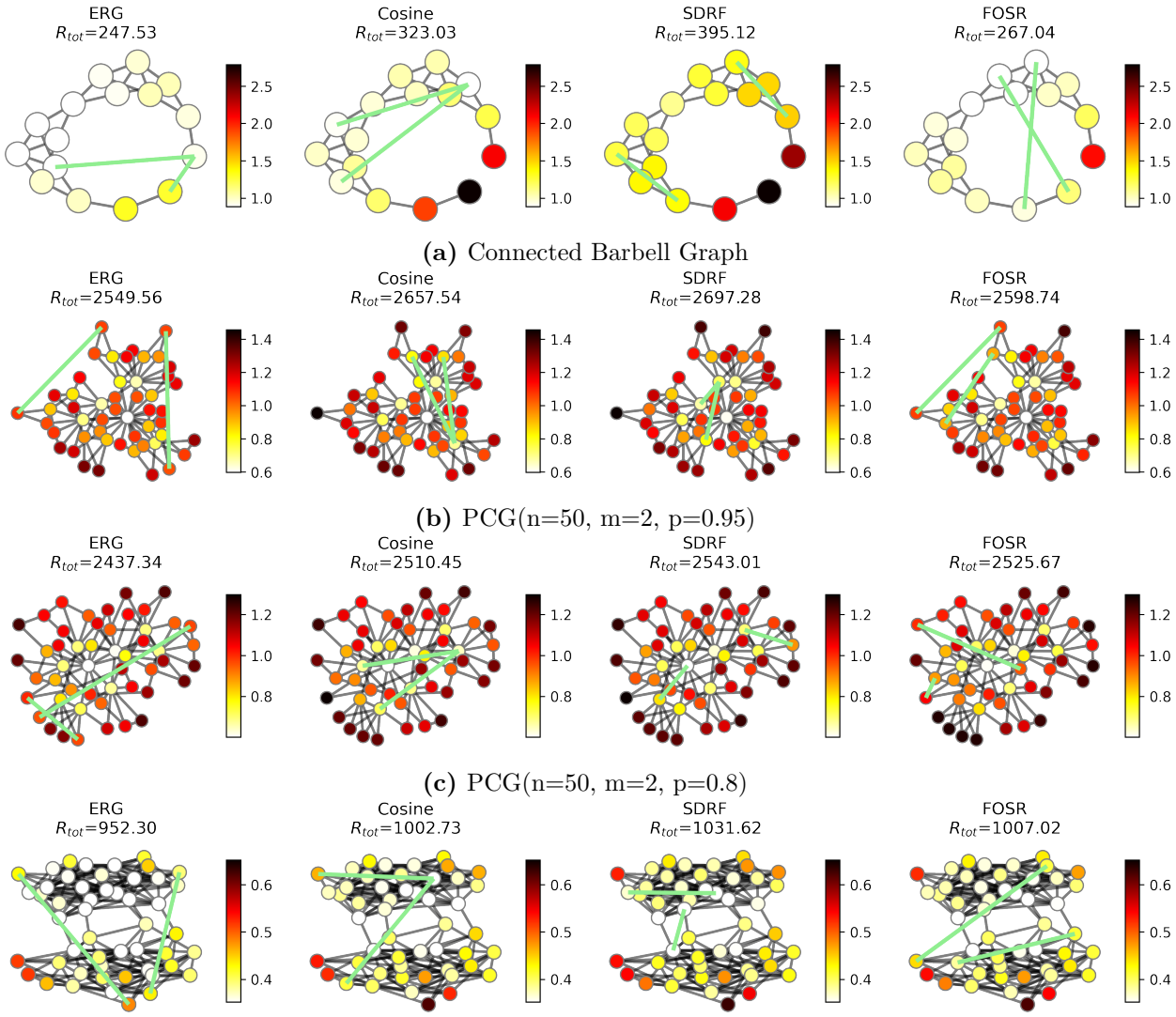
Table 15. Added edges per second on the three datasets

B.5 Notation

For the seek of clarity, Table 16 summarizes the notation used in Chapter 3.

Symbol	Description	Definition
$G = (\mathcal{V}, \mathcal{E})$	Graph = (Nodes, Edges)	
$n = \lvert\mathcal{V}\rvert$	Number of nodes	
$\mathbf{A}$	Adjacency matrix: $\mathbf{A} \in \mathbb{R}^{n \times n}$	$A_{u,v} = 1$ if $(u, v) \in \mathcal{E}$ and 0 o/w
v or u	Node $v \in \mathcal{V}$ or $u \in \mathcal{V}$	
e	Edge $e \in \mathcal{E}$	
d_v	Degree of node, *i.e.*, number of neighbors v	$d_v = \sum_{u \in V} A_{v,u}$
$\mathbf{D}$	Degree diagonal matrix where d_v in D_{vv}	$\mathbf{D} = \text{diag}(d_0, \ldots, d_{\mathcal{V}})$
$\text{vol}(G)$	Sum of the degrees of the graph	$\text{vol}(G) = \sum_{u \in V} d_u = 2\lvert\mathcal{E}\rvert = \text{Tr}[\mathbf{D}]$
$\mathcal{N}(u)$	Neighbors of u	$\mathcal{N}(u) = \{v : (u, v) \in \mathcal{E}\}$
S_i	Subset of nodes	$S \subseteq V$
SA	Set of sensitive attributes	$SA = \{sa_1, sa_2, \ldots, sa_{\lvert SA\rvert}\}$
$SA(v)$	Value of the sensitive attribute of the node v	
sa_i	Specific value of a sensitive attribute, *e.g.*, sa_i=*female*.	
S_i	Set of nodes defined by their sensitive attribute	$S_i = \{v \in V \lvert SA(v) = sa_i\}$
S_d	Set of nodes defined by their sensitive attribute with the highest level of isolation $\mathsf{R}_{\mathsf{tot}}(S_i)$	
$\mathbf{L}$	Graph Laplacian	$\mathbf{L} = \mathbf{D} - \mathbf{A} = \mathbf{\Phi \Lambda \Phi}^\top$
$\mathbf{\Lambda}$	Eigenvalue matrix of $\mathbf{L}$	
Φ	Matrix of eigenvectors of $\mathbf{L}$	
λ_i	The i-th smallest eigenvalue of $\mathbf{L}$	
$\mathbf{f}_i$	Eigenvector associated with the i-th smallest eigenvalue of $\mathbf{L}$	
$\mathbf{L}^+$	The pseudo-inverse of $\mathbf{L}$	$\mathbf{L}^+ = \sum_{i>1} \lambda_i^{-1} \phi\phi^\top$
h_G	Cheeger constant	Eq. 31
$\mathbf{e}_u$	Unit vector with unit value at u and 0 elsewhere	
R_{uv}	Effective resistance between nodes u and v	$R_{uv} = (\mathbf{e}_u - \mathbf{e}_v)\mathbf{L}^+(\mathbf{e}_u - \mathbf{e}_v)$
$\mathbf{R}$	Effective resistance matrix where the i, j entry corresponds to R_{ij}	$\mathbf{R} = \mathbf{1}\text{diag}(\mathbf{L}^+)^\top + \text{diag}(\mathbf{L}^+)\mathbf{1}^\top - 2\mathbf{L}^+$
$\mathbf{Z}$	Commute Time Embedding matrix	$\mathbf{Z} = \sqrt{vol(G)}\mathbf{\Lambda}^{-1/2}\mathbf{\Phi}^\top$
$\mathbf{z}_u$	Commute times embedding of node $Z_{u,:}$	
$\mathsf{CT}(u, v)$	Commute time	$\mathsf{CT}(u, v) = \text{vol}(G)R_{u.v}$
$\mathsf{R}_{\mathsf{tot}}$	Total Effective Resistance of G	$\mathsf{R}_{\mathsf{tot}} = \frac{1}{2}\mathbf{1}^\top\mathbf{R}\mathbf{1}$
$\mathcal{R}_{\text{diam}}$	Resistance Diameter of G	$\mathcal{R}_{\text{diam}} = \max_{u,v \in \mathcal{V}} R_{u,v}$
$\mathsf{R}_{\mathsf{tot}}(u)$	Node Isolation or Total Effective Resistance	$\mathsf{R}_{\mathsf{tot}}(u) = \sum_{v \in V} R_{uv}$
$\mathcal{R}_{\text{diam}}(u)$	Node Resistance Diameter	$\mathcal{R}_{\text{diam}}(u) = \max_{v \in V} R_{uv}$
$\mathsf{B}_{\mathsf{R}}(u)$	Node Control or Resistance Betweenness	$\mathsf{B}_{\mathsf{R}}(u) = \sum_{v \in \mathcal{N}(u)} R_{uv}$
$\mathsf{R}_{\mathsf{tot}}(S_i)$	Group Isolation or Total Effective Resistance	$\mathsf{R}_{\mathsf{tot}}(S_i) = \lvert S\rvert^{-1} \sum_{u \in S} \mathsf{R}_{\mathsf{tot}}(u)$
$\mathcal{R}_{\text{diam}}(S_i)$	Group Resistance Diameter	$\mathcal{R}_{\text{diam}}(S) = \lvert S\rvert^{-1} \sum_{u \in S} \mathcal{R}_{\text{diam}}(u)$
$\mathsf{B}_{\mathsf{R}}(S_i)$	Group Control or average Betweenness	$\mathsf{B}_{\mathsf{R}}(S) = \lvert S\rvert^{-1} \sum_{u \in S} \mathsf{B}_{\mathsf{R}}(u)$

Table 16. Table of Notation of Chapter 3.



(a) Connected Barbell Graph

(b) PCG(n=50, m=2, p=0.95)

(c) PCG(n=50, m=2, p=0.8)

(d) SBM(intra=0.5, inter=0.1)

Figure 39. Examples of the added links by the `ERG-Link`, Cosine, SDRF and FOSR algorithms on synthetic graphs. Nodes are colored by $\mathsf{R}_{\mathsf{tot}}(u)$. Edge augmentation via `ERG-Link` consistently yields the largest improvement in information flow in the graph and hence the smallest values of R_{tot}.

Appendix C

Appendix of Chapter 4

C.1 Existence of an Optimal Integral Solution to the Linear Program

To show that the linear program defined in Equation (23) has an integral optimal solution $\mathbf{v}^* = (v^*_{ir})_{i\in\mathcal{I},r\in\mathcal{R}} \in \{0,1\}^{n\cdot k}$, it is sufficient to verify that the matrix A associated with the constraints of the linear program is totally unimodular.

First, recall that $|\mathcal{I}| = n$ and $|\mathcal{R}| = k$. The constraints of the linear program in Equation (23) can be written in matrix formulation as $A\mathbf{v} \leq \mathbf{d}, \mathbf{v} \geq 0$ for $A \in \{-1,0,+1\}^{(n+k+2)\times n\cdot k}$ and $\mathbf{d} \in \mathbb{R}^{n+k+2}$ with

$$A_{i,j} = \begin{cases} 1 & \text{if } i \leq n \text{ and } i\cdot k \leq j \leq i\cdot k + k \\ 1 & \text{if } n \leq i \leq n+k \text{ and } i \equiv j \mod k \\ 1 & \text{if } i = n+k+1 \\ -1 & \text{if } i = n+k+2 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \mathbf{d} = \begin{pmatrix} \mathbf{1} \\ \mathbf{c} \\ \max(n-b,0) \\ -\max(n-b,0) \end{pmatrix}, \tag{60}$$

where, in the vector $\mathbf{v}$, we stack all the variables v_{ir} associated to each individual $i \in \mathcal{I}$ in turn. Below we give an example of matrix A when $|\mathcal{I}| = 3$ and $|\mathcal{R}| = 2$,

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 \end{pmatrix}.$$

According to Hoffman and Kruskal's theorem [344], if A is an integral matrix, then A is totally unimodular if and only if the polyhedron $\{x \,|\, x \geq 0; Ax \leq \mathbf{d}\}$ is integral for each integral vector $\mathbf{d}$. As $\mathbf{d}$ is integral in Equation (60), it follows that, if A is totally unimodular, then the linear program in Equation (23) has an integral optimal solution.

To show that A is totally unimodular, we use the characterization of Ghoulia-Houri [344]. In this characterization, a matrix A is totally unimodular if and only if each collection of

rows of A can be separated into two subsets such that the sum of the rows in one subset minus the sum in the other subset is a (row) vector with entries only in {-1,0,1}.[32]

Let A_i denote the i-th row of A, and let $A_{\mathcal{I}}$ ($A_{\mathcal{R}}$) denote the subset of rows corresponding to the matching (capacity) constraints for each individual (resource), *i.e.*, $A_{\mathcal{I}} = \{A_1, \ldots, A_n\}$ and $A_{\mathcal{R}} = \{A_{n+1}, \ldots, A_{n+k}\}$. Let T be an arbitrary subset of rows from A, we show that there exists a partition of T into T^+ and T^- such that

$$\alpha := \left(\sum_{a \in T^+} a + \sum_{a' \in T^-} -a' \right) \in \{-1, 0, 1\}^{n \cdot k}.$$

Let α^+ denote the result of the left sum over subset T^+ and α^- denote the result of the right sum over subset T^-. We distinguish three cases.

- If $\{A_{n+k+1}, A_{n+k+2}\} \subseteq T$ or $\{A_{n+k+1}, A_{n+k+2}\} \cap T = \emptyset$, let $T^+ = T \cap (A_{\mathcal{I}} \cup \{A_{n+k+1}, A_{n+k+2}\})$ and $T^- = T \cap A_{\mathcal{R}}$. Note that, this is a valid partition of T as $A = A_{\mathcal{I}} \cup A_{\mathcal{R}} \cup \{A_{n+k+1}, A_{n+k+2}\}$. Since rows A_{n+k+1} and A_{n+k+2} are either not in T or they sum up to vector $\mathbf{0}^T$, we have that

  $$\alpha^+ = \sum_{a \in T^+} a = \sum_{a \in T \cap A_{\mathcal{I}}} a \,.$$

  Now, observe that for any $j \in [n \cdot k]$, there is only one row $a \in A_{\mathcal{I}}$ and only one row $a' \in A_{\mathcal{R}}$ such that a_j, respectively a'_j, is non zero (*i.e.*, $a_j = a'_j = 1$). Thus, it follows that

  $$\alpha^+_j \in \{0, 1\} \quad \text{and} \quad \alpha^-_j \in \{0, -1\}$$

  implying that $\alpha \in \{-1, 0, 1\}^{n \cdot k}$.

- If $A_{n+k+1} \in T$ and $A_{n+k+2} \notin T$, let $T^+ = T \cap (A_{\mathcal{I}} \cup A_{\mathcal{R}})$ and $T^- = \{A_{n+k+1}\}$. By the same reasoning as above, we have that $\alpha^+_j \in \{0, 1, 2\}$ for all $j \in [n \cdot k]$. Since $\alpha^- = -A_{n+k+1} = -\mathbf{1}^T$, it follows that $\alpha \in \{-1, 0, 1\}^{n \cdot k}$.

- If $A_{n+k+1} \notin T$ and $A_{n+k+2} \in T$, let $T^+ = T \cap (A_{\mathcal{I}} \cup A_{\mathcal{R}} \cup \{A_{n+k+2}\})$ and $T^- = \emptyset$. As $A_{n+k+2} = -A_{n+k+1}$, it follows analogously to the previous case distinction that $\alpha^+ \in \{-1, 0, 1\}^{n \cdot k}$. Thus, we have that $\alpha \in \{-1, 0, 1\}^{n \cdot k}$ as $\alpha^- = 0$.

In all cases $\alpha \in \{-1, 0, 1\}^{n \cdot k}$, which proves that A is totally unimodular, so the linear program in Equation (23) has an integral optimal solution.

C.2 Data Generation Process

In this subsection, we describe the synthetic data generation process used in the human subject study to generate the matching problems. We simulate a scenario in which patients must be matched to available time slots in a hospital, with each patient–slot pair associated with an individual success score indicating the likelihood that the patient will attend the assigned appointment. Each matching task consists of a pool $\mathcal{I}$ of $n = 20$ patients to be

[32] In the reference, this characterization is written for columns. However, as the transpose of a totally unimodular matrix is also totally unimodular, this characterization is also valid for rows.

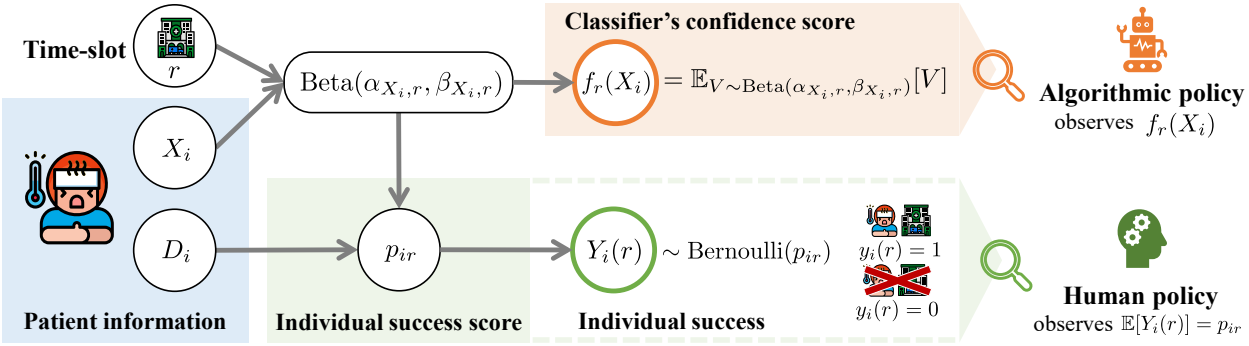


Figure 40. Data generation process. Human policy observes both X_i and D_i, and can thus use the patient's expected success $\mathbb{E}[Y_i(r)] = p_{ir} = Q_{(\alpha_{X_i,r},\beta_{X_i,r})}(D_i)$ to perform the matching. On the other hand, the algorithmic policy does not observes D_i and uses the classifier's confidence score $f_r(X_i) = \mathbb{E}_{V\sim\text{Beta}(\alpha_{X_i,r},\beta_{X_i,r})}[V]$ to perform the matching.

assigned across a set $\mathcal{R}$ of $k = 10$ time slots, each slot $r \in \mathcal{R}$ has a capacity of $c_r = 2$ available appointments.[33] An illustration of the data generation process is available in Figure 40.

Patient information. A widely accepted assumption in the human-AI collaboration literature is that human experts often have access to richer contextual information than what is available to machine learning models—since not all relevant information can be captured or encoded in the input features provided to the classifier. To reflect this, our data generation process is designed to produce more accurate confidence scores for human participants, while providing less accurate scores to the algorithmic matching. Each party—humans and the algorithm—then solves the matching task using its respective confidence scores, with the same objective of maximizing the number of patients who make use of their assigned appointments.

To model this, we represent the information of each patient $i \in \mathcal{I}$ as a pair $Z_i = (X_i, D_i)$, where X_i denotes the observable features available to both humans and the algorithm, and D_i denotes additional contextual information accessible only to the human policy.[34] The algorithmic policy generates confidence scores based solely on $X_i = \phi(Z_i)$, where the projection ϕ discards the variable D_i from Z_i, while the human policy uses the full information (X_i, D_i), resulting in more accurate confidence scores. The variable X_i is drawn from a categorical distribution $\mathcal{X} = \{0, 1, 2\}$, with probabilities $P(X_i = 0) = 0.20$, $P(X_i = 1) = 0.45$, $P(X_i = 2) = 0.35$. The variable D_i is sampled from a uniform distribution: $D_i \sim U(0, 1)$.

Individual success. We define individual success based on whether patient $i \in \mathcal{I}$ attends their assigned appointment at time slot r, *i.e.*, $y_i(r) = 1$ if the patient attends, and $y_i(r) = 0$ otherwise. We model $Y_i(r)$ as a Bernoulli random variable, whose probability p_{ir} depends on the patient's features $Z_i = (X_i, D_i)$ and the time slot r. More precisely, for every feature $x \in \mathcal{X}$ and time slot $r \in \mathcal{R}$, we define a Beta distribution Beta$(\alpha_{x,r}, \beta_{x,r})$, whose parameters $(\alpha_{x,r}, \beta_{x,r})$ depend on the specific combination of x and r. Table 17 lists the parameters of the Beta$(\alpha_{x,r}, \beta_{x,r})$ distribution for each $x \in \mathcal{X}$ and $r \in \mathcal{R}$, while Figure 41 illustrates the corresponding distributions.

For each patient and time slot pair $(i, r) \in \mathcal{I} \times \mathcal{R}$, we set the variable p_{ir} (i.e., the success

[33]Each time slot $r \in \{1, \ldots, 10\}$ represent half-day period spanning from Monday morning to Friday afternoon *i.e.*, {Monday-am, Monday-pm, Tuesday-am, ..., Friday-pm}, as illustrated in Figure 23.

[34]Recall that capital letters denote random variables, while lowercase letters represent their realizations.

Table 17. Beta parameters $(\alpha_{xr}, \beta_{xr})$. Rows correspond to feature $x \in \{0, 1, 2\}$ and columns to time slots $r \in \{1, \ldots, 10\}$.

	$r \in \{1, \ldots, 10\}$, where each r represents a time slot.									
	1	2	3	4	5	6	7	8	9	10
$x = 0$	(0.2,0.3)	(20,20)	(0.2,0.3)	(20,17)	(17,20)	(20,20)	(0.2,0.4)	(19,12)	(0.2,0.3)	(0.15,0.2)
$x = 1$	(20,20)	(0.2,0.4)	(19,12)	(0.2,0.3)	(0.15,0.2)	(0.2,0.3)	(20,20)	(0.2,0.3)	(20,17)	(17,20)
$x = 2$	(1,10)	(1,10)	(5,10)	(5,2)	(3.1,4)	(19,12)	(0.2,0.3)	(0.15,0.2)	(0.2,0.3)	(20,20)

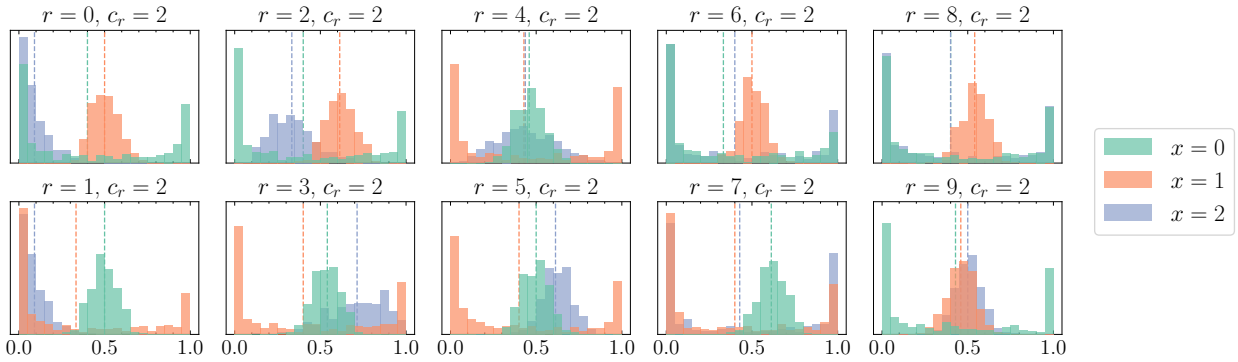


Figure 41. Beta distributions for the parameters defined in Table 17. Dashed vertical lines mark the classifier's confidence scores $f_r(x)$ used by the algorithmic policy given x.

probability for the Bernoulli distribution of $Y_i(r)$) as the D_i-quantile of the corresponding beta distribution: $p_{ir} = Q_{(\alpha_{X_i,r}, \beta_{X_i,r})}(D_i)$, where $Q_{(\alpha,\beta)}$ is the quantile function (*i.e.*, inverse cumulative distribution function) of the Beta(α, β) distribution.[35] Finally, we sample $Y_i(r)$ from a Bernoulli distribution with parameter p_{ir}. In sum:

$$p_{ir} = Q_{(\alpha_{X_i,r}, \beta_{X_i,r})}(D_i), \qquad Y_i(r) \sim \text{Bernoulli}(p_{ir}), \tag{61}$$

where the expected value for success is $\mathbb{E}[Y_i(r)] = p_{ir}$.

In this study, the human participants can access $Z_i = (X_i, D_i)$. Therefore, they observe the expected individual success score, $\mathbb{E}[Y_i(r)] = p_{ir}$, which is an example of an accurate confidence score.

Classifier's confidence scores. As discussed earlier, the classifier's confidence scores $f_r(X_i)$ for each $i \in \mathcal{I}, r \in \mathcal{R}$ are generated without access to the subset of patient features D_i, and are instead computed using only the observable features X_i and the time slot r. Each score is estimated as the expected value of a Beta distribution parameterized by X_i and time slot r, as follows.

$$f_r(X_i) = \mathbb{E}_{V \sim \text{Beta}(\alpha_{X_i,r}, \beta_{X_i,r})}[V] = \frac{\alpha_{X_i,r}}{\alpha_{X_i,r} + \beta_{X_i,r}}. \tag{62}$$

[35] That is, $Q_{(\alpha,\beta)} \colon [0, 1] \to [0, 1]$ returns the value p such that $\Pr\{V \leq p\} = D_i$ for a random variable $V \sim \text{Beta}(\alpha, \beta)$.

Part V

Supporting Work

Appendix D

DiffWire: Inductive Graph Rewiring via the Lovász Bound

This work is based on the following publication:

> [23] Adrian Arnaiz-Rodriguez, Ahmed Begga, Francisco Escolano, and Nuria Oliver. "DiffWire: Inductive Graph Rewiring via the Lovász Bound". In: *Proceedings of the First Learning on Graphs Conference*. Vol. 198. Proceedings of Machine Learning Research. PMLR, Dec. 2022, 15:1–15:27. URL: https://proceedings.mlr.press/v198/arnaiz-rodri-guez22a.html

D.1 Introduction

Graph Neural Networks (GNNs) [195, 342] are a class of deep learning models applied to graph structured data. They have been shown to achieve state-of-the-art results in many graph-related tasks, such as node and graph classification [190, 245], link prediction [244] and node and graph clustering [91, 379], and in a variety of domains, including image or molecular structure classification, recommender systems and social influence prediction [408].

Most GNNs use a message passing framework and thus are referred to as Message Passing Neural Networks (MPNNs) [190] . In these networks, every node in each layer receives a message from its adjacent neighbors. All the incoming messages at each node are then aggregated and used to update the node's representation via a learnable non-linear function –which is typically implemented by means of a neural network. The final node representations (called node embeddings) are used to perform the graph-related task at hand (*e.g.*, graph classification). MPNNs are extensible, simple and have proven to yield competitive empirical results. Examples of MPNNs include GCN [245], GAT [391], GATv2 [80], GIN [409] and GraphSAGE [203]. However, they typically use transductive learning, *i.e.*, the model observes both the training and testing data during the training phase, which might limit their applicability to graph classification tasks.

However, MPNNs also have important limitations due to the inherent complexity of graphs. Despite such complexity, the literature has reported best results when MPNNs have a small number of layers, because networks with many layers tend to suffer from *over-smoothing* [264] and *over-squashing* [14]. However, this models fail to capture information that depends on

the entire structure of the graph [274] and prevent the information flow to reach distant nodes. This phenomenon is called *under-reaching* [40] and occurs when the MPNN's depth is smaller than the graph's diameter.

Over-smoothing [210, 308, 408, 425] takes place when the embeddings of nodes that belong to different classes become indistinguishable. It tends to occur in MPNNs with many layers that are used to tackle short-range tasks, *i.e.*, tasks where a node's correct prediction mostly depends on its local neighborhood. Given this local dependency, it makes intuitive sense that adding layers to the network would not help the network's performance.

Conversely, long-range tasks require as many layers in the network as the range of the interaction between the nodes. However, as the number of layers in the network increases, the number of nodes feeding into each of the node's receptive field also increases exponentially, leading to *over-squashing* [14, 383]: the information flowing from the receptive field composed of many nodes is compressed in fixed-length node vectors, and hence the graph fails to correctly propagate the messages coming from distant nodes. Thus, over-squashing emerges due to the distortion of information flowing from distant nodes due to graph bottlenecks that emerge when the number of k-hop neighbors grows exponentially with k.

Graph pooling and *graph rewiring* have been proposed in the literature as solutions to address these limitations [14]. Given that the main infrastructure for message passing in MPNNs are the edges in the graph, and given that many of these edges might be noisy or inadequate for the downstream task [390], graph rewiring aims to identify such edges and edit them.

Many graph rewiring methods rely on edge sampling strategies: first, the edges are assigned new weights according to a *relevance function* and then they are re-sampled according to the new weights to retain the most relevant edges (*i.e.*, those with larger weights). Edge relevance might be computed in different ways, including randomly [334], based on similarity [237] or on the edge's curvature [383].

Due to the diversity of possible graphs and tasks to be performed with those graphs, optimal graph rewiring should include a *variety of strategies* that are suited not only to the task at hand but also to the nature and structure of the graph.

Motivation. State-of-the-art edge sampling strategies have three significant **limitations**. First, most of the proposed methods **fail to preserve the global topology of the graph**. Second, most graph rewiring methods are neither **differentiable** nor **inductive** [383]. Third, relevance functions that depend on a diffusion measure (typically in the spectral domain) are **not parameter-free**, which adds a layer of complexity in the models. In this chapter, we address these three limitations.

Contributions and Outline. The main contribution of this work is to propose a theoretical framework called DiffWire for graph rewiring in GNNs that is principled, differentiable, inductive, and parameter-free by leveraging the Lovász bound [274] given by Equation (63). This bound is a mathematical expression of the relationship between the *commute times* (*effective resistance distance*) and the network's *spectral gap*. Given an unseen test graph, DiffWire predicts the optimal graph structure for the task at hand without any parameter tuning. Given the recently reported connection between commute times and curvature [125], and between curvature and the spectral gap [383], the proposed framework provides a unified theory linking these concepts. Our aim is to leverage diffusion and curvature theories

to propose a new approach for graph rewiring that preserves the graph's structure.

We first propose using the CT as a relevance function for edge re-weighting. Moreover, we develop a differentiable, parameter-free layer in the GNN (CT-Layer) to learn the CT. Second, we propose an alternative graph rewiring approach by adding a layer in the network (GAP-Layer) that optimizes the spectral gap according to the nature of the network and the task at hand. Finally, we empirically validate the proposed layers with state-of-the-art benchmark datasets in a graph classification task. We test our approach on a graph classification task to emphasize the inductive nature of DiffWire: the layers in the GNN (CT-Layer or GAP-Layer) are trained to predict the CTs embedding or minimize the spectral gap for unseen graphs, respectively. This approach gives a great advantage when compared to SoTA methods that require optimizing the parameters of the models for each graph. CT-Layer and GAP-Layer learn the weights during training to predict the optimal changes in the topology of any unseen graph in test time. Finally, we also perform preliminary node classification experiments in heterophilic and homophilic graphs using CT-Layer.

The chapter is organized as follows: Appendix D.2 provides a summary of the most relevant related literature. Our core technical contribution is described in Appendix D.3, followed by our experimental evaluation and discussion in Appendix D.4. Finally, Appendix D.5 is devoted to conclusions and an outline of our future lines of research.

D.2 Related Work

In this section we provide an overview of the most relevant works that have been proposed in the literature to tackle the challenges of over-smoothing, over-squashing and under-reaching in MPNNs by means of graph rewiring and pooling.

Graph Rewiring in MPNNs. *Rewiring* is a process of changing the graph's structure to control the information flow and hence improve the ability of the network to perform the task at hand (*e.g.*, node or graph classification, link prediction...). Several approaches have been proposed in the literature for graph rewiring, such as connectivity diffusion [247] or evolution [383], adding new bridge-nodes [47] and multi-hop filters [175], and neighborhood [203], node [313] and edge [334] sampling.

Edge sampling methods sample the graph's edges based on their weights or relevance, which might be computed in different ways. Rong et al. [334] show that randomly dropping edges during training improves the performance of GNNs. Klicpera, Weißenberger, and Günnemann [247], define edge relevance according to the coefficients of a parameterized diffusion process over the graph and then edges are selected using a threshold rule. For Kazi et al. [237], edge relevance is given by the similarity between the nodes' attributes. In addition, a reinforcement learning process rewards edges leading to a correct classification and penalizes the rest.

Edge sampling-based rewiring has been proposed to tackle over-smoothing and over-squashing in MPNNs. Over-smoothing may be relieved by removing inter-class edges [101]. However, this strategy is only valid when the graph is homophilic, *i.e.*, connected nodes tend to share similar attributes. Otherwise, removing these edges could lead to over-squashing [383] if their removal obstructs the message passing between distant nodes belonging to the same class (heterophily). Increasing the size of the bottlenecks of the graph via rewiring has been shown to improve node classification performance in heterophilic graphs, but not in homophilic graphs [383]. Recently, Topping et al. [383] propose an edge relevance function

given by the edge curvature to mitigate over-squashing. They identify the bottleneck of the graph by computing the Ricci curvature of the edges. Next, they remove edges with high curvature and add edges around minimal curvature edges.

Graph Structure Learning (GSL). GSL methods [430] aim to learn an optimized graph structure and its corresponding representations *at the same time*. DiffWire could be seen from the perspective of GSL: CT-Layer, as a metric-based, neural approach, and GAP-Layer, as a direct-neural approach to optimize the structure of the graph to the task at hand.

Graph Pooling. *Pooling* layers simplify the original graph by compressing it into a smaller graph or a vector via pooling operators, which range from simple [289] to more sophisticated approaches, such as DiffPool [413] and MinCut pool [56]. Although graph pooling methods do not consider the edge representations, there is a clear relationship between pooling methods and rewiring since both of them try to quantify the flow of information through the graph's bottleneck.

Positional Encodings (PEs) A Positional Encoding is a feature that describes the global or local position of the nodes in the graph. These features are related to random walk measures and the Laplacian's eigenvectors [328]. Commute Times embeddings (CTEs) may be considered an expressive form of PEs due to their spectral properties, *i.e.*, their relation with the shortest path, the spectral gap or Cheeger constant. Velingker et al. [392] recently proposed use the CTEs as PE or commute times (CT) as edge feature. They pre-compute the CTEs and CT and add it as node or edge features to improve the structural expressiveness of the GNN. PEs are typically pre-computed and then used to build more expressive graph architectures, either by concatenating them to the node features or by building transformer models [135, 267]. Our work is related to PEs as CT-Layer learns the original PEs from the input $\mathbf{X}$ and the adjacency matrix $\mathbf{A}$ instead of pre-computing and potentially modifying them, as previous works do [135, 262, 267, 392]. Thus, CT-Layer may be seen as a method to automatically learn the PEs for graph rewiring.

D.3 DiffWire: Inductive Graph Rewiring

DiffWire provides a unified theory for graph rewiring by proposing two new, complementary layers in MPNNs: first, CT-Layer, a layer that learns the commute times and uses them as a relevance function for edge re-weighting; and second, GAP-Layer, a layer to optimize the spectral gap, depending on the nature of the network and the task at hand.

This section presents the theoretical foundations for the definitions of CT-Layer and GAP-Layer. First, we introduce the bound that our approach is based on: The Lovász bound. Table 23 in Appendix D.6.1 summarizes the notation used in the chapter.

D.3.1 The Lovász Bound

The Lovász bound, given by Equation (63), was derived by Lovász in [274] as a means of linking the spectrum governing a random walk in an undirected graph $G = (V, E)$ with the *hitting time* H_{uv} between any two nodes u and v of the graph. H_{uv} is the expected number of steps needed to reach (or hit) v from u; H_{vu} is defined analogously. The sum of both hitting times between the two nodes, v and u, is the *commute time* $CT_{uv} = H_{uv} + H_{vu}$. Thus, CT_{uv}

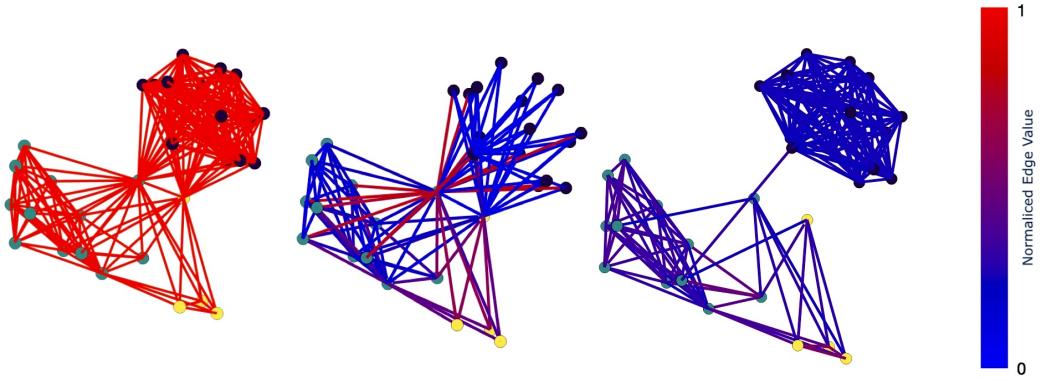


Figure 42. DIFFWIRE. Left: Original graph from COLLAB (test set). Center: Rewired graph after CT-LAYER. Right: Rewired graph after GAP-LAYER. Colors indicate the strength of the edges.

is the expected number of steps needed to hit v from u and go back to u. According to the Lovász bound:

$$\left|\frac{1}{vol(G)}CT_{uv} - \left(\frac{1}{d_u} + \frac{1}{d_v}\right)\right| \leq \frac{1}{\lambda'_2}\frac{2}{d_{min}} \tag{63}$$

where $\lambda'_2 \geq 0$ is the *spectral gap*, *i.e.*, the first non-zero eigenvalue of $\mathcal{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ (normalized Laplacian [108], where $\mathbf{D}$ is the degree matrix and $\mathbf{A}$, the adjacency matrix); $vol(G)$ is the volume of the graph (sum of degrees); d_u and d_v are the degrees of nodes u and v, respectively; and d_{min} is the minimum degree of the graph.

The term $CT_{uv}/vol(G)$ in Equation (63) is referred to as the *effective resistance*, R_{uv}, between nodes u and v. The bound states that the effective resistance between two nodes in the graph converges to or diverges from $(1/d_u + 1/d_v)$, depending on whether the graph's spectral gap diverges from or tends to zero. The larger the spectral gap, the closer $CT_{uv}/vol(G)$ will be to $\frac{1}{d_u} + \frac{1}{d_v}$ and hence the less informative the commute times will be.

We propose two novel GNNs layers based on each side of the inequality in Equation (63): CT-LAYER, focuses on the left-hand side, and GAP-LAYER, on the right-hand side. The use of each layer depends on the nature of the network and the task at hand. In a graph classification task (our focus), CT-LAYER is expected to yield good results when the graph's spectral gap is small; conversely, GAP-LAYER would be the layer of choice in graphs with large spectral gap.

The Lovász bound was later refined by Luxburg, Radl, and Hein [279]. Appendix D.6.2 presents this bound along with its relationship with R_{uv} as a global measure of node similarity. Once we have defined both sides of the Lovász bound, we proceed to describe their implications for graph rewiring.

D.3.2 CT-Layer: Commute Times for Graph Rewiring

We focus first on the left-hand side of the Lovász bound which concerns the effective resistances $CT_{uv}/vol(G) = R_{uv}$ (or commute times)[36] between any two nodes in the graph.

[36] We use commute times and effective resistances interchangeably as per their use in the literature

Spectral Sparsification leads to Commute Times. Graph sparsification in undirected graphs may be formulated as finding a graph $H = (V, E')$ that is *spectrally similar* to the original graph $G = (V, E)$ with $E' \subset E$. Thus, the spectra of their Laplacians, $\mathbf{L}_G$ and $\mathbf{L}_H$ should be similar.

Theorem D.1 (Spielman and Srivastava [358])**.** *Let* `Sparsify(G, q) -> G'` *be a sampling algorithm of graph* $G = (V, E)$*, where edges* $e \in E$ *are sampled with probability* $q \propto R_e$ *(proportional to the effective resistance). For* $n = |V|$ *sufficiently large and* $1/\sqrt{n} < \epsilon \leq 1$*,* $O(n \log n/\epsilon^2)$ *samples are needed to satisfy* $\forall \mathbf{x} \in \mathbb{R}^n : \ (1-\epsilon)\mathbf{x}^T\mathbf{L}_G\mathbf{x} \leq \mathbf{x}^T\mathbf{L}_{G'}\mathbf{x} \leq (1+\epsilon)\mathbf{x}^T\mathbf{L}_G\mathbf{x}$ *, with probability* $\geq 1/2$*.*

The above theorem has a simple explanation in terms of Dirichlet energies, $\mathcal{E}(\mathbf{x})$. The Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A} \succcurlyeq 0$, *i.e.*, it is positive semi-definite (all its eigenvalues are non-negative). Then, if we consider $\mathbf{x} : V \to \mathbb{R}$ as a real-valued function of the n nodes of $G = (V, E)$, we have that $\mathcal{E}(\mathbf{x}) := \mathbf{x}^T\mathbf{L}_G\mathbf{x} = \sum_{e=(u,v)\in E}(\mathbf{x}_u - \mathbf{x}_v)^2 \geq 0$ for any $\mathbf{x}$. In particular, the eigenvectors $\mathbf{f} := \{\mathbf{f}_i : \mathbf{L}\mathbf{f}_i = \lambda_i\mathbf{f}_i\}$ are the set of special functions that minimize the energies $\mathcal{E}(\mathbf{f}_i)$, *i.e.*, they are the mutually orthogonal and normalized functions with the minimal variabilities achievable by the topology of G. Therefore, any minimal variability of G' is bounded by $(1 \pm \epsilon)$ times that of G if we sample enough edges with probability $q \propto R_e$. In addition, $\lambda_i = \frac{\mathcal{E}(\mathbf{f}_i)}{\mathbf{f}_i^T\mathbf{f}_i}$.

This first result implies that edge sampling based on commute times is a principled way to rewire a graph while preserving its original structure and it is bounded by the Dirichlet energies. Next, we present what the commute times embedding is and how it can be spectrally computed.

Commute Times Embedding (CTE) The choice of effective resistances in Theorem D.1 is explained by the fact that R_{uv} can be computed from $R_{uv} = (\mathbf{e}_u - \mathbf{e}_v)^T\mathbf{L}^+(\mathbf{e}_u - \mathbf{e}_v)$, where $\mathbf{e}_u$ is the unit vector with a unit value at u and zero elsewhere. $\mathbf{L}^+ = \sum_{i\geq 2}\lambda_i^{-1}\mathbf{f}_i\mathbf{f}_i^T$, where $\mathbf{f}_i, \lambda_i$ are the eigenvectors and eigenvalues of $\mathbf{L}$, is the pseudo-inverse or Green's function of $G = (V, E)$ if it is connected. The Green's function leads to envision R_{uv} (and therefore CT_{uv}) as *metrics* relating pairs of nodes of G. As a result, the CTE will preserve the commute times distance in a Euclidean space. Note that this latent space of the nodes can not only be described spectrally but also in a *parameter free*-manner, which is not the case for other spectral embeddings, such as heat kernel or diffusion maps as they rely on a time parameter t. More precisely, the embedding matrix $\mathbf{Z}$ whose columns contain the nodes' commute times embeddings is spectrally given by:

$$\mathbf{Z} := \sqrt{vol(G)}\Lambda^{-1/2}\mathbf{F}^T = \sqrt{vol(G)}\Lambda'^{-1/2}\mathbf{G}^T\mathbf{D}^{-1/2} \tag{64}$$

where Λ is the diagonal matrix of the unnormalized Laplacian $\mathbf{L}$ eigenvalues and $\mathbf{F}$ is the matrix of their associated eigenvectors. Similarly, Λ' contains the eigenvalues of the normalized Laplacian $\mathcal{L}$ and $\mathbf{G}$ the eigenvectors. We have $\mathbf{F} = \mathbf{G}\mathbf{D}^{-1/2}$ or $\mathbf{f}_i = \mathbf{g}_i\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is the degree matrix.

Finally, the commute times are given by the Euclidean distances between the embeddings $CT_{uv} = \|\mathbf{z}_u - \mathbf{z}_v\|^2$. The spectral calculation of commute times distances is given by:

$$R_{uv} = \frac{CT_{uv}}{vol(G)} = \frac{\|\mathbf{z}_u - \mathbf{z}_v\|^2}{vol(G)} = \sum_{i=2}^{n}\frac{1}{\lambda_i}\left(\mathbf{f}_i(u) - \mathbf{f}_i(v)\right)^2 = \sum_{i=2}^{n}\frac{1}{\lambda'_i}\left(\frac{\mathbf{g}_i(u)}{\sqrt{d_u}} - \frac{\mathbf{g}_i(v)}{\sqrt{d_v}}\right)^2 \tag{65}$$

Commute Times as an Optimization Problem. In this section, we demonstrate how the CTs may be computed as an optimization problem by means of a differentiable layer in a GNN. Constraining neighboring nodes to have a similar embedding leads to

$$\mathbf{Z} = \arg\min_{\mathbf{Z}^T\mathbf{Z}=\mathbf{I}} \frac{\sum_{u,v} \|\mathbf{z}_u - \mathbf{z}_v\|^2 \mathbf{A}_{uv}}{\sum_{u,v} \mathbf{Z}_{uv}^2 d_u} = \frac{\sum_{(u,v)\in E} \|\mathbf{z}_u - \mathbf{z}_v\|^2}{\sum_{u,v} \mathbf{Z}_{uv}^2 d_u} = \frac{Tr[\mathbf{Z}^T\mathbf{L}\mathbf{Z}]}{Tr[\mathbf{Z}^T\mathbf{D}\mathbf{Z}]}\,, \tag{66}$$

which reveals that CTs embeddings result from a Laplacian regularization down-weighted by the degree. As a result, *frontier* nodes or hubs –*i.e.*, nodes with inter-community edges– which tend to have larger degrees than those lying inside their respective communities will be embedded far away from their neighbors, increasing the *distance* between communities. Note that the above *quotient of traces* formulation is easily differentiable and different from $Tr[\frac{\mathbf{Z}^T\mathbf{L}\mathbf{Z}}{\mathbf{Z}^T\mathbf{D}\mathbf{Z}}]$ proposed in [325].

With the above elements we define CT-Layer, the first rewiring layer proposed in this chapter. See Figure 43 for a graphical representation of the layer.

Definition D.1 (CT-Layer)**.** *Given the matrix* $\mathbf{X}_{n\times F}$ *encoding the features of the nodes after any message passing (MP) layer,* $\mathbf{Z}_{n\times O(n)} = \tanh(\text{MLP}(\mathbf{X}))$ *learns the association* $\mathbf{X} \to \mathbf{Z}$ *while* $\mathbf{Z}$ *is optimized according to the loss* $L_{CT} = \frac{Tr[\mathbf{Z}^T\mathbf{L}\mathbf{Z}]}{Tr[\mathbf{Z}^T\mathbf{D}\mathbf{Z}]} + \left\| \frac{\mathbf{Z}^T\mathbf{Z}}{\|\mathbf{Z}^T\mathbf{Z}\|_F} - \mathbf{I}_n \right\|_F$*. This results in the following* resistance diffusion $\mathbf{T}^{CT} = \mathbf{R}(\mathbf{Z}) \odot \mathbf{A}$, i.e., *the Hadamard product between the resistance distance and the adjacency matrix, providing as input to the subsequent MP layer a learnt convolution matrix. We set* $\mathbf{R}(\mathbf{Z})$ *to the pairwise Euclidean distances of the node embeddings in* $\mathbf{Z}$ *divided by* $vol(G)$*.*

Thus, CT-Layer learns the CTs and rewires an input graph according to them: the edges with maximal resistance will tend to be the most important edges so as to preserve the topology of the graph.

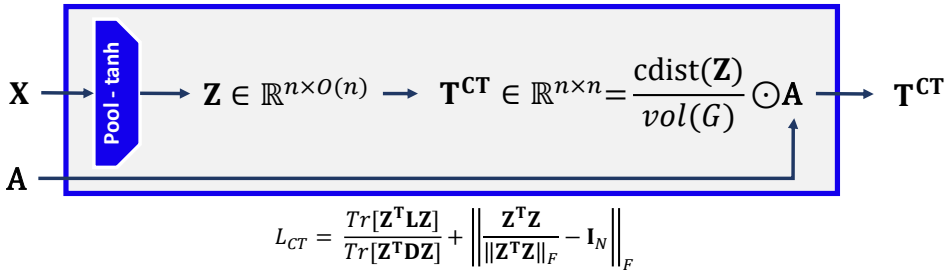


Figure 43. Detailed depiction of CT-Layer, where `cdist` refers to the matrix of pairwise Euclidean distances between the node embeddings in $\mathbf{Z}$.

Below, we present the relationship between the CTs and the graph's bottleneck and curvature.

$\mathbf{T}^{CT}$ and Graph Bottlenecks. Beyond the principled sparsification of $\mathbf{T}^{CT}$ (Theorem D.1), this layer rewires the graph $G = (E, V)$ in such a way that edges with maximal resistance will tend to be the most critical to preserve the topology of the graph. More precisely,

although $\sum_{e \in E} R_e = n - 1$, the bulk of the resistance distribution will be located at graph bottlenecks, if they exist. Otherwise, their magnitude is upper-bounded and the distribution becomes more uniform.

Graph bottlenecks are controlled by the *graph's conductance* or Cheeger constant, $h_G = min_{S \subseteq V} h_S$, where: $h_S = \frac{|\partial S|}{\min(vol(S), vol(\bar{S}))}$, $\partial S = \{e = (u, v) : u \in S, v \in \bar{S}\}$ and $vol(S) = \sum_{u \in S} d_u$.

The interplay between the graph's conductance and effective resistances is given by:

Theorem D.2 (Alev et al. [12]). *Given a graph $G = (V, E)$, a subset $S \subseteq V$ with $vol(S) \leq vol(G)/2$,*

$$h_S \geq \frac{c}{vol(S)^{1/2-\epsilon}} \iff |\partial S| \geq c \cdot vol(S)^{1/2-\epsilon}, \tag{67}$$

for some constant c and $\epsilon \in [0, 1/2]$. Then, $R_{uv} \leq \left(\frac{1}{d_u^{2\epsilon}} + \frac{1}{d_v^{2\epsilon}}\right) \cdot \frac{1}{\epsilon \cdot c^2}$ for any pair u, v.

According to this theorem, the larger the graph's bottleneck, the tighter the bound on R_{uv} are. Moreover, $\max(R_{uv}) \leq 1/h_S^2$, *i.e.*, the resistance is bounded by the square of the bottleneck.

This bound partially explains the rewiring of the graph in Figure 42-center. As seen in the Figure 42-center, rewiring using CT-Layer sparsifies the graph and assigns larger weights to the edges located in the graph's bottleneck. The interplay between Theorem D.2 and Theorem D.1 is described in Appendix D.6.1.

Recent work has proposed using curvature for graph rewiring. We outline below the relationship between CTs and curvature.

Effective Resistances and Curvature. Topping et al. [383] propose an approach for graph rewiring, where the relevance function is given by the Ricci curvature. However, this measure is non-differentiable. More recent definitions of curvature [125] have been formulated based on resistance distances that would be differentiable using our approach. The resistance curvature of an edge $e = (u, v)$ is $\kappa_{uv} := 2(p_u + p_v)/R_{uv}$ where $p_u := 1 - \frac{1}{2}\sum_{u \sim w} R_{uv}$ is the node's curvature. Relevant properties of the edge resistance curvature are discussed in Appendix D.6.1, along with a related Theorem proposed in Devriendt and Lambiotte [125].

D.3.3 GAP-Layer: Spectral Gap Optimization for Graph Rewiring

The right-hand side of the Lovász bound in Equation (63) relies on the graph's spectral gap λ_2', such that the larger the spectral gap, the closer the commute times would be to their non-informative regime. Note that the spectral gap is typically large in commonly observed graphs –such as communities in social networks which may be bridged by many edges [1]– and, hence, in these cases it would be desirable to rewire the adjacency matrix $\mathbf{A}$ so that λ_2' is minimized.

In this section, we explain how to rewire the graph's adjacency matrix A to minimize the spectral gap. We propose using the gradient of λ_2 wrt each component of $\tilde{\mathbf{A}}$. Then, we can compute these gradient either using Laplacians ($\mathbf{L}$, with Fiedler λ_2) or normalized Laplacians ($\mathcal{L}$, with Fiedler λ_2'). We also present an approximation of the Fiedler vectors needed to compute those gradients, and propose computing them as a GNN Layer called the GAP-Layer. A detailed schematic of GAP-Layer is shown in Figure 44.

Rewiring using a Ratio-cut (Rcut) Approximation. We propose to rewire the adjacency matrix, $\mathbf{A}$, so that λ_2 is minimized. We consider a matrix $\tilde{\mathbf{A}}$ close to $\mathbf{A}$ that satisfies $\tilde{\mathbf{L}}\mathbf{f}_2 = \lambda_2\mathbf{f}_2$, where $\mathbf{f}_2$ is the solution to the ratio-cut relaxation [83]. Following [233], the gradient of λ_2 wrt each component of $\tilde{\mathbf{A}}$ is given by

$$\nabla_{\tilde{\mathbf{A}}}\lambda_2 := Tr\left[(\nabla_{\tilde{\mathbf{L}}}\lambda_2)^T \cdot \nabla_{\tilde{\mathbf{A}}}\tilde{\mathbf{L}}\right] = \text{diag}(\mathbf{f}_2\mathbf{f}_2^T)\mathbf{1}\mathbf{1}^T - \mathbf{f}_2\mathbf{f}_2^T \tag{68}$$

where $\mathbf{1}$ is the vector of n ones; and $[\nabla_{\tilde{\mathbf{A}}}\lambda_2]_{ij}$ is the gradient of λ_2 wrt $\tilde{\mathbf{A}}_{uv}$. The driving force of this gradient relies on the correlation $\mathbf{f}_2\mathbf{f}_2^T$. Using this gradient to minimize λ_2 results in breaking the graph's bottleneck while preserving simultaneously the inter-cluster structure. We delve into this matter in Appendix D.6.2.

Rewiring using a Normalized-cut (Ncut) Approximation. Similarly, considering now λ_2' for rewiring leads to

$$\begin{aligned}\nabla_{\tilde{\mathbf{A}}}\lambda_2' := Tr\left[(\nabla_{\tilde{\mathcal{L}}}\lambda_2)^T \cdot \nabla_{\tilde{\mathbf{A}}}\tilde{\mathcal{L}}\right] \quad = \\ \mathbf{d}'\left\{\mathbf{g}_2^T\tilde{\mathbf{A}}^T\tilde{\mathbf{D}}^{-1/2}\mathbf{g}_2\right\}\mathbf{1}^T + \mathbf{d}'\left\{\mathbf{g}_2^T\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{g}_2\right\}\mathbf{1}^T \quad + \quad \tilde{\mathbf{D}}^{-1/2}\mathbf{g}_2\mathbf{g}_2^T\tilde{\mathbf{D}}^{-1/2}\end{aligned} \tag{69}$$

where $\mathbf{d}'$ is a $n \times 1$ vector including derivatives of degree wrt adjacency and related terms. This gradient relies on the Fiedler vector $\mathbf{g}_2$ (the solution to the normalized-cut relaxation), and on the incoming and outgoing one-hop random walks. This approximation breaks the bottleneck while preserving the global topology of the graph (Figure 42-left). Proof and details are included in Appendix D.6.2.

We present next an approximation of the Fiedler vector, followed by a proposed new layer in the GNN called the GAP-Layer to learn how to minimize the spectral gap of the graph.

Approximating the Fiedler vector. Given that $\mathbf{g}_2 = \tilde{\mathbf{D}}^{1/2}\mathbf{f}_2$, we can obtain the normalized-cut gradient in terms of $\mathbf{f}_2$. From [210] we have that

$$\mathbf{f}_2(u) = \begin{cases} +1/\sqrt{n} & \text{if } u \text{ belongs to the first cluster} \\ -1/\sqrt{n} & \text{if } u \text{ belongs to the second cluster} \end{cases} + O\left(\frac{\log n}{n}\right) \tag{70}$$

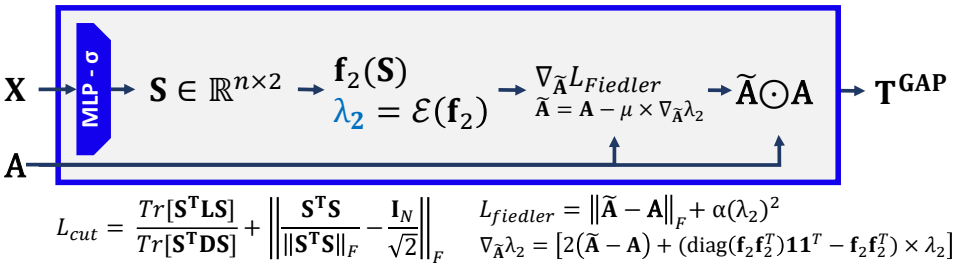


Figure 44. GAP-Layer (Rcut). For GAP-Layer (Ncut), substitute $\nabla_{\tilde{\mathbf{A}}}\lambda_2$ by Appendix D.3.3

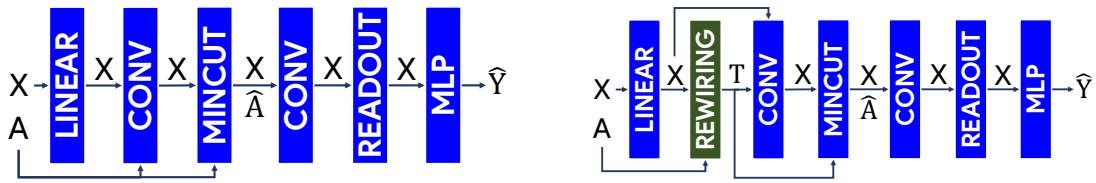
Definition D.2 (GAP-Layer)**.** *Given the matrix $\mathbf{X}_{n\times F}$ encoding the features of the nodes after any message passing (MP) layer, $\mathbf{S}_{n\times 2} = Softmax(\text{MLP}(\mathbf{X}))$ learns the association*

$\mathbf{X} \to \mathbf{S}$ *while* $\mathbf{S}$ *is optimized according to the loss* $L_{Cut} = -\frac{Tr[\mathbf{S}^T\mathbf{A}\mathbf{S}]}{Tr[\mathbf{S}^T\mathbf{D}\mathbf{S}]} + \left\| \frac{\mathbf{S}^T\mathbf{S}}{\|\mathbf{S}^T\mathbf{S}\|_F} - \frac{\mathbf{I}_n}{\sqrt{2}} \right\|_F$. *Then the Fiedler vector* $\mathbf{f}_2$ *is approximated by appyling a softmaxed version of Appendix D.3.3 and considering the loss* $L_{Fiedler} = \|\tilde{\mathbf{A}} - \mathbf{A}\|_F + \alpha(\lambda_2^*)^2$, *where* $\lambda_2^* = \lambda_2$ *if we use the ratio-cut approximation (and gradient) and* $\lambda_2^* = \lambda_2'$ *if we use the normalized-cut approximation and gradient. This returns* $\tilde{\mathbf{A}}$ *and the GAP diffusion* $\mathbf{T}^{GAP} = \tilde{\mathbf{A}}(\mathbf{S}) \odot \mathbf{A}$ *results from minimizing* $L_{GAP} := L_{Cut} + L_{Fiedler}$.

D.4 Experiments and Discussion

D.4.1 Graph Classification

In this section, we study the properties and performance of CT-Layer and GAP-Layer in a graph classification task with several benchmark datasets. To illustrate the merits of our approach, we compare CT-Layer and GAP-Layer with 3 state-of-the-art diffusion and curvature-based graph rewiring methods. Note that the aim of the evaluation is to shed light on the properties of both layers and illustrate their inductive performance, not to perform a benchmark comparison with all previously proposed graph rewiring methods.



(a) MinCut baseline **(b)** CT-Layer or GAP-Layer

Figure 45. GNN models used in the experiments. Left: MinCut Baseline model. Right: CT-Layer or GAP-Layer models, depending on what method is used for rewiring.

Baselines:. The first baseline architecture is based on **MinCut Pool** [56] and it is shown in Figure 45a. It is the base GNN that we use for graph classification without rewiring. MinCut Pool layer learns $(\mathbf{A}_{n\times n}, \mathbf{X}_{n\times F}) \to (\mathbf{A}'_{k\times k}, \mathbf{X}_{k\times F})$, being $k < n$ the new number of node clusters. The first baseline strategy using graph rewiring is k-NN graphs [322], where weights of the edges are computed based on feature similarity. The next two baselines are graph rewiring methods that belong to the same family of methods as DiffWire, *i.e.*, methods based on diffusion and curvature, namely **DIGL** (PPR) [247] and **SDRF** [383]. DIGL is a diffusion-based preprocessing method within the family of metric-based GSL approaches. We set the teleporting probability $\alpha = 0.001$ and ϵ is set to keep the same average degree for each graph. Once preprocessed with DIGL, the graphs are provided as input to the MinCut Pool (Baseline1) arquitecture. The third baseline model is SDRF, which performs curvature-based rewiring. SDRF is also a preprocessing method which has 3 parameters that are highly graph-dependent. We set these parameters to $\tau = 20$ and $C^+ = 0$ for all experiments as per [383]. The number of iterations is estimated dynamically according to $0.7 * |V|$ for each graph.

Both DIGL and SDRF aim to preserve the global topology of the graph but require optimizing their parameters for each input graph via hyper-parameter search. In a graph

Table 18. Experimental results on common graph classification benchmarks. **Red** denotes the best model row-wise and **Blue** marks the runner-up. '*' means degree as node feature.

	MinCutPool	k-NN	DIGL	SDRF	CT-Layer	GAP-Layer (R)	GAP-Layer (N)
REDDIT-B*	66.53±4.4	64.40±3.8	76.02±4.3	65.3±7.7	**78.45**±4.5	**77.63**±4.9	76.00±5.3
IMDB-B*	60.75±7.0	55.20±4.3	59.35±7.7	59.2±6.9	**69.84**±4.6	**69.93**±3.3	68.80±3.1
COLLAB*	58.00±6.2	58.33±11	57.51±5.9	56.60±10	**69.87**±2.4	64.47±4.0	**65.89**±4.9
MUTAG	84.21±6.3	**87.58**±4.1	85.00±5.6	82.4±6.8	**87.58**±4.4	**86.90**±4.0	**86.90**±4.0
PROTEINS	74.84±2.3	**76.76**±2.5	74.49±2.8	74.4±2.7	**75.38**±2.9	75.03±3.0	**75.34**±2.1
SBM*	53.00±9.9	50.00±0.0	56.93±12	54.1±7.1	81.40±11	**90.80**±7.0	**92.26**±2.9
Erdös-Rényi*	81.86±6.2	63.40±3.9	**81.93**±6.3	73.6±9.1	79.06±9.8	79.26±10	**82.26**±3.2

classification task, this search is $O(n^3)$ per graph. Details about the parameter tuning in these methods can be found in Appendix D.6.3.

To shed light on the performance and properties of CT-Layer and GAP-Layer, we add the corresponding layer in between Linear($\mathbf{X}$) $\xrightarrow{*}$ Conv1($\mathbf{A}, \mathbf{X}$). We build 3 different models: CT-Layer, GAP-Layer (Rcut), GAP-Layer (Ncut), depending on the layer used. For CT-Layer, we learn $\mathbf{T}^{CT}$ which is used as a convolution matrix afterwards. For GAP-Layer, we learn $\mathbf{T}^{GAP}$ either using the Rcut or the Ncut approximations. A schematic of the architectures is shown in Figure 45b and in Appendix D.6.3.

As shown in Table 18, we use in our experiments common benchmark datasets for graph classification. We select datasets both with features and featureless, in which case we use the degree as the node features. These datasets are diverse regarding the topology of their networks: Reddit-B, Imdb-B and Collab contain truncate scale-free graphs (social networks), whereas Mutag and Proteins contain graphs from biology or chemistry. In addition, we use two synthetic datasets with 2 classes: Erdös-Rényi with $p_1 \in [0.3, 0.5]$ and $p_2 \in [0.4, 0.8]$ and Stochastic block model (SBM) with parameters $p_1 = 0.8$, $p_2 = 0.5$, $q_1 \in [0.1, 0.15]$ and $q_2 \in [0.01, 0.1]$. More details about the datasets in Appendix D.6.3. In addition, Table 18 reports average accuracies and standard deviation on 10 random data splits, using 85/15 stratified train-test split, training during 60 epochs and reporting the results of the last epoch for each random run. We use Pytorch Geometric [170] and the code is available in a public repository[37].

The experiments support our hypothesis that rewiring based on CT-Layer and GAP-Layer improves the performance of the baselines on graph classification. Since both layers are differentiable, they learn how to inductively rewire unseen graphs. The improvements are significant in graphs where social components arise (RedditB, ImdbB, Collab), *i.e.*, graphs with small world properties and power-law degree distributions with a topology based on hubs and authorities. These are graphs where bottlenecks arise easily and our approach is able to properly rewire the graphs. However, the improvements observed in planar or grid networks (Mutag and Proteins) are more limited: the bottleneck does not seem to be critical for the graph classification task.

Moreover, CT-Layer and GAP-Layer perform better in graphs with featureless nodes than graphs with node features because it is able to leverage the information encoded in the topology of the graphs. Note that in attribute-based graphs, the weights of the attributes typically overwrite the graph's structure in the classification task, whereas in graphs without node features, the information is encoded in the graph's structure. Thus, k-NN rewiring

[37] https://github.com/AdrianArnaiz/DiffWire

outperforms every other rewiring method in graph classification where graphs has node features.

Appendix D.6.3 contains an in-depth analysis of the comparison between the spectral node CT embeddings (CTEs) given by Equation (64), and the learned node CTEs as predicted by CT-Layer. We find that the CTEs that are learned in CT-Layer are able to better preserve the original topology of the graph while shifting the distribution of the effective resistances of the edges towards an asymmetric distribution where few edges have very large weights and a majority of edges have low weights.

In addition, Appendix D.6.3 also includes the analysis of the graphs latent space of the readout layer produced by each model. Finally, we analyze the performance of the proposed layers in graphs with different structural properties in Appendix D.6.3. We analyze the correlation between accuracy, the graph's assortativity, and the graph's bottleneck (λ_2).

CT-Layer vs GAP-Layer. The datasets explored in this chapter are characterized by mild bottlenecks from the perspective of the Lovász bound. For completion, we have included two synthetic datasets (Stochastic Block Model and Erdös-Rényi) where the Lovász bound is very restrictive. As a result, CT-Layer is outperformed by GAP-Layer in SBM. Note that the results on the synthetic datasets suffer from large variability. As a general rule of thumb, the smaller the graph's bottleneck (defined as the ratio between the number of inter-community edges and the number of intra-community edges), the more useful the CT-Layer is because the rewired graph will be sparsified in the communities but will preserve the edges in the gap. Conversely, the larger the bottleneck, the more useful the GAP-Layer is.

D.4.2 Node Classification using CT-Layer

CT-Layer and GAP-Layer are mainly designed to perform graph classification tasks. However, we identify two potential areas to apply CT-Layer for node classification.

First, the new $\mathbf{T}^{\mathbf{CT}}$ diffusion matrix learned by CT-Layer gives more importance to edges that connect different communities, *i.e.*, edges that connect distant nodes in the graph. This behaviour of CT-Layer is aligned to solve long-range and *heterophilic* node classification tasks using fewer number of layers and thus avoiding under-reaching, over-smoothing and over-squashing.

Second, there is an increasingly interest in the community in using PEs in the nodes to develope more expressive GNN. PEs tend to help in node classification in *homophilic* graphs, as nearby nodes will be assigned similar PEs. However, the main limitation is that PEs are usually pre-computed before the GNN training due to their high computational cost. CT-Layer provides a solution to this problem, as it *learns* to predict the commute times embedding ($\mathbf{Z}$) of a given graph (see Figure 43 and Definition D.1). Hence, CT-Layer is able to learn and predict PEs from $\mathbf{X}$ and $\mathbf{A}$ inside a GNN without needing to pre-compute them.

We empirically validate CT-Layer in a node classification task on benchmark homophilic (Cora, Pubmed and Citeseer) and heterophilic (Cornell, Actor and Wisconsin) graphs. The results are depicted in Table 19 comparing three models: (1) the baseline model consists of a 1-layer-GCN; (2) *model 1* is a 1-layer-GCN where the CTEs are concatenated to the node features as PEs ($\mathbf{X} \,\|\, \mathbf{Z}$); (3) Finally, *model 2* is a 1-layer-GCN where $\mathbf{T}^{\mathbf{CT}}$ is used as a diffusion matrix ($\mathbf{A} = \mathbf{T}^{\mathbf{CT}}$). More details can be found in Appendix D.6.3.

As seen in the Table, the proposed models outperform the baseline GCN model: using CTEs as features (model 1) yields competitive results in homophilic graphs, whereas using $\mathbf{T}^{\mathbf{CT}}$ as a matrix for message passing (model 2) performs well in heterophilic graphs. Note that in our experiments, the CTEs are learned by CT-Layer instead of being precomputed. A promising direction of future work would be to explore how to combine these two approaches (model 1 and model 2) to leverage the best of each of the methods on a wide range of graphs for node classification tasks.

Table 19. Results in node classification

Dataset	GCN (baseline)	*model 1:* $\mathbf{X} \parallel \mathbf{Z}$	*model 2:* $\mathbf{A} = \mathbf{T}^{\mathbf{CT}}$	Homophily
Cora	$82.01_{\pm 0.8}$	$\mathbf{83.66}_{\pm 0.6}$	$67.96_{\pm 0.8}$	*81.0%*
Pubmed	$81.61_{\pm 0.3}$	$\mathbf{86.07}_{\pm 0.1}$	$68.19_{\pm 0.7}$	*80.0%*
Citeseer	$70.81_{\pm 0.5}$	$\mathbf{72.26}_{\pm 0.5}$	$66.71_{\pm 0.6}$	*73.6%*
Cornell	$59.19_{\pm 3.5}$	$58.02_{\pm 3.7}$	$\mathbf{69.04}_{\pm 2.2}$	*30.5%*
Actor	$29.59_{\pm 0.4}$	$29.35_{\pm 0.4}$	$\mathbf{31.98}_{\pm 0.3}$	*21.9%*
Wisconsin	$68.05_{\pm 6.2}$	$69.25_{\pm 5.1}$	$\mathbf{79.05}_{\pm 2.1}$	*19.6%*

D.5 Conclusion and Future Work

In this chapter, we have proposed DiffWire, a unified framework for graph rewiring that links the two components of the Lovász bound: CTs and the spectral gap. We have presented two novel, fully differentiable and inductive rewiring layers: CT-Layer and GAP-Layer. We have empirically evaluated these layers on benchmark datasets for graph classification with competitive results when compared to SoTA baselines, especially in graphs where the nodes have no attributes and have small-world properties. We have also performed preliminary experiments in a node classification task, showing that the CT Embeddings and the CT distances benefit GNN architectures in homophilic and heterophilic graphs, respectively.

In future work, we plan to test the proposed approach in other graph-related tasks and intend to apply DiffWire to large-scale graphs and real-world applications, particularly in social networks, which have unique topology, statistics, and direct implications in society.

D.6 Additional Content and Experiments

In this section, we provide additional analysis of the proposed GNN layers, and supplementary experiments to shed light on the empirical behavior of them. First, we provide an analysis of the diffusion and its relationship with curvature in Appendix D.6.1. Secondly, we study in detail GAP-Layer and the implications of the proposed spectral gradients in Appendix D.6.2. In addition, Appendix D.6.3 reports statistics and characteristics of the datasets used in the experimental section, provides more information about the experiments results, describes additional experimental results, and includes a summary of the computing infrastructure used in our experiments. Finally, we include Table 23 with the notation used in Appendix D.

D.6.1 In-detail Analysys of CT-Layer

Analysis of Commute Times rewiring

First, we provide an answer to the following question:

Is resistance diffusion via $\mathbf{T}^{CT}$ *a principled way of preserving the Cheeger constant?*

We answer the question above by linking Theorems D.1 and D.2 in Appendix D with the Lovász bound. The outline of our explanation follows three steps.

- **Proposition 1:** Theorem D.1 (**Sparsification**) provides a principled way to bias the adjacency matrix so that the edges with the largest weights in the rewired graph correspond to the edges in graph's bottleneck.
- **Proposition 2:** Theorem D.2 (**Cheeger vs Resistance**) can be used to demonstrate that increasing the effective resistance leads to a mild reduction of the Cheeger constant.
- **Proposition 3:** (Conclusion) The effectiveness of the above theorems to contain the Cheeger constant is constrained by the Lovász bound.

Next, we provide a thorough explanation of each of the propositions above.

Proposition D.1 (Biasing)**.** *Let G' =* `Sparsify`*(G, q) be a sampling algorithm of graph* $G = (V, E)$*, where edges* $e \in E$ *are sampled with probability* $q \propto R_e$ *(proportional to the effective resistance). This choice is* necessary *to retain the global structure of G,* i.e.*, to satisfy*

$$\forall \mathbf{x} \in \mathbb{R}^n : \ (1-\epsilon)\mathbf{x}^T\mathbf{L}_G\mathbf{x} \leq \mathbf{x}^T\mathbf{L}_{G'}\mathbf{x} \leq (1+\epsilon)\mathbf{x}^T\mathbf{L}_G\mathbf{x} \ , \tag{71}$$

with probability at least $1/2$ *by sampling* $O(n \log n/\epsilon^2)$ *edges , with* $1/\sqrt{n} < \epsilon \leq 1$*, instead of* $O(m)$*, where* $m = |E|$*. In addition, this choice* biases *the uniform distribution in favor of critical edges in the graph.*

Proof. We start by expressing the Laplacian $\mathbf{L}$ in terms of the edge-vertex incidence matrix $\mathbf{B}_{m\times e}$:

$$\mathbf{B}_{eu} = \begin{cases} 1 & \text{if } u \text{ is the head of } e \\ -1 & \text{if } u \text{ is the tail of } e \\ 0 & \text{otherwise .} \end{cases} \tag{72}$$

where edges in undirected graphs are counted once, *i.e.*, $e = (u, v) = (v, u)$. Then, we have $\mathbf{L} = \mathbf{B}^T\mathbf{B} = \sum_e \mathbf{b}_e\mathbf{b}_e^T$, where $\mathbf{b}_e$ is a row vector (*incidence vector*) of $\mathbf{B}$, with $\mathbf{b}_{e=(u,v)} = (\mathbf{e}_u - \mathbf{e}_v)$. In addition, the Dirichlet energies can be expressed as norms:

$$\mathcal{E}(\mathbf{x}) = \mathbf{x}^T\mathbf{L}\mathbf{x} = \mathbf{x}^T\mathbf{B}^T\mathbf{B}\mathbf{x} = \|\mathbf{B}\mathbf{x}\|_2^2 = \sum_{e=(u,v)\in E} (\mathbf{x}_u - \mathbf{x}_v)^2 \ . \tag{73}$$

As a result, the effective resistance R_e between the two nodes of an edge $e = (u, v)$ can be defined as

$$R_e = (\mathbf{e}_u - \mathbf{e}_v)^T\mathbf{L}^+(\mathbf{e}_u - \mathbf{e}_v) = \mathbf{b}_e^T\mathbf{L}^+\mathbf{b}_e \tag{74}$$

Next, we reformulate the spectral constraints in Equation (71), *i.e.*, $(1-\epsilon)\mathbf{L}_G \preccurlyeq \mathbf{L}_{G'} \preccurlyeq (1+\epsilon)\mathbf{L}_G$ as

$$\mathbf{L}_G \preccurlyeq \mathbf{L}_{G'} \preccurlyeq \Gamma\mathbf{L}_G \,, \Gamma = \frac{1+\epsilon}{1-\epsilon} \,. \tag{75}$$

This simplifies the analysis, since the above expression can be interpreted as follows: the Dirichlet energies of $\mathbf{L}_{G'}$ are lower-bounded by those of $\mathbf{L}_G$ and upper-bounded by Γ times the energies of $\mathbf{L}_G$. Considering that the energies define hyper-ellipsoids, the hyper-ellipsoid associated with $\mathbf{L}_{G'}$ is between the hyper-ellipsoids of $\mathbf{L}_G$ and Γ times the $\mathbf{L}_G$.

The hyper-ellipsoid analogy provides a framework to proof that the inclusion relationships are preserved under scaling: $M\mathbf{L}_G M \preccurlyeq M\mathbf{L}_{G'}M \preccurlyeq M\Gamma\mathbf{L}_G M$ where M can be a matrix. In this case, if we set $M := (\mathbf{L}_G^+)^{1/2} = \mathbf{L}_G^{+/2}$ we have:

$$\mathbf{L}_G^{+/2}\mathbf{L}_G\mathbf{L}_G^{+/2} \preccurlyeq \mathbf{L}_G^{+/2}\mathbf{L}_{G'}\mathbf{L}_G^{+/2} \preccurlyeq \mathbf{L}_G^{+/2}\Gamma\mathbf{L}_G^{+/2} \,, \tag{76}$$

which leads to

$$\mathbf{I}_n \preccurlyeq \mathbf{L}_G^{+/2}\mathbf{L}_{G'}\mathbf{L}_G^{+/2} \preccurlyeq \Gamma\mathbf{I}_n \,. \tag{77}$$

We seek a Laplacian $\mathbf{L}_{G'}$ satisfying the *similarity constraints* in Equation (75). Since $E' \subset E$, *i.e.*, we want to remove structurally irrelevant edges, we can design $\mathbf{L}_{G'}$ in terms of considering *all* the edges E:

$$\mathbf{L}_{G'} := \mathbf{B}_G^T\mathbf{B}_G = \sum_e s_e \mathbf{b}_e\mathbf{b}_e^T \tag{78}$$

and let the similarity constraint define the sampling weights and the choice of e (setting $s_e \geq 0$ propertly). More precisely:

$$\mathbf{I}_n \preccurlyeq \mathbf{L}_G^{+/2}\sum_e \mathbf{b}_e\mathbf{b}_e^T\mathbf{L}_G^{+/2} \preccurlyeq \Gamma\mathbf{I}_n \,. \tag{79}$$

Then if we define $\mathbf{v}_e := \mathbf{L}_G^{+/2}\mathbf{b}_e$ as the *projected incidence vector*, we have

$$\mathbf{I}_n \preccurlyeq \sum_e s_e\mathbf{v}_e\mathbf{v}_e^T \preccurlyeq \Gamma\mathbf{I}_n \,. \tag{80}$$

Consequently, a spectral sparsifier must find $s_e \geq 0$ so that the above similarity constraint is satisfied. Since there are m edges in E, s_e must be zero for most of the edges. But, what are the best candidates to retain? Interestingly, the similarity constraint provides the answer. From Equation (74) we have

$$\mathbf{v}_e^T\mathbf{v}_e = \|\mathbf{v}_e\|^2 = \|\mathbf{L}_G^{+/2}\mathbf{b}_e\|_2^2 = \mathbf{b}_e^T\mathbf{L}_G^+\mathbf{b}_e = R_e \,. \tag{81}$$

This result explains why sampling the edges with probability $q \propto R_e$ leads to a ranking of m edges of $G = (V, E)$ such that edges with large $R_e = \|\mathbf{v}_e\|^2$ are preferred[38].

Algorithm 5 implements a deterministic greedy version of `Sparsify`(G, q), where we build incrementally $E' \subset E$ by creating a budget of decreasing resistances $R_{e_1} \geq R_{e_2} \geq \ldots \geq R_{e_{O(n\log n/\epsilon^2)}}$. □

[38] Although some of the elements of this section are derived from [46], we note that the Nikhil Srivastava's lectures at The Simons Institute (2014) are by far more clarifying.

Note that this rewiring strategy preserves the spectral similarities of the graphs, *i.e.*, the global structure of $G = (V, E)$ is captured by $G' = (V, E')$.

Moreover, the maximum R_e in each graph determines an upper bound on the Cheeger constant and hence an upper bound on the size of the graph's bottleneck, as per the following proposition.

Algorithm 5: GREEDYSparsify

Input : $G = (V, E)$,$\epsilon \in (1/\sqrt{n}, 1]$, $n = |V|$.
Output: $G' = (V, E')$ with $E' \subset E$ such that $|E'| = O(n \log n/\epsilon^2)$.

$L \leftarrow \text{List}(\{\mathbf{v}_e : e \in E\})$
$Q \leftarrow \text{Sort}(L, \text{descending}, \text{criterion}=\|\mathbf{v}_e\|^2)$ `# Sort candidate edges by descending Resistance`
$E' \leftarrow \emptyset$
$\mathcal{I} \leftarrow \mathbf{0}_{n\times n}$
Repeat
  $\mathbf{v}_e \leftarrow \text{pop}(Q)$ `# Remove the head of the queue`
  $\mathcal{I} \leftarrow \mathcal{I} + \mathbf{v}_e\mathbf{v}_e^T$
  if $\mathcal{I} \preccurlyeq \Gamma\mathbf{I}_n$ **then**
    $E' \leftarrow E' \cup \{e\}$ `# Update the current budget of edges`
  else
    return $G' = (V, E')$
Until $Q = \emptyset$

Proposition D.2 (Resistance Diameter)**.** *Let G' =* `Sparsify`*(G, q) be a sampling algorithm of graph $G = (V, E)$, where edges $e \in E$ are sampled with probability $q \propto R_e$ (proportional to the effective resistance). Consider the* resistance diameter $\mathcal{R}_{diam} := \max_{u,v} R_{uv}$*. Then, for the pair of (u, v) does exist an edge $e = (u, v) \in E'$ in $G' = (V, E')$ such that $R_e = \mathcal{R}_{diam}$. A a result the Cheeger constant of G h_G is upper-bounded as follows:*

$$h_G \leq \frac{\alpha^\epsilon}{\sqrt{\mathcal{R}_{diam} \cdot \epsilon}} vol(S)^{\epsilon-1/2}, \tag{82}$$

with $0 < \epsilon < 1/2$ and $d_u \geq 1/\alpha$ for all $u \in V$.

Proof. The fact that the maximum resistance $\mathcal{R}_{diam}$ is located in an edge is derived from two observations: a) Resistance is upper bounded by the shortest-path distance; and b) edges with maximal resistance are prioritized in (Proposition D.1).

Theorem D.2 states that any attempt to increase the graph's bottleneck in a multiplicative way (*i.e.*, multiplying it by a constant $c \geq 0$) results in decreasing the effective resistances as follows:

$$R_{uv} \leq \left(\frac{1}{d_u^{2\epsilon}} + \frac{1}{d_v^{2\epsilon}}\right) \cdot \frac{1}{\epsilon \cdot c^2} \tag{83}$$

with $\epsilon \in [0, 1/2]$. This equation is called the *resistance bound*. Therefore, a multiplicative increase of the bottleneck leads to a quadratic decrease of the resistances.

Following Corollary 2 of [12], we obtain an upper bound of any h_S, *i.e.*, the Cheeger constant for $S \subseteq V$ with $vol(S) \leq vol(G)/2$ – by defining c properly. In particular we are seeking a value of c that would lead to a contradiction, which is obtained by setting

$$c = \sqrt{\frac{\left(\frac{1}{d_{u^*}^{2\epsilon}} + \frac{1}{d_{v^*}^{2\epsilon}}\right)}{\mathcal{R}_{diam} \cdot \epsilon}}\,, \tag{84}$$

where (u^*, v^*) is a pair of nodes with maximal resistance, *i.e.*, $R_{u^*v^*} = \mathcal{R}_{diam}$.

Consider now any other pair of nodes (s, t) with $R_{st} < \mathcal{R}_{diam}$. Following Theorem D.2, if the bottleneck of h_S is multiplied by c, we should have

$$R_{st} \leq \left(\frac{1}{d_s^{2\epsilon}} + \frac{1}{d_s^{2\epsilon}}\right) \cdot \frac{1}{\epsilon \cdot c^2} = \left(\frac{1}{d_s^{2\epsilon}} + \frac{1}{d_s^{2\epsilon}}\right) \cdot \frac{\mathcal{R}_{diam}}{\left(\frac{1}{d_{u^*}^{2\epsilon}} + \frac{1}{d_{v^*}^{2\epsilon}}\right)}\,. \tag{85}$$

However, since $\mathcal{R}_{diam} \leq \left(\frac{1}{d_{u^*}^{2\epsilon}} + \frac{1}{d_{v^*}^{2\epsilon}}\right)$ we have that R_{st} can satisfy

$$R_{st} > \left(\frac{1}{d_s^{2\epsilon}} + \frac{1}{d_s^{2\epsilon}}\right) \cdot \frac{1}{\epsilon \cdot c^2} \tag{86}$$

which is a contradiction and enables

$$h_S \leq \frac{c}{vol(S)^{1/2-\epsilon}} \iff |\partial S| \leq c \cdot vol(S)^{1/2-\epsilon}. \tag{87}$$

Using c as defined in Equation (84) and $d_u \geq 1/\alpha$ we obtain

$$c = \sqrt{\frac{\left(\frac{1}{d_{u^*}^{2\epsilon}} + \frac{1}{d_{v^*}^{2\epsilon}}\right)}{\mathcal{R}_{diam} \cdot \epsilon}} \leq \sqrt{\frac{\alpha^\epsilon}{\mathcal{R}_{diam} \cdot \epsilon}} \leq \frac{\alpha^\epsilon}{\sqrt{\mathcal{R}_{diam} \cdot \epsilon}}\,. \tag{88}$$

Therefore,

$$h_S \leq \frac{c}{vol(S)^{1/2-\epsilon}} \leq \frac{\frac{\alpha^\epsilon}{\sqrt{\mathcal{R}_{diam} \cdot \epsilon}}}{vol(S)^{1/2-\epsilon}} = \frac{\alpha^\epsilon}{\sqrt{\mathcal{R}_{diam} \cdot \epsilon}} \cdot vol(S)^{\epsilon-1/2}. \tag{89}$$

As a result, the Cheeger constant of $G = (V, E)$ is mildly reduced (by the square root of the maximal resistance). □

Proposition D.3 (Conclusion)**.** *Let (u^*, v^*) be a pair of nodes (may be not unique) in $G = (V, E)$ with maximal resistance,* i.e., *$R_{u^*v^*} = \mathcal{R}_{diam}$. Then, the Cheeger constant h_G relies on the ratio between the maximal resistance $\mathcal{R}_{diam}$ and its* uninformative approximation *$\left(\frac{1}{d_u^*} + \frac{1}{d_v^*}\right)$. The closer this ratio is to the unit, the easier it is to contain the Cheeger constant.*

Proof. The referred ratio above is the ratio leading to a proper c in Proposition D.2. This is consistent with a Lovász regime where the spectral gap λ_2' has a moderate value. However, for regimes with very small spectral gaps, *i.e.*, $\lambda_2' \to 0$, according to the Lovász bound, $\mathcal{R}_{diam} \gg \left(\frac{1}{d_u^*} + \frac{1}{d_v^*}\right)$ and hence the Cheeger constant provided by Proposition D.2 will tend to zero. □

We conclude that we can always find an moderate upper bound for the Cheeger constant of $G = (V, E)$, provided that the regime of the Lovász bound is also moderate. Therefore, as the global properties of $G = (V, E)$ are captured by $G' = (V, E')$, a moderate Cheeger constant, when achievable, also controls the bottlenecks in $G' = (V, E')$.

Our methodology has focused on first exploring the properties of the commute times / effective resistances in $G = (V, E)$. Next, we have leveraged the spectral similarity to reason about the properties –particularly the Cheeger constant– of $G = (V, E')$. In sum, we conclude that resistance diffusion via $\mathbf{T}^{CT}$ is a principled way of preserving the Cheeger constant of $G = (V, E)$.

Resistance-based Curvatures

We refer to recent work by Devriendt and Lambiotte [125] to complement the contributions of Topping et al. [383] regarding the use of curvature to rewire the edges in a graph.

Theorem D.3 (Devriendt and Lambiotte [125])**.** *The edge resistance curvature has the following properties: (1) It is bounded by* $(4 - d_u - d_v) \leq \kappa_{uv} \leq 2/R_{uv}$*, with equality in the lower bound iff all incident edges to* u *and* v *are cut links; (2) It is upper-bounded by the Ollivier-Ricci curvature* $\kappa^{OR}_{uv} \geq \kappa_{uv}$*, with equality if* (u, v) *is a cut link; and (3) Forman-Ricci curvature is bounded as follows:* $\kappa^{FR}_{uv}/R_{uv} \leq \kappa_{uv}$ *with equality in the bound if the edge is a cut link.*

The new definition of curvature given in [383] is related to the resistance distance and thus it is learnable with the proposed framework (CT-Layer). Actually, the Balanced-Forman curvature (Definition 1 in [383]) relies on the uniformative approximation of the resistance distance.

Figure 46 illustrates the relationship between effective resistances / commute times and curvature on an exemplary graph from the Collab dataset.

As seen in the Figure, effective resistances prioritize the edges connecting outer nodes with hubs or central nodes, while the intra-community connections are de-prioritized. This observation is consistent with the aforementioned theoretical explanations about preserving the bottleneck while breaking the intra-cluster structure. In addition, we also observe that the original edges between hubs have been deleted o have been extremely down-weighted.

Regarding curvature, hubs or central nodes have the lowest node curvature (this curvature increases with the number of nodes in a cluster/community). Edge curvatures, which rely on node curvatures, depend on the long-term neighborhoods of the connecting nodes. In general, edge curvatures can be seen as a smoothed version –since they integrate node curvatures– of the inverse of the resistance distances.

We observe that edges linking nodes of a given community with hubs tend to have similar edge-curvature values. However, edges linking nodes of different communities with hubs have different edge curvatures (Figure 46-right). This is due to the different number of nodes belonging to each community, and to their different average degree inside their respective communities (property 1 of Theorem D.3).

Finally, note that the range of edge curvatures is larger than that of resistance distances. The sparsifier transforms a uniform distribution of the edge weights into a less entropic one: in the example of Figure 46 we observe a power-law distribution of edge resistances. As a result, $\kappa_{uv} := 2(p_u + p_v)/\mathbf{T}^{CT}_{uv}$ becomes very large on average (edges with infinite curvature

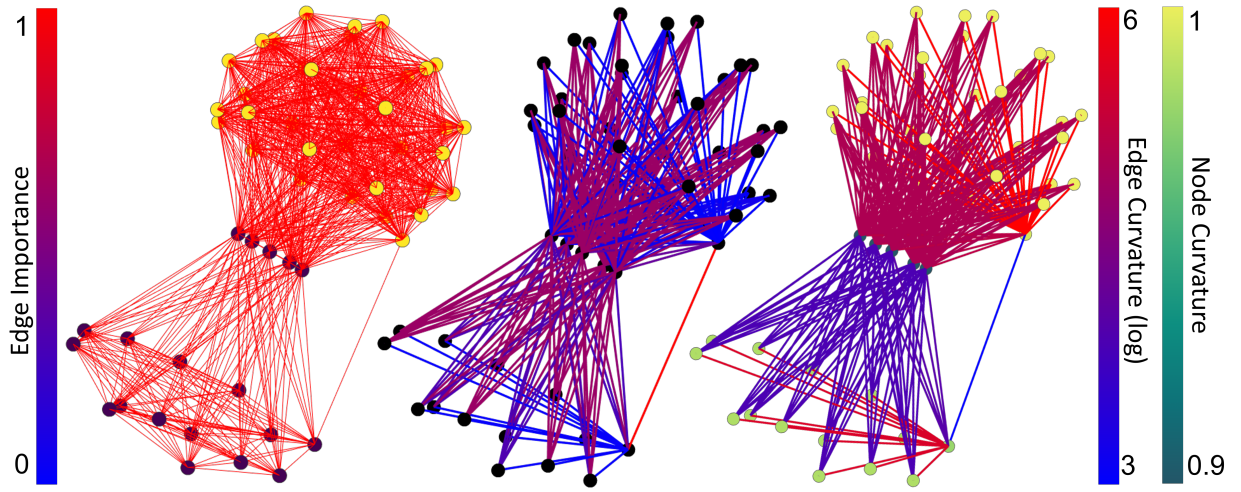


Figure 46. Left: Original graph with nodes colored as Louvain communities. Middle: $\mathbf{T}^{CT}$ learnt by CT-Layer with edges colors as node importance [0,1]. Right: Node and edge curvature: $\mathbf{T}^{CT}$ using $p_u := 1 - \frac{1}{2}\sum_{u\sim w} \mathbf{T}^{CT}_{uv}$ and $\kappa_{uv} := 2(p_u + p_v)/\mathbf{T}^{CT}_{uv}$ with edge an node curvatures as color. Graph from Reddit-B dataset.

are not shown in the plot) and a log scale is needed to appreciate the differences between edge resistances and edge curvatures.

D.6.2 In-detail Analysys of GAP-Layer

Spectral Gradients

The proposed GAP-Layer relies on gradients wrt the Laplacian eigenvalues, and particularly the spectral gap (λ_2 for $\mathbf{L}$ and λ_2' wrt $\mathcal{L}$). Although the GAP-Layer inductively rewires the adjacency matrix $\mathbf{A}$ so that λ_2 is minimized, the gradients derived in this section may also be applied for gap maximization.

Note that while our cost function $L_{Fiedler} = \|\tilde{\mathbf{A}} - \mathbf{A}\|_F + \alpha(\lambda_2^*)^2$, with $\lambda_2^* \in \{\lambda_2, \lambda_2'\}$, relies on an eigenvalue, we *do not compute it explicitly*, as its computation has a complexity of $O(n^3)$ and would need to be computed in every learning iteration. Instead, we learn an approximation of λ_2's eigenvector $\mathbf{f}_2$ and use its Dirchlet energy $\mathcal{E}(\mathbf{f}_2)$ to approximate the eigenvalue. In addition, since $\mathbf{g}_2 = \mathbf{D}^{1/2}\mathbf{f}_2$, we first approximate $\mathbf{g}_2$ and then approximate λ_2' from $\mathcal{E}(\mathbf{g}_2)$.

Gradients of the Ratio-cut Approximation. Let $\mathbf{A}$ be the adjacency matrix of $G = (V, E)$; and $\tilde{\mathbf{A}}$, a matrix similar to the original adjacency but with minimal λ_2. Then, the gradient of λ_2 wrt each component of $\tilde{\mathbf{A}}$ is given by

$$\nabla_{\tilde{\mathbf{A}}}\lambda_2 := Tr\left[(\nabla_{\tilde{\mathbf{L}}}\lambda_2)^T \cdot \nabla_{\tilde{\mathbf{A}}}\tilde{\mathbf{L}}\right] = \mathrm{diag}(\mathbf{f}_2\mathbf{f}_2^T)\mathbf{1}\mathbf{1}^T - \mathbf{f}_2\mathbf{f}_2^T\ , \tag{90}$$

where $\mathbf{1}$ is the vector of n ones; and $[\nabla_{\tilde{\mathbf{A}}}\lambda_2]_{ij}$ is the gradient of λ_2 wrt $\tilde{\mathbf{A}}_{uv}$. The above formula is an instance of the network derivative mining mining approach [233]. In this framework,

λ_2 is seen as a function of $\tilde{\mathbf{A}}$ and $\nabla_{\tilde{\mathbf{A}}}\lambda_2$, the gradient of λ_2 wrt $\tilde{\mathbf{A}}$, comes from the chain rule of the matrix derivative $Tr\left[(\nabla_{\tilde{\mathbf{L}}}\lambda_2)^T \cdot \nabla_{\tilde{\mathbf{A}}}\tilde{\mathbf{L}}\right]$. More precisely,

$$\nabla_{\tilde{\mathbf{L}}}\lambda_2 := \frac{\partial \lambda_2}{\partial \tilde{\mathbf{L}}} = \mathbf{f}_2\mathbf{f}_2^T \ , \tag{91}$$

is a matrix relying on an outer product (correlation). In the proposed GAP-Layer, since $\mathbf{f}_2$ is approximated by:

$$\mathbf{f}_2(u) = \begin{cases} +1/\sqrt{n} & \text{if } u \text{ belongs to the first cluster} \\ -1/\sqrt{n} & \text{if } u \text{ belongs to the second cluster} \end{cases} , \tag{92}$$

i.e., we discard the $O\left(\frac{\log n}{n}\right)$ from Appendix D.6.2 (the non-liniarities conjectured in [210]) in order to simplify the analysis. After reordering the entries of $\mathbf{f}_2$ for the sake of clarity, $\mathbf{f}_2\mathbf{f}_2^T$ is the following block matrix:

$$\mathbf{f}_2\mathbf{f}_2^T = \left[\begin{array}{c|c} 1/n & -1/n \\ \hline -1/n & 1/n \end{array}\right] \text{ whose diagonal matrix is } \operatorname{diag}(\mathbf{f}_2\mathbf{f}_2^T) = \left[\begin{array}{c|c} 1/n & 0 \\ \hline 0 & 1/n \end{array}\right] \tag{93}$$

Then, we have

$$\nabla_{\tilde{\mathbf{A}}}\lambda_2 = \left[\begin{array}{c|c} 1/n & 1/n \\ \hline 1/n & 1/n \end{array}\right] - \left[\begin{array}{c|c} 1/n & -1/n \\ \hline -1/n & 1/n \end{array}\right] = \left[\begin{array}{c|c} 0 & 2/n \\ \hline 2/n & 0 \end{array}\right] \tag{94}$$

which explains the results in Figure 42-left: edges linking nodes belonging to the same cluster remain unchanged whereas inter-cluster edges have a gradient of $2/n$. This provides a simple explanation for $\mathbf{T}^{GAP} = \tilde{\mathbf{A}}(\mathbf{S}) \odot \mathbf{A}$. The additional masking added by the adjacency matrix ensures that we do not create new links.

Gradients Normalized-cut Approximation. Similarly, using λ_2' for graph rewiring leads to the following complex expression:

$$\begin{aligned} \nabla_{\tilde{\mathbf{A}}}\lambda_2' := Tr\left[(\nabla_{\tilde{\mathcal{L}}}\lambda_2)^T \cdot \nabla_{\tilde{\mathbf{A}}}\tilde{\mathcal{L}}\right] \ = \\ \mathbf{d}'\left\{\mathbf{g}_2^T\tilde{\mathbf{A}}^T\tilde{\mathbf{D}}^{-1/2}\mathbf{g}_2\right\}\mathbf{1}^T + \mathbf{d}'\left\{\mathbf{g}_2^T\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{g}_2\right\}\mathbf{1}^T \ + \ \tilde{\mathbf{D}}^{-1/2}\mathbf{g}_2\mathbf{g}_2^T\tilde{\mathbf{D}}^{-1/2} \ . \end{aligned} \tag{95}$$

However, since $\mathbf{g}_2 = \mathbf{D}^{1/2}\mathbf{f}_2$ and $\mathbf{f}_2 = \mathbf{D}^{-1/2}\mathbf{g}_2$, the gradient may be simplified as follows:

$$\begin{aligned} \nabla_{\tilde{\mathbf{A}}}\lambda_2' := Tr\left[(\nabla_{\tilde{\mathcal{L}}}\lambda_2)^T \cdot \nabla_{\tilde{\mathbf{A}}}\tilde{\mathcal{L}}\right] \ = \\ \mathbf{d}'\left\{\mathbf{f}_2^T\tilde{\mathbf{D}}^{1/2}\tilde{\mathbf{A}}^T\mathbf{f}_2\right\}\mathbf{1}^T + \mathbf{d}'\left\{\mathbf{f}_2^T\tilde{\mathbf{D}}^{1/2}\tilde{\mathbf{A}}\mathbf{f}_2\right\}\mathbf{1}^T \ + \ \tilde{\mathbf{D}}^{-1/2}\mathbf{f}_2\mathbf{f}_2^T\tilde{\mathbf{D}}^{-1/2} \ . \end{aligned} \tag{96}$$

In addition, considering symmetry for the undirected graph case, we obtain:

$$\begin{aligned} \nabla_{\tilde{\mathbf{A}}}\lambda_2' := Tr\left[(\nabla_{\tilde{\mathcal{L}}}\lambda_2)^T \cdot \nabla_{\tilde{\mathbf{A}}}\tilde{\mathcal{L}}\right] = \\ 2\mathbf{d}'\left\{\mathbf{f}_2^T\tilde{\mathbf{D}}^{1/2}\tilde{\mathbf{A}}\mathbf{f}_2\right\}\mathbf{1}^T + \tilde{\mathbf{D}}^{-1/2}\mathbf{f}_2\mathbf{f}_2^T\tilde{\mathbf{D}}^{-1/2} \ . \end{aligned} \tag{97}$$

where $\mathbf{d}'$ is a $n \times 1$ negative vector including derivatives of degree wrt adjacency and related terms. The obtained gradient is composed of two terms.

The first term contains the matrix $\tilde{\mathbf{D}}^{1/2}\tilde{\mathbf{A}}$ which is the adjacency matrix weighted by the square root of the degree; $\mathbf{f}_2^T\tilde{\mathbf{D}}^{1/2}\tilde{\mathbf{A}}\mathbf{f}_2$ is a quadratic form (similar to a Dirichlet energy for the Laplacian) which approximates an eigenvalue of $\tilde{\mathbf{D}}^{1/2}\tilde{\mathbf{A}}$. We plan to further analyze the properties of this term in future work.

The second term, $\tilde{\mathbf{D}}^{-1/2}\mathbf{f}_2\mathbf{f}_2^T\tilde{\mathbf{D}}^{-1/2}$, downweights the correlation term for the Ratio-cut case $\mathbf{f}_2\mathbf{f}_2^T$ by the degrees as in the normalized Laplacian. This results in a normalization of the Fiedler vector: $-1/n$ becomes $-\sqrt{d_u d_v}/n$ at the uv entry and similarly for $1/n$, *i.e.*, each entry contains the average degree assortativity.

Beyond the Lovász Bound: the von Luxburg et al. bound

The Lovász bound was later refined by Luxburg, Radl, and Hein [279] via a new, tighter bound which replaces d_{min} by d_{min}^2 in Equation (63). Given that $\lambda_2' \in (0, 2]$, as the number of nodes in the graph ($n = |V|$) and the average degree increase, then $R_{uv} \approx 1/d_u + 1/d_v$. This is likely to happen in certain types of graphs, such as Gaussian similarity-graphs –graphs where two nodes are linked if the neg-exponential of the distances between the respective features of the nodes is large enough; ϵ-graphs –graphs where the Euclidean distances between the features in the nodes are $\leq \epsilon$; and $k-$NN graphs with large k wrt n. The authors report a linear collapse of R_{uv} with the density of the graph in scale-free networks, such as social network graphs, whereas a faster collapse of R_{uv} has been reported in community graphs –congruent graphs with Stochastic Block Models (SBMs) [1].

Given the importance of the effective resistance, R_{uv}, as a *global* measure of node similarity, the von Luxburg et al.'s refinement motivated the development of *robust effective resistances*, mostly in the form of $p-$resistances given by $R_{uv}^p = \arg\min_{\mathbf{f}}\{\sum_{e\in E} r_e|f_e|^p\}$, where $\mathbf{f}$ is a unit-flow injected in u and recovered in v; and $r_e = 1/w_e$ with w_e being the edge's weight [10]. For $p = 1$, R_{uv}^p corresponds to the shortest path; $p = 2$ results in the effective resistance; and $p \to \infty$ leads to the inverse of the unweighted u-v-mincut[39]. Note that the optimal p value depends on the type of graph [10] and $p-$resistances may be studied from the perspective of $p-$Laplacians [10, 83].

While R_{uv} could be unbounded by minimizing the spectral gap λ_2', this approach has received little attention in the literature of mathematical characterization of graphs with small spectral gaps [54][360], *i.e.*, instead of tackling the daunting problem of explicitly minimizing the gap, researchers in this field have preferred to find graphs with small spectral gaps.

D.6.3 Additional Experiments

In this section, we provide details about the graphs contained in each of the datasets used in our experiments, a detailed clarification about architectures and experiments, and, finally, report additional experimental results.

Datasets Statistics

Table 20 depicts the number of nodes, edges, average degree, assortativity, number of triangles, transitivity and clustering coefficients (mean and standard deviation) of all the graphs

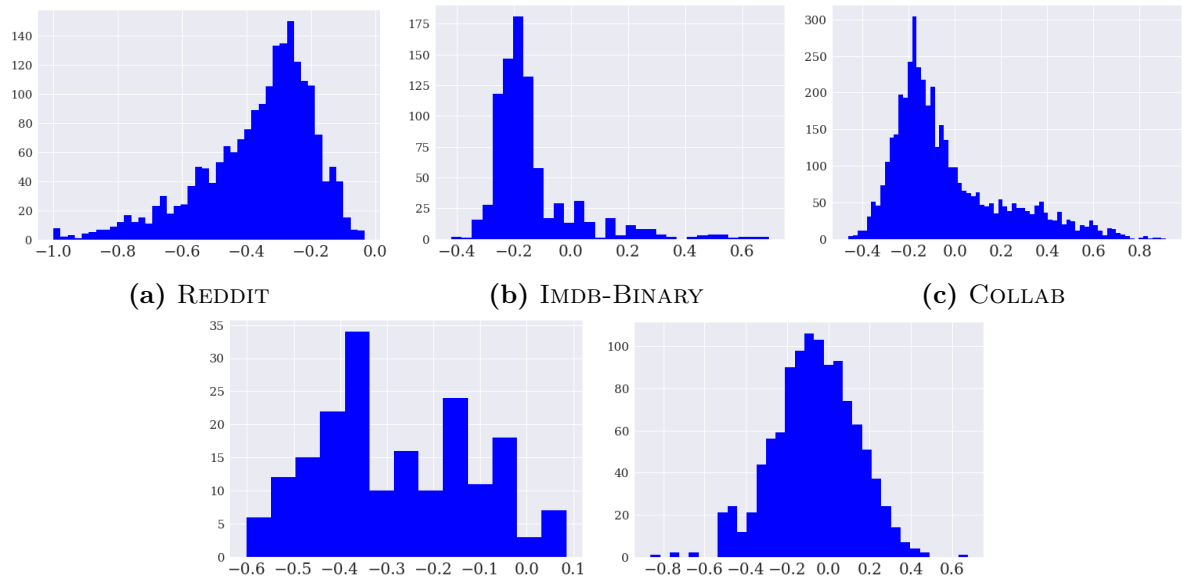
[39] The link between CTs and mincuts is leveraged in Appendix D as an essential element of our approach.

contained in each of the benchmark datasets used in our experiments. As seen in the Table, the datasets are very diverse in their characteristics. In addition, we use two synthetic datasets with 2 classes: Erdös-Rényi with $p_1 \in [0.3, 0.5]$ and $p_2 \in [0.4, 0.8]$ and Stochastic block model (SBM) with parameters $p_1 = 0.8$, $p_2 = 0.5$, $q_1 \in [0.1, 0.15]$ and $q_2 \in [0.01, 0.1]$.

Table 20. Dataset statistics. Parenthesis in *Assortativity* column denotes number of complete graphs (assortativity is undefined).

	Nodes	Egdes	AVG Degree	Triangles	Transitivity	Clustering	Assortativity
REDDIT-B	429.6 ± 554	497.7 ± 622	2.33 ± 0.3	24 ± 41	0.01 ± 0.02	0.04 ± 0.06	-0.364 ± 0.17 (0)
IMDB-B	19.7 ± 10	96.5 ± 105	8.88 ± 5.0	391 ± 868	0.77 ± 0.15	0.94 ± 0.03	-0.135 ± 0.16 (139)
COLLAB	74.5 ± 62	2457 ± 6438	37.36 ± 44	12×10^4 $\pm 48\times10^4$	0.76 ± 0.21	0.89 ± 0.08	-0.033 ± 0.24 (680)
MUTAG	2.2 ± 0.1	19.8 ± 5.6	2.18 ± 0.1	0.00 ± 0.0	0.00 ± 0.00	0.00 ± 0.00	-0.279 ± 0.17 (0)
PROTEINS	39.1 ± 45.8	72.8 ± 84.6	3.73 ± 0.4	27.4 ± 30	0.48 ± 0.20	0.51 ± 0.23	-0.065 ± 0.2 (13)

In addition, Figure 47 depicts the histograms of the assortativity for all the graphs in each of the eight datasets used in our experiments. As shown in Table 20 assortativity is undefined in complete graphs (constant degree, all degrees are the same). Assortativity is defined as the normalized degree correlation. If the graph is complete, then both correlation and its variance is 0, so assortativity will be 0/0.



(a) Reddit **(b)** Imdb-Binary **(c)** Collab

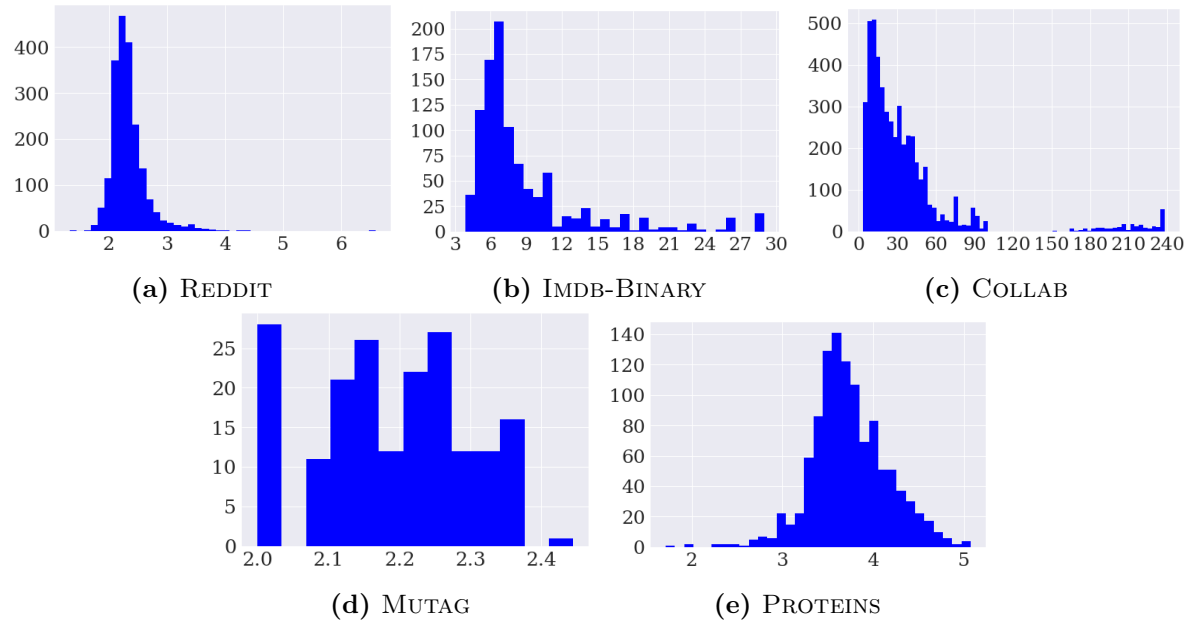
(d) Mutag **(e)** Proteins

Figure 47. Histogram of the Assortativity of all the graphs in each of the datasets.

In addition, Figure 48 depicts the histograms of the average node degrees for all the graphs in each of the eight datasets used in our experiments. The datasets are also very diverse in terms of topology, corresponding to social networks, biochemical networks and meshes.

Graph Classification GNN Architectures

Figure 49 shows the specific GNN architectures used in the experiments explained in Appendix D.4. Although the specific calculation of $\mathbf{T}^{GAP}$ and $\mathbf{T}^{CT}$ are given in Definitions D.1



(a) REDDIT **(b)** IMDB-BINARY **(c)** COLLAB

(d) MUTAG **(e)** PROTEINS

Figure 48. Degree histogram of the average degree of all the graphs in each of the datasets.

and D.2, we also provide a couple of pictures for a better intuition.

Training Parameters

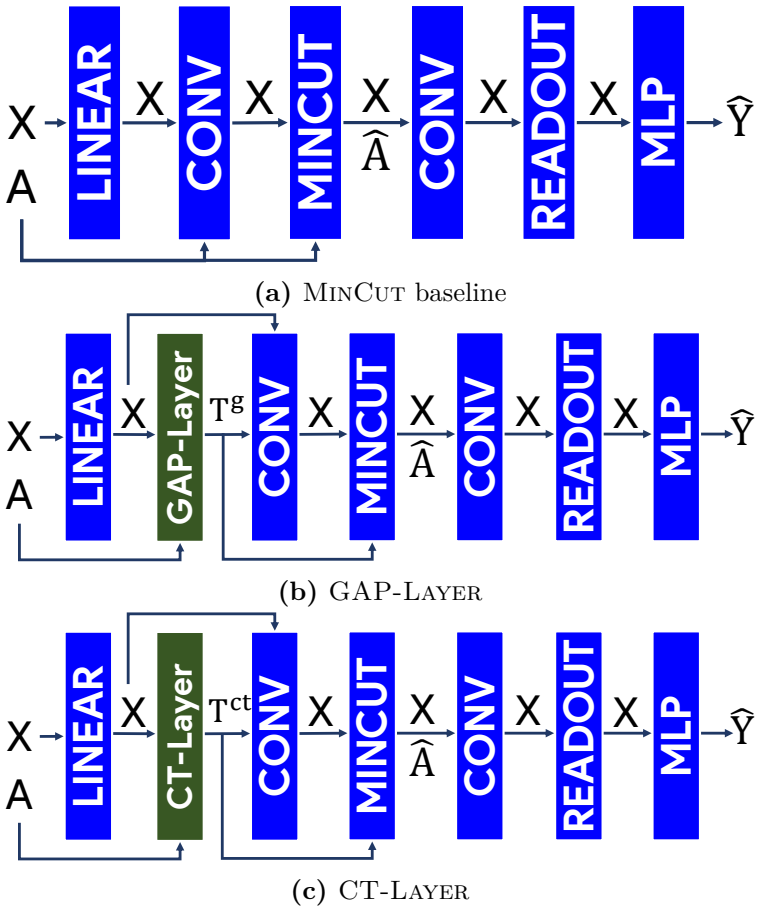
The value of the hyperparameters used in the experiments are the ones by default in the code repository.[40] We report average accuracies and standard deviation on 10 random iterations, using different 85/15 train-test stratified split (we do not perform hyperparameter search), training during 60 epochs and reporting the results of the last epoch for each random run. We have used an Adam optimizer, with a learning rate of $5e-4$ and weight decay of $1e-4$. In addition, the batch size used for the experiments are shown in Table 21. Regarding the synthetic datasets, the parameters are: Erdös-Rényi with $p_1 \in [0.3, 0.5]$ and $p_2 \in [0.4, 0.8]$ and Stochastic block model (SBM) $p_1 = 0.8$, $p_2 = 0.5$, $q_1 \in [0.1, 0.15]$ and $q_2 \in [0.01, 0.1]$.

Table 21. Dataset Batch size

	Batch	Dataset size
REDDIT-BINARY	64	1000
IMDB-BINARY	64	2000
COLLAB	64	5000
MUTAG	32	188
PROTEINS	64	1113
SBM	32	1000
Erdös-Rényi	32	1000

For the k-nn graph baseline, we choose k such that the main degree of the original graph is maintained, *i.e.*, k equal to average degree. Our experiments also use 2 preprocessing

[40] https://github.com/AdrianArnaiz/DiffWire



(a) MinCut baseline

(b) GAP-Layer

(c) CT-Layer

Figure 49. Diagrams of the GNNs used in the experiments.

methods DIGL and SDRF. Unlike our proposed methods, both SDRF [383] and DIGL [247] use a set of hyperparamerters to optimize for each specific graph, because both are also not inductive. This approach could be manageable for the task of node classification, where you only have one graph. However, when it comes to graph classification, the number of graphs are huge (Table 21) and it is nor computationally feasible optimize parameters for each specific graph. For DIGL, we use a fixed $\alpha = 0.001$ and ϵ based on keeping the same average degree for each graph, *i.e.*, we use a different dynamically chosen ϵ for each graph in each dataset which maintain the same number of edges as the original graph. In the case of SDRF, the parameters define how stochastic the edge addition is (τ), the graph edit distance upper bound (number of iterations) and optional Ricci upper-bound above which an edge will be removed each iteration (C^+). We set the parameters $\tau = 20$ (the edge added is always near the edge of lower curvature), $C^+ = 0$ (to force one edge is removed every iteration), and number of iterations dynamic according to $0.7 * |V|$. Thus, we maintain the same number of edges in the new graph ($\tau = 20$ and $C^+ = 0$), *i.e.*, same average degree, and we keep the graph distance to the original bounded by $0.7 * |V|$.

Latent Space Analysis

In this section, we analyze the two latent spaces produced by the models.

- First, we compare the CT Embedding computed spectrally ($\mathbf{Z}$ in Equation (64)) with the CT Embedding predicted by our CT-Layer ($\mathbf{Z}$ in Definition D.1) for a given graph, where each point is a node in the graph.

- Second, we compare the graph readout output for every model defined in the experiments (Figure 45) where each point is a graph in the dataset.

Spectral CT Embedding vs CT Embeddings Learned by CT-Layer The well-known embeddings based on the Laplacian positional encodings (PE) are typically computed beforehand and appended to the input vector $\mathbf{X}$ as additional features [135, 392]. This task requires an expensive computation $O(n^3)$ (see Equation (64)). Conversely, we propose a GNN Layer that learns how to predict the CT embeddings (CTEs) for unseen graphs (Definition D.1 and Figure 43) with a loss function that optimizes such CTEs. Note that we do not explicitly use the CTE features (PE) for the nodes, but we use the CTs as a new diffusion matrix for message passing (given by $\mathbf{T}^{\mathbf{CT}}$ in Definition D.1). Note that we could also use $\mathbf{Z}$ as positional encodings in the node features, such that CT-Layer may be seen as a novel approach to learn Positonal Encodings.

In this section, we perform a comparative analysis between the spectral commute times embeddings (spectral CTEs, $\mathbf{Z}$ in Equation (64)) and the CTEs that are predicted by our CT-Layer ($\mathbf{Z}$ in Definition D.1). As seen in Figure 50 (top), both embeddings respect the original topology of the graph, but they differ due to (1) orthogonality restrictions, and more interestingly to (2) the simplification of the original spectral loss function in Alev et al. [12]: the spectral CTEs minimize the trace of a quotient, which involves computing an inverse, whereas the CTEs learned in CT-Layer minimize the quotient of two traces which is computationally simpler (see L_{CT} loss in Definition 1). Two important properties of the first term in Definition D.1 are: (1) the learned embedding $\mathbf{Z}$ has minimal Dirichlet energy (numerator) and (2) large degree nodes will be separated (denominator). Figure 50 (top) illustrates how the CTEs that are learned in CT-Layer are able to better preserve the original topology of the graph (note how the nodes are more compactly embedded when compared to the spectral CTEs).

Figure 50 (bottom) depicts a histogram of the effective resistances or commute times (CTs) (see Appendix D.3.2) of the edges according to CT-Layer or the spectral CTEs. The histogram is computed from the upper triangle of the $\mathbf{T}^{\mathbf{CT}}$ matrix defined in Definition D.1. Note that the larger the effective resistance of an edge, the more important that edge will be considered (and hence the lower the probability of being removed [246]). We observe how in the histogram of CTEs that are learned in CT-Layer there is a "small club" of edges with very large values and a large number of edges with low values yielding a power-law-like profile. However, the histogram of the effective resistances computed by the spectral CTEs exhibits a profile similar to a Gaussian distribution. From this result, we conclude that the use of L_{CT} in the learning process of the CT-Layer shifts the distribution of the effective resistances of the edges towards an asymmetric distribution where few edges have very large weights and a majority of edges have low weights.

Graph Readout Latent Space Analysis To delve into the analysis of the latent spaces produced by our layers and model, we also inspect the latent space produced by the models (Figure 45) that use MinCutPool (Figure 49a), GAP-Layer (Figure 49b) and CT-

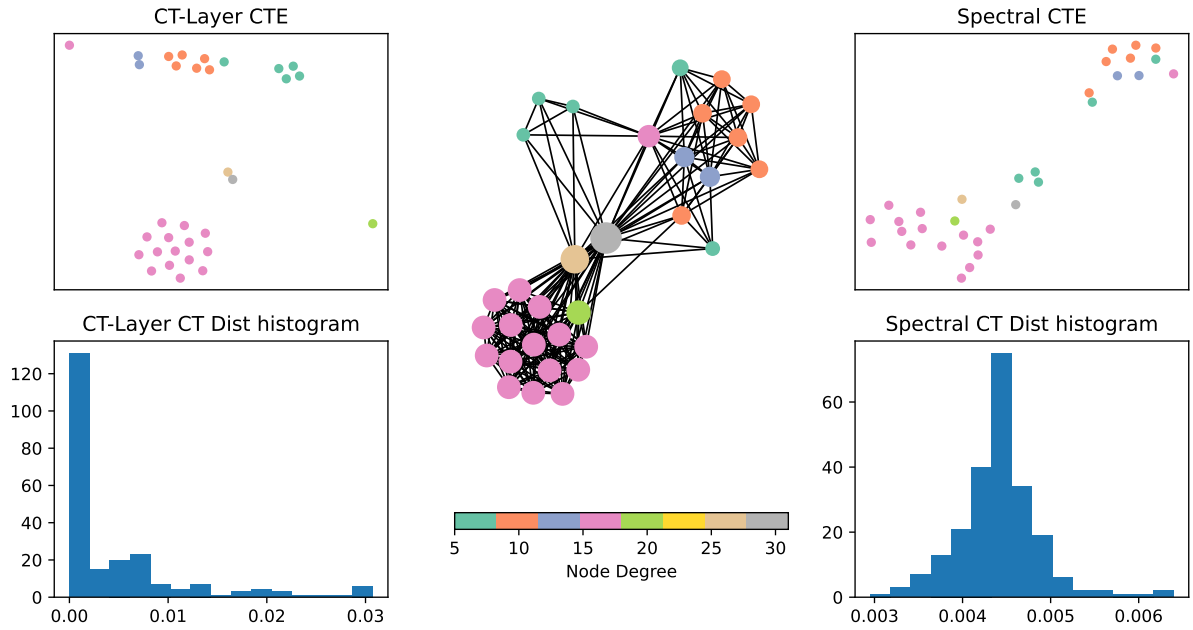


Figure 50. Top: CT embeddings predicted by CT-Layer (left) and spectral CT embeddinggs (right). Bottom: Histogram of normalized effective resistances (*i.e.*, CT distances or upper triangle in $\mathbf{T}^{\mathbf{CT}}$) computed from the above CT embeddings. Middle: original graph from the COLLAB dataset. Colors correspond to node degree. CT-Layer CTEs reduced from 75 to 32 dimensions using Johnson-Lindenstrauss. Finally, both CTEs reduced from 32 to 2 dimensions using T-SNE.

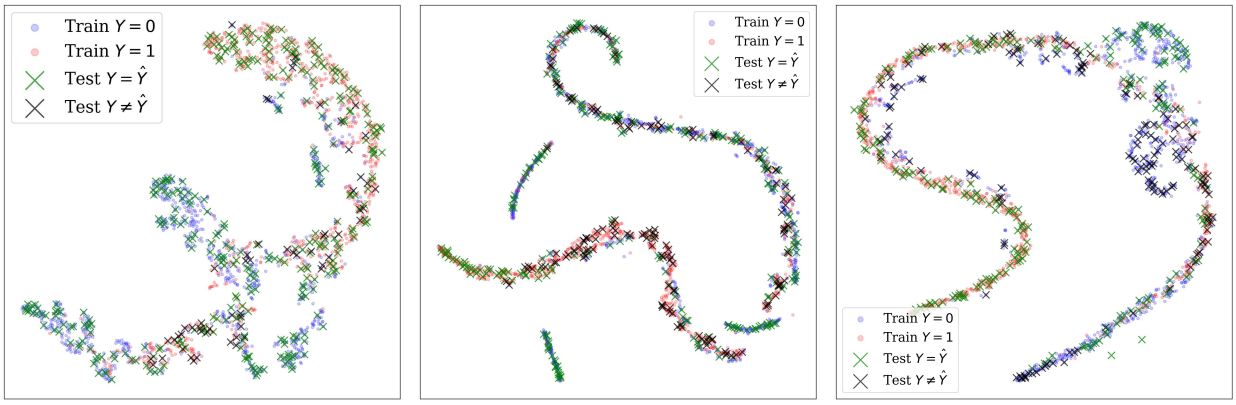
Layer (Figure 49c). Each point is a graph in the dataset, corresponding to the graph embedding of the readout layer. We plot the output of the readout layer for each model, and then perform dimensionality reduction with TSNE.

Observing the latent space of the Reddit-Binary dataset (Figure 51), CT-Layer creates a disperse yet structured latent space for the embeddings of the graphs. This topology in latent spaces show that this method is able to capture different topological details. The main reason is the expressiveness of the commute times as a distance metric when performing rewiring, which has been shown to be a optimal metric to measure node structural similarity. In addition, GAP-Layer creates a latent space where, although the 2 classes are also separable, the embeddings are more compressed, due to a more aggressive –yet still informative– change in topology. This change in topology is due to the change in bottleneck size that GAP-Layer applies to the graph. Finally, MinCut creates a more squeezed and compressed embedding, where both classes lie in the same spaces and most of the graphs have collapsed representations, due to the limited expressiveness of this architecture.

Architectures and Details of Node Classification Experiments

The application of our framework for a node classification task entails several considerations. First, this first implementation of our method works with dense $\mathbf{A}$ and $\mathbf{X}$ matrices, whereas node classification typically uses sparse representations of the edges. Thus, the implementation of our proposed layers is not straightforward for sparse graph representations. We are planning to work on the sparse version of this method in future work.

Note that we have chosen benchmark datasets that are manageable with our dense im-



(a) CT-Layer **(b)** MinCut **(c)** GAP-Layer

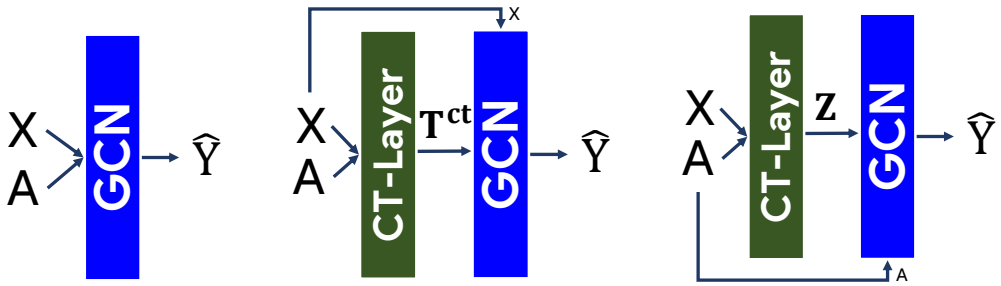
Figure 51. Reddit embeddings produced by GAP-Layer (Ncut) CT-Layer and MinCut.

plementation. In addition, we have chosen a basic baseline with 1 GCN layer to show the ability of the approaches to avoid under-reaching, over-smoothing and over-squashing.

The baseline GCN is a 1-layer-GCN, and the 2 compared models are:

- 1 CT-Layer for calculating $\mathbf{Z}$ followed by 1 GCN Layer using $\mathbf{A}$ for message passing and $\mathbf{X} \parallel \mathbf{Z}$ as features. This approach is a combination of Velingker et al. [392] and our method. See Figure 52c.

- 1 CT-Layer for calculating $\mathbf{T}^{\mathbf{CT}}$ followed by 1 GCN Layer using that $\mathbf{T}^{\mathbf{CT}}$ for message passing and $\mathbf{X}$ as features. See Figure 52b.



(a) GCN baseline **(b)** $\mathbf{A} = \mathbf{T}^{\mathbf{CT}}$ **(c)** $\mathbf{X} \parallel \mathbf{Z}$

Figure 52. Diagrams of the GNNs used in the experiments for node classification.

A promising direction of future work would be to explore how to combine both approaches to leverage the best of each of the methods on a wide range of graphs for node classification tasks. In addition, using this learnable CT distance for modulating message passing in more sophisticated ways is planned for future work.

Analysis of Correlation between Structural Properties and CT-Layer Performance

To analyze the performance of our model in graphs with different structural properties, we analyze the correlation between accuracy, the graph's assortativity, and the graph's bottleneck (λ_2) in COLLAB and REDDIT datasets. If the error is consistent along all levels of accuracy and gaps, the layer can generalize along different graph topologies.

As seen in Figure 55, Figure 53 (middle), and Figure 54 (middle), we do not identify any correlation or systematic pattern between graph classification accuracy, assortativity, and bottleneck with CT-Layer-based rewiring, since the proportion of wrong and correct predictions are regular for all levels of assortativity and bottleneck size.

In addition, note that while there is a systematic error of the model over-predicting class 0 in the COLLAB dataset (see Figure 53), this behavior is not explained by assortativity or bottleneck size, but by the unbalanced number of graphs in each class.

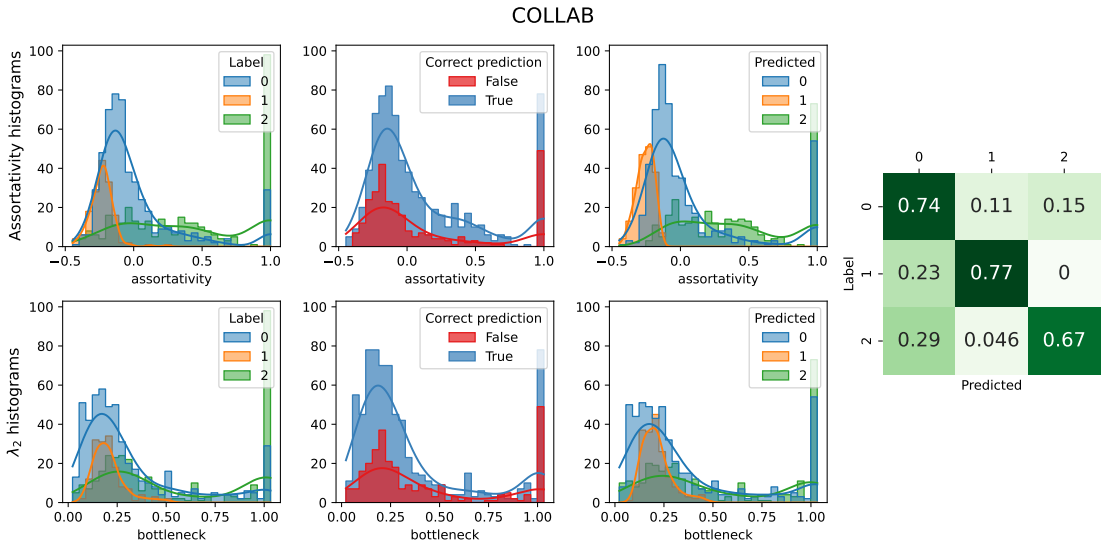


Figure 53. Analysis of assortativity, bottleneck and accuracy for COLLAB dataset. Top: Histograms of assortativity. Bottom: Histograms of bottleneck size (λ_2). Both are grouped by actual label of the graph (left), by correct or wrong predictions (middle) and by predicted label (right).

Computing Infrastructure

Table 22 summarizes the computing infrastructure used in our experiments.

D.7 Notation

The Table 23 summarizes the notation used in Appendix D.

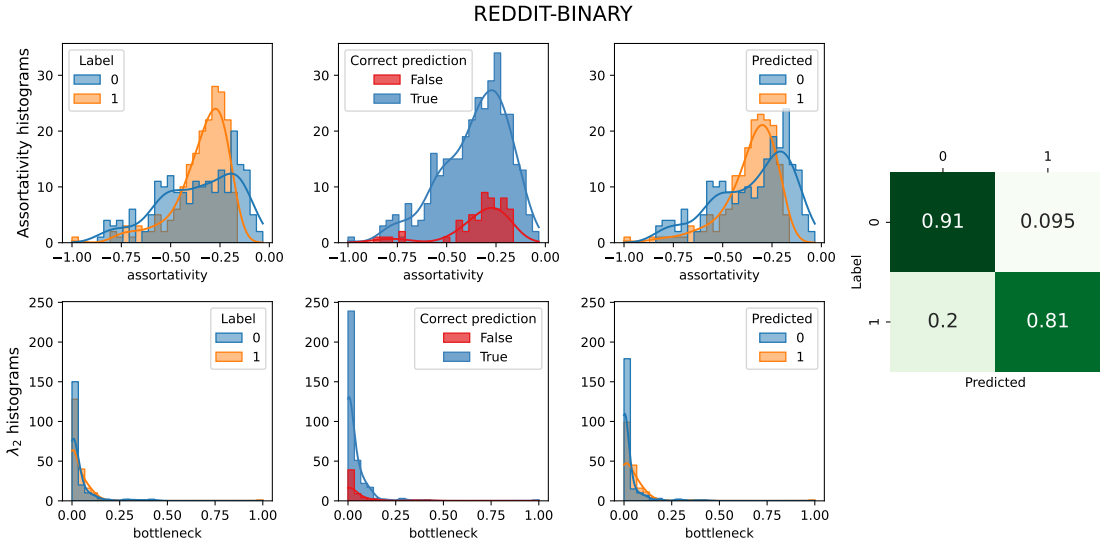
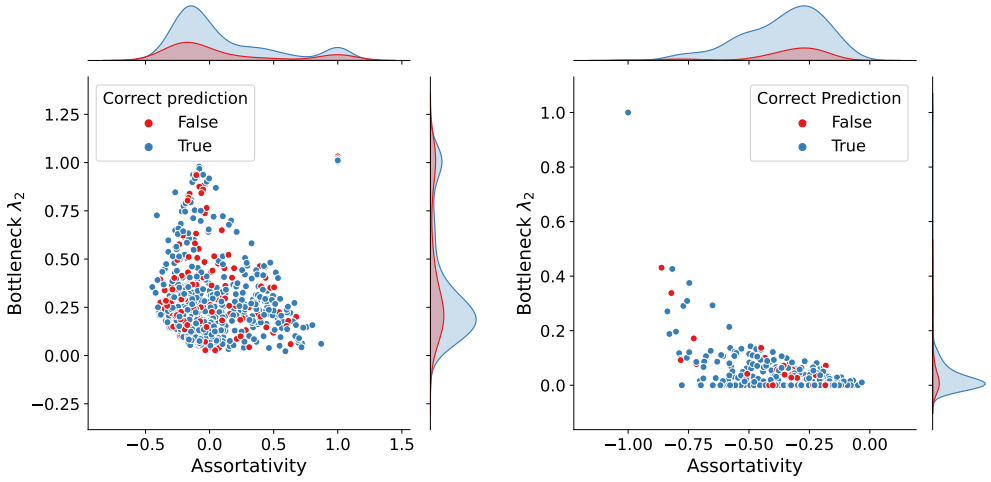


Figure 54. Analysis of assortativity, bottleneck and accuracy for REDDIT-B dataset. Top: Histograms of assortativity. Bottom: Histograms of bottleneck size (λ_2). Both are grouped by actual label of the graph (left), by correct or wrong predictions (middle) and by predicted label (right).

Table 22. Computing infrastructure.

Component	Details
GPU	2x A100-SXM4-40GB
RAM	1 TiB
CPU	255x AMD 7742 64-Core @ 2.25 GHz
OS	Ubuntu 20.04.4 LTS



(a) COLLAB **(b)** REDDIT-B

Figure 55. Correlation between assortativity, λ_2 and accuracy for CT-Layer. Histograms shows that the proportion of correct and wrong predictions are regular for all levels of assortativity (x axis) and bottleneck size (y axis). For the sake of clarity, these visualizations, a and b, are the combination of the 2 histograms in the middle column of Figure 53 and Figure 54 respectively.

Symbol	Description
$G = (V, E)$	Graph = (Nodes, Edges)
$\mathbf{A}$	Adjacency matrix: $\mathbf{A} \in \mathbb{R}^{n \times n}$
$\mathbf{X}$	Feature matrix: $\mathbf{X} \in \mathbb{R}^{n \times F}$
v	Node $v \in V$ or $u \in V$
e	Edge $e \in E$
x	Features of node v: $x \in X$
n	Number of nodes: $n = \|V\|$
F	Number of features
$\mathbf{D}$	Degree diagonal matrix where d_v in D_{vv}
d_v	Degree of node v
$vol(G)$	Sum of the degrees of the graph $vol(G) = Tr[D]$
$\mathbf{L}$	Laplacian: $\mathbf{L} = \mathbf{D} - \mathbf{A}$
$\mathbf{B}$	Signed edge-vertex incidence matrix
$\mathbf{b}_e$	Incidence vector: Row vector of $\mathbf{B}$, with $\mathbf{b}_{e=(u,v)} = (\mathbf{e}_u - \mathbf{e}_v)$
$\mathbf{v}_e$	Projected incidence vector: $\mathbf{v}_e = \mathbf{L}^{+/2}\mathbf{b}_e$
Γ	Ratio $\Gamma = \frac{1+\epsilon}{1-\epsilon}$
$\mathcal{E}$	Dirichlet Energy wrt $\mathbf{L}$: $\mathcal{E}(\mathbf{x}) := \mathbf{x}^T\mathbf{L}\mathbf{x}$
$\mathcal{L}$	Normalized Laplacian: $\mathcal{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$
Λ	Eigenvalue matrix of $\mathbf{L}$
Λ'	Eigenvalue matrix of $\mathcal{L}$
λ_i	i-th eigenvalue of $\mathbf{L}$
λ_2	Second eigenvalue of $\mathbf{L}$: Spectral gap
λ'_i	i-th eigenvalue of $\mathcal{L}$
λ'_2	Second eigenvalue of $\mathcal{L}$: Spectral gap
$\mathbf{F}$	Matrix of eigenvectors of $\mathbf{L}$
$\mathbf{G}$	Matrix of eigenvectors of $\mathcal{L}$
$\mathbf{f}_i$	i eigenvector of $\mathbf{L}$
$\mathbf{f}_2$	Second eigenvector of $\mathbf{L}$: Fiedler vector
$\mathbf{g}_i$	i eigenvector of $\mathcal{L}$
$\mathbf{g}_2$	Second eigenvector of $\mathcal{L}$: Fiedler vector
$\tilde{\mathbf{A}}$	New Adjacency matrix
E'	New edges
H_{uv}	Hitting time between u and v
CT_{uv}	Commute time: $CT_{uv} = H_{uv} + H_{vu}$
R_{uv}	Effective resistance: $R_{uv} = CT_{uv}/vol(G)$
$\mathbf{Z}$	Matrix of commute times embeddings for all nodes in G
$\mathbf{z}_u$	Commute times embedding of node u
$\mathbf{T}^{CT}$	Resistance diffusion or commute times diffusion
$\mathbf{R}(\mathbf{Z})$	Pairwise Euclidean distance of embedding $\mathbf{Z}$ divided by $vol(G)$
$\mathbf{S}$	Cluster assignment matrix: $\mathbf{S} \in \mathbb{R}^{n \times 2}$
$\mathbf{T}^{GAP}$	GAP diffusion
$\mathbf{e}_u$	Unit vector with unit value at u and 0 elsewhere
$\nabla_{\tilde{\mathbf{A}}}\lambda_2$	Gradient of λ_2 wrt $\tilde{\mathbf{A}}$
$[\nabla_{\tilde{\mathbf{A}}}\lambda_2]_{ij}$	Gradient of λ_2 wrt $\tilde{\mathbf{A}}_{uv}$
p_u	Node curvature: $p_u := 1 - \frac{1}{2}\sum_{u \sim w} R_{uv}$
κ_{uv}	Edge curvature: $\kappa_{uv} := 2(p_u + p_v)/R_{uv}$
$\|$	Concatenation

Table 23. Notation of Appendix D.

Appendix E

Oversmoothing, "Oversquashing", Heterophily, Long-Range, and more: Demystifying Common Beliefs in Graph Machine Learning

This work is based on the following pre-print:

> [20] **Adrian Arnaiz-Rodriguez** and Federico Errica. "Oversmoothing, "Oversquashing", Heterophily, Long-Range, and more: Demystifying Common Beliefs in Graph Machine Learning". In: *22nd International Workshop on Mining and Learning with Graphs (MLG 2025) at ECML-PKDD 2025*. July 2025. URL: https://arxiv.org/abs/2505.15547

E.1 Introduction

The last decade has seen an increasing scholarly interest in machine learning for graph-structured data [31, 195, 291, 357]. After an initial focus on the design of various message-passing architectures [190], inheriting from the recurrent [342] and convolutional [290] Deep Graph Networks (DGN), together with the analysis of their expressive power [296, 409], researchers later turned their attention to the intrinsic limitations of the message-passing strategy and the relation between the graph, the task, and the attainable performance. We refer, in particular, to the fact that node embeddings may become increasingly similar to each other as more message-passing layers are used [337], the loss of information that results from aggregating too many messages onto a single node embedding [14], the presence of topological bottlenecks [383], the existence of neighbors of different classes [399], and the propagation of information between far ends of a graph [136]. Addressing these limitations makes a difference when applying message-passing models, including foundational ones [50], to large and topologically varying graphs at different scales, from proteins with hundreds of thousands of atoms [421] to dynamically evolving social networks [273] with billions of users, where such limitations manifest together.

A pace of research so rapid can sometimes lead, however, to the premature consolidation of

ideas and beliefs that have not been thoroughly verified. There are several reasons for this to happen: the (perhaps too) intense pressure to publish, follow the latest scientific trends, and demonstrate state-of-the-art performance. As a result, we may end up putting the spotlight on positive findings but overlooking contradictory evidence, eventually accepting hypotheses as canon.

In this chapter, we argue and provide evidence that this is potentially the case for the aforementioned issues of *oversmoothing* (OSM), *oversquashing* (OSQ), *heterophily*, and *long-range dependencies*, driving researchers away by the difficulty of circumventing common beliefs while concurrently introducing novel contributions. Scientific progress is therefore slowed down both in terms of reduced workforce and clarity of the problems to be addressed; failure to acknowledge existing inconsistencies may well lead to a reiterated spreading of questionable claims.

While reviewing the literature, we identify and isolate nine common beliefs that, in our opinion, cause great confusion and ambiguities in the field. We then demystify such beliefs by providing *simple* and possibly memorable counterexamples that should be easy to recall. By encouraging critical thinking around these issues and separating the different research questions, we hope to foster further advancements in the graph machine learning field.

Disclaimer: We remark that the goal of this chapter is not explicitly pinpointing criticalities in previous works; on the contrary, **these works were fundamental to forming our current understanding and ultimately producing this analysis**. We also acknowledge that the list of referenced works cannot be exhaustive, due to the field's size, and it mainly serves to support our arguments.

Table 24 summarizes our findings about common beliefs in the literature, together with the list of papers where we could find mentions of them.[41] We logically divide common beliefs about OSM, OSQ, and the homophily-heterophily dichotomy. In the following sections, we discuss each belief, provide counterexamples, and summarize our arguments with take-home messages.

E.2 Common Definitions and Metrics

This section provides a brief introduction to message-passing models for readers who are less familiar with the topic and its definitions. We also offer a non-exhaustive review of the most commonly used metrics for measuring OSM and OSQ. Additionally, we provide an alternative definition of the computational bottleneck that explicitly links the OSQ problem to the computational graph employed in the message passing mechanism.

Deep Graph Networks

We provide a brief excursus into Deep Graph Networks for readers new to the topic.

We can define a graph as a tuple $g = (\mathcal{V}_g, \mathcal{E}_g, \mathcal{X}_g, \mathcal{A}_g)$, with $\mathcal{V}_g$ the set of nodes, $\mathcal{E}_g$ the set of edges (oriented or not oriented) connecting pairs of nodes. $\mathcal{E}_g$ encodes the topological information of the graph and can be represented as an adjacency matrix: a binary square

[41] Sometimes we found overly general claims in the first sections, later refined, which nonetheless contribute to the spreading of common beliefs.

	Common Belief	References
OSM	1. OSM is the cause of performance degradation.	[60, 88, 89, 101, 104, 105, 129, 141, 204, 207, 214, 216, 222, 234, 265, 269, 284, 305, 334, 336–338, 361, 385, 396, 400, 406, 423, 426, 427]
	2. OSM is a property of all DGNs	[2, 9, 14, 23, 26, 28, 29, 89, 104, 105, 127, 129, 143, 169, 185, 191, 204, 207, 214, 216, 222, 242, 260, 305, 336–338, 348, 383, 400, 405, 406, 410, 423, 426, 427]
Hom-Het	3. Homophily is good, heterophily is bad.	[23, 28, 29, 55, 60, 128, 193, 266, 271, 275, 315, 320, 336, 396, 424, 427, 429]
	4. Long-range propagation is evaluated on heterophilic graphs.	[9, 23, 28, 59, 191, 213, 284, 384, 385]
	5. Different classes imply different features.	[4, 28, 29, 55, 213, 266, 275, 276, 280, 315, 337, 370, 399, 424]
OSQ	6. OSQ synonym of a topological bottleneck.	[9, 23, 26, 28, 29, 34, 35, 39, 59, 105, 124, 126, 127, 129, 143, 169, 182, 191, 197, 213, 222, 234, 270, 305, 348, 350, 356, 361, 370, 383–385, 410, 414]
	7. OSQ synonym of computational bottleneck.	[2, 9, 14, 23, 26, 28, 29, 34, 35, 39, 59, 105, 124, 129, 136, 143, 191, 197, 200, 213, 234, 305, 348, 350, 356, 361, 370, 383, 385]
	8. OSQ problematic for long-range tasks.	[2, 9, 14, 26, 29, 35, 39, 59, 89, 105, 124, 126, 127, 143, 169, 182, 197, 213, 234, 270, 305, 350, 356, 361, 383, 385, 414]
	9. Topological bottlenecks associated with long-range problems.	[9, 59, 169, 234, 270, 350, 383, 385]

Table 24. List of common beliefs together with a non-exhaustive list of papers that make those claims.

matrix $\mathbf{A}$ where $\mathbf{A}_{uv}$ is 1 if there is an edge between u and v, and it is 0 otherwise. Additional node and edge features belong are represented by $\mathbf{x}_v \in \mathcal{X}_g$ and $\mathbf{a}_{uv} \in \mathcal{A}_g$, respectively. $\mathcal{X}_g$ can be, for instance, $\mathbb{R}^d, d \in \mathbb{N}^+$.

The neighborhood of a node v is the set of nodes that are connected to v by an oriented edge, *i.e.*, $\mathcal{N}_v = \{u \in \mathcal{V}_g | (u, v) \in \mathcal{E}_g\}$. If the graph is undirected, we convert each non-oriented edge into two oriented but opposite ones.

The main mechanism of DGNs is the repeated aggregation of neighbors' information, which gives rise to the spreading of local information across the graph. The process is simple: i) at iteration ℓ, each node receives "messages" (usually just node representations) from the neighbors and processes them into a single new message; ii) the message is used to update the representation of that node. Both steps involve learnable functions, so DGNs can learn to capture the relevant correlations in the graph.

Most DGNs implement a synchronous message-passing mechanism, meaning each node always receives information from all neighbors at every iteration step. This local and iterative processing is at the core of DGNs' efficiency since computation can be easily parallelized across nodes. In addition, being local means being independent of the graph's size. When one learns the same function for all message passing iterations, we talk about *recurrent* architectures, as the GNN of Gori, Monfardini, and Scarselli [195] and Scarselli et al. [342]; on the contrary, when one learns a separate parametrization for a finite number of iterations (also known as layers), we talk about *convolutional* architectures as the NN4G of Micheli

[290] and Micheli and Sestito [291].

The neighborhood aggregation is usually implemented using permutation-invariant functions, which make learning possible on cyclic graphs that have no consistent topological ordering of their nodes. A rather general and classical neighborhood aggregation mechanism for node v at layer/step $\ell+1$ is the following:

$$\mathbf{h}_v^{\ell+1} = \phi^{\ell+1}\Big(\mathbf{h}_v^{\ell},\ \Psi(\{\psi^{\ell+1}(\mathbf{h}_u^{\ell}) \mid u \in \mathcal{N}_v\})\Big) \tag{98}$$

where $\mathbf{h}_u^{\ell}$ is the node embedding of u at layer/step ℓ, ϕ and ψ implement learnable functions, and Ψ is a permutation invariant aggregation function. Note that $\mathbf{h}_v^0 = \mathbf{x}_v$. For instance, the Graph Convolutional Network of Kipf and Welling [245] implements the following aggregation, which is a special case of the above equation:

$$\mathbf{h}_v^{\ell+1} = \sigma(\mathbf{W}^{\ell+1} \sum_{u \in \mathcal{N}(v)} \hat{\mathbf{L}}_{uv} \mathbf{h}_u^{\ell}), \tag{99}$$

with $\hat{\mathbf{L}}$ being the normalized graph Laplacian, $\mathbf{W}$ is a learnable weight matrix and σ is a non-linear activation function.

Some OSM Definitions

In order to keep the work self-contained and to illustrate the different ways OSM has been defined, we briefly review the most common definitions. The following subsection summarizes the most commonly used OSM metrics. Some of these metrics are used in the main text.

Let $D = \mathrm{diag}(d_1, \ldots, d_n)$ be the degree matrix with $d_u = \sum_v A_{uv}$. The combinatorial graph Laplacian is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, with eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. In addition, the symmetric normalized Laplacian, defined by $\hat{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{1/2}\mathbf{A}\mathbf{D}^{1/2}$, normalizes each entry to remove the influence of node degrees and its spectrum lies in $[0, 2]$.

DE [108] was used to measure OSM in Cai and Wang [88] is defined as

$$\hat{\mathrm{DE}}(\mathbf{H}^{\ell}) = \mathrm{Tr}((\mathbf{H}^{\ell})^T \hat{\mathbf{L}} \mathbf{H}^{\ell}) = \frac{1}{2} \sum_{u,v \in \mathcal{E}} \left\| \frac{\mathbf{h}_u}{\sqrt{d_u}} - \frac{\mathbf{h}_v}{\sqrt{d_v}} \right\|_2^2 \tag{100}$$

Note that DE can also be computed using the non-normalized Laplacian, $\mathbf{L}$:

$$\mathrm{DE}(\mathbf{H}^{\ell}) = \mathrm{Tr}((\mathbf{H}^{\ell})^T \mathbf{L} \mathbf{H}^{\ell}) = \frac{1}{2} \sum_{u,v \in \mathcal{E}} \|\mathbf{h}_u - \mathbf{h}_v\|_2^2 \tag{101}$$

Other OSM metrics include with respect to the norm of node embeddings. For instance, Rayleigh Quotient (RQ) [108], which can be seen as normalized DE, is defined by

$$\mathrm{RQ} = \frac{\mathrm{Tr}((\mathbf{H}^{\ell})^T \hat{\mathbf{L}} \mathbf{H}^{\ell})}{\|\mathbf{H}^{\ell}\|_F^2} \tag{102}$$

first proposed for OSM analysis in [88] and later used by [128, 284, 336, 411]. RQ discerns whether embeddings become *relatively* smoother, independent of magnitude.

Mean Absolute Deviation (MAD) [101] averages cosine dissimilarity between a node and its neighbors.

$$\mathrm{MAD}_G(\mathbf{H}^\ell) = \frac{1}{n}\sum_{v\in\mathcal{V}}\sum_{u\in\mathcal{N}_v} 1 - \frac{\mathbf{h}_v^{\ell\,T}\mathbf{h}_u^\ell}{\|\mathbf{h}_v^\ell\|\,\|\mathbf{h}_u^\ell\|} \tag{103}$$

The smoothness metric SMV [269] captures a global node-distance average over all node pairs:

$$\mathrm{SMV} = \frac{1}{n}\sum_{u\in\mathcal{V}}\frac{1}{n-1}\sum_{v\neq u\in\mathcal{V}}\frac{1}{2}\left\|\frac{\mathbf{h}_u}{\|\mathbf{h}_u\|} - \frac{\mathbf{h}_v}{\|\mathbf{h}_v\|}\right\| \tag{104}$$

In the main text, we primarily consider DE and RQ. However, all notions convey the same intuition, loss of discriminative variation, yet, as we show, can disagree in practice.

Convergence-rate results. Known theoretical bounds show $\mathrm{DE}(H^k)$ decays exponentially with depth k (*e.g.*, the rate depends on weight spectra and graph eigenvalues) [214, 308]. Such convergence rate results primarily make assumptions on the architecture, weight matrix, and activation functions, and may be altered by skip connections, normalization layers, or simple rescaling such as $2W$ [336].

For instance Cai and Wang [88] propose a bound on the DE of two consecutive message passing layers (similar bounds found in [308, 426])

$$\hat{\mathrm{DE}}(\mathbf{H}^\ell) \leq (1-\lambda_2)^2 s_{\max}^\ell \hat{\mathrm{DE}}(\mathbf{H}^{\ell-1})$$

being $s_{\max}^\ell$ the square of the maximum singular value of W^ℓ, and λ_2 the second smallest eigenvalue of the Laplacian, *i.e.*, the spectral gap. The proof holds when $s_{\max}^\ell < 1/(1-\lambda_2)$.

In addition, Di Giovanni et al. [128] further relate Laplacian eigenvalues and weight spectra to predict whether RQ converges to 0 (collapse) or to $\lambda_{\max}$ (no collapse).

Some OSQ Definitions

For completeness, we summarize some of the most commonly used metrics in the OSQ literature and their relationships. These quantities mainly measure three different aspects of the graph: (i) *sensitivity/Jacobian* measures that capture how information from a distant node u affects a target node v after K message-passing layers; (ii) *topological bottleneck* proxies such as Cheeger-type cut ratios, graph spectrum or curvature scores; and (iii) *distance-based* quantities such as effective resistance that upper-bound information flow.

Sensitivity For a K-layer GNN let $h_u^{(k)}$ denote the embedding of node u at layer k. A first proxy for OSQ is the *Influence Score* of Xu et al. [410],

$$I(u,v) = \left\|\frac{\partial \mathbf{h}_u^K}{\partial \mathbf{h}_v^0}\right\|. \tag{105}$$

Black et al. [59] sum these sensitivities over *all* unordered pairs,

$$\sum_{u\neq v\in V}\left\|\frac{\partial \mathbf{h}_u^K}{\partial \mathbf{h}_v^0}\right\|, \tag{106}$$

to obtain a graph-level indicator of how much information is lost.

Di Giovanni et al. [128] propose a *symmetric Jacobian obstruction* that removes self-influence and degree bias. They define the Jacobian obstruction of node v with respect to node u at layer m as

$$\tilde{\mathbf{J}}_k^{(m)}(v,u) := \left(\frac{1}{d_v}\frac{\partial \mathbf{h}_v^{(m)}}{\partial \mathbf{h}_v^{(k)}} - \frac{1}{\sqrt{d_v d_u}}\frac{\partial \mathbf{h}_v^{(m)}}{\partial \mathbf{h}_u^{(k)}}\right) + \left(\frac{1}{d_u}\frac{\partial \mathbf{h}_u^{(m)}}{\partial \mathbf{h}_u^{(k)}} - \frac{1}{\sqrt{d_v d_u}}\frac{\partial \mathbf{h}_u^{(m)}}{\partial \mathbf{h}_v^{(k)}}\right), \tag{107}$$

being the extension to the Jacobian obstruction of node v with respect to node u after m layers defined as

$$\tilde{\mathsf{O}}^m(u,v) = \sum_{k=0}^{m} \left\| \tilde{\mathbf{J}}_k^{(m)}(v,u) \right\|. \tag{108}$$

Topological Bottlenecks Many OSQ papers measure topological (structural) bottlenecks using spectral or curvature quantities. Note that here we give an intuition based on the spectral metrics [23, 35, 234], but a significant part of the literature uses metrics based on curvature [191, 270, 305, 383].

First, the topological bottleneck can be measured by Cheeger Constant [108], which is the size of the min-cut of the graph.

$$h_G = \min_{S \subset V} \frac{|\{e = (u,v) : u \in S, v \in \bar{S}\}|}{\min\{\mathrm{vol}(S),\ \mathrm{vol}(\bar{S})\}}$$

A small h_G means one can separate G into two large-volume parts by removing only a few edges, *i.e.*, a severe *topological bottleneck*. Cheeger's inequality links h_G to the spectrum of G:

$$\frac{h_G^2}{2} \leq \lambda_2 \leq 2h_G,$$

where λ_2 is the second eigenvalue of the normalized Laplacian.

Pairwise Distances The commute time between two nodes [274] is defined as the expected number of steps that a random walker needs to go from node u to v and come back. The Effective Resistance between two nodes [99], R_{uv}, is the commute time divided by the volume of the graph [246], which is the sum of the degrees of all nodes in the graph. The effective resistance between two nodes is computed as

$$R_{u,v} = L^+_{ii} + L^+_{jj} - 2L^+_{ij}$$

being $\mathbf{L}^+ = \sum_{i>0} \frac{1}{\lambda_i}\phi_i\phi_i^T$ the pseudoinverse of $\mathbf{L}$

Then, some measures derived from this metric can be connected with the topological bottleneck [99, 108, 325]. For instance, the maximum effective resistance of a graph is connected with the Cheeger constant as per $R_{\max} = \max_{u,v\in\mathcal{V}} R_{uv}$

$$R_{\max} \leq \frac{1}{h_G^2}$$

and thus also bounded by the spectral gap as

$$\frac{1}{n\lambda_2} \leq R_{\max} \leq \frac{2}{\lambda_2}.$$

In addition, the total effective resistance $R_{\text{tot}} = 1/2 \sum_{u,v \in \mathcal{V}} R_{uv}$ is bounded to the spectral gap [140]:

$$\frac{n}{\lambda_2} \leq R_{\text{tot}} \leq \frac{n(n-1)}{\lambda_2}$$

Note that the total effective resistance also equals the sum of the spectrum of $\mathbf{L}^+$ $R_{\text{tot}} = n \sum_2^n 1/\lambda_n$.

Connecting Sensitivity and Topological Distances The larger Total Effective Resistance ($R_{tot} = \sum_{(u,v)\in V} R_{uv}$) is, the lower the sum of pairwise sensitivities [59]:

$$\sum_{u,v \in V \times V} \left\| \frac{\partial h_v^{(r)}}{\partial h_u^{(0)}} \right\| \leq c(b - R_{tot}) \tag{109}$$

The larger the Effective Resistance is, the higher the Symmetric Jacobian Obstruction [128]:

$$\tilde{\mathsf{O}}^m(u,v) = \sum_{k=0}^{m} \left\| \tilde{\mathbf{J}}_k^{(m)}(v,u) \right\| \leq c \; R_{u,v} \tag{110}$$

Computational Bottleneck

Oversquashing can also be seen through the perspective of the message-passing computational graph: each message-passing layer expands the set of nodes whose features can influence a target node. If this *receptive field* grows fast, any fixed-width DGN "squash" many signals into a single vector.

Receptive Field Following [14, 102] the receptive field was defined recursively as:

$$\mathcal{N}_v^K := \mathcal{N}_v^{K-1} \cup \{w \mid w \in \mathcal{N}_u \wedge u \in \mathcal{N}_v^{K-1}\} \quad \text{and} \quad \mathcal{N}_v^1 = \mathcal{N}_v \tag{111}$$

which can be also seen as the set of K-hop neighbors neighbors, *i.e.*, nodes that are reachable from v within K hops The number of nodes in each node's receptive field can grow exponentially with the number of layers $|\mathcal{N}_v^K| = \mathcal{O}\left(\exp(K)\right)$ [102]. For instance, in a rooted binary tree each layer has exactly b^{K-1} new neighbors, so $|\mathcal{N}_v^K| = 1 + b + b^2 + \cdots + b^{K-1} = \Theta(b^K)$.

When evaluating the actual *computational* graph resulting from message-passing, duplicates matter: a node that appears in several branches of the computation tree contributes multiple times, since each distinct walk contributes a separate message. We therefore use the multiset notation to define the computational tree for a node v:

$$\mathcal{M}_v^1 := \mathcal{N}_v, \quad \mathcal{M}_v^K := \mathcal{M}_v^{K-1} \uplus \left\{ \biguplus_{u \in \mathcal{M}_v^{K-1}} \mathcal{N}_u \right\}. \tag{112}$$

Therefore, we can define the notion of an "exponentially-growing receptive field" [14] as follows.

Definition E.1 (Computational Bottleneck)**.** *For a given node v and number of message passing layers K, the computational bottleneck of node v is defined as $|\mathcal{M}_v^K|$.*

The size of the computational bottleneck (multiset receptive field) at node v, can be computed as:

$$|\mathcal{M}_v^K| := \sum_{\ell=1}^{K} \|A^\ell[v,:]\|_1 = \sum_{\ell=1}^{K} \sum_{u \in \mathcal{V}} (A^\ell)_{u,v} \tag{113}$$

This definition counts every distinct length-ℓ walk from v to any node u. Equation (113) is exactly the row–sum of the powers of the adjacency matrix; it therefore matches the size of the *computational tree*.

Note that the size of the set-based receptive field corresponds to the support of the multiset $\mathcal{M}_v^K$, denoted $\mathcal{N}_v^K := \text{supp}(\mathcal{M}_v^K)$. Therefore, the multiset size $|\mathcal{M}_v^K|$ is always greater than or equal to the size of the support, $|\mathcal{M}_v^K| \geq |\mathcal{N}_v^K|$, since it accounts for path multiplicity.

In early deep-graph networks literature, Micheli [290] introduced the idea by using the term "contextual window": deeper layers aggregate exponentially many paths unless skip connections or global pooling curb the growth. The multiset perspective in Equation (113) makes this explosion explicit and the matrix computation directly links to matrix-power interpretations of message passing.

Figure 56 visualises the difference between the set size $|\mathcal{N}_v^K|$ and the multiset size $|\mathcal{M}_v^K|$ on a toy graph and on a stochastic block model (SBM).

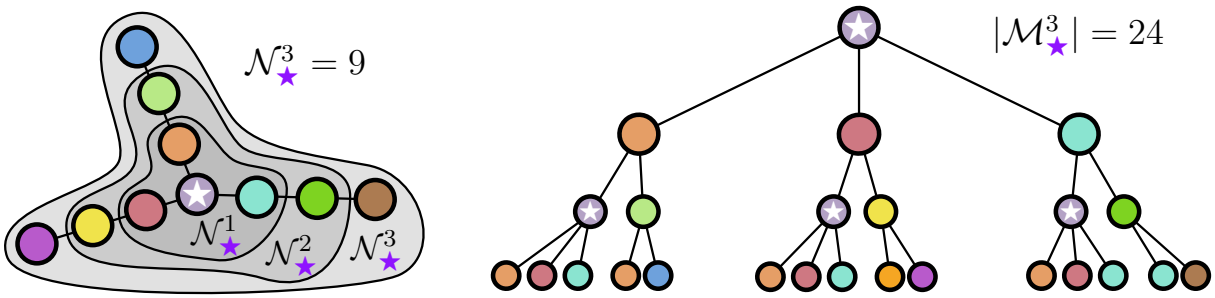


Figure 56. Computational Bottleneck. Illustration of a receptive field – defined with sets (K-hop neighborhood) – and the definition of computational bottleneck measured as the size of the computational graph – defined with multisets.

In conclusion, we note that in message-passing, the computational bottleneck is driven not by how many distinct vertices are in the K-hop neighborhood, but by the size of the computational graph.

E.3 Is Oversmoothing Really a Problem?

Oversmoothing broadly refers to the phenomenon where, as we stack more message-passing layers in a DGN, the node embeddings become increasingly similar to each other eventually collapsing into a low-dimensional [214, 308]–or even single-vector [88]–subspace. This creates an almost constant representation, independent of the original node-feature distribution, and can **potentially** result in loss of discriminative power along the way [88, 265, 308].

Formally, et $H^\ell \in \mathbb{R}^{n \times d}$ be the matrix of node embeddings after ℓ message-passing layers in a DGN, where n is the number of nodes and d is the hidden dimension. Consider a similarity

(or separation) function $\pi\colon \mathcal{H} \to \mathbb{R}$, where $\mathcal{H} = \mathbb{R}^{n\times d}$ is the space of all possible embedding matrices. We say that a DGN experiences OSM if

$$\lim_{\ell\to\infty} \pi(H^{\ell}) \;=\; c. \tag{114}$$

where c is some constant indicating a collapse of embeddings. Deviation metrics (*e.g.*, Dirichlet Energy [88], MAD [101]) and subspace-collapse criteria [213, 308, 336] all measure the same OSM intuition: as depth increases, node embeddings shrink toward a nearly constant subspace.

E.3.1 Belief: OSM is a Property of All DGNs.

A widespread claim in the literature is that *OSM happens regardless of the specific architecture or the underlying graph.* Early theoretical works support this view by analyzing message-passing propagation as a diffusion process: iterated normalized-Laplacian updates converge to a degree-weighted stationary distribution [88, 191], while heat-kernel diffusion converges to a constant vector [22, 308]. The resulting bounds quantify the rate of OSM in terms of the singular values of the feature transform W and the eigenvalues of the graph structure G.

These conclusions, however, rely on restrictive assumptions. Later work has relaxed them by introducing learnable feature transforms, non-linear activations, and more elaborate architectures [88, 308, 406], yet no existing proof shows inevitable collapse under realistic training regimes. In practice, remedies such as residual/skip connections, normalization layers, or gating mechanisms are explicitly architectural changes designed to maintain local distinctions, calling into question the universality of this OSM claim.

In addition, many studies probe OSM with *untrained* (weights frozen at initialization) linear GCN stacks, which is an experimental choice that hides the effect of learning and may lead to the wrong conclusions, as noted by Zhang et al. [419]. Indeed, Cong, Ramezani, and Mahdavi [112, Figure 2] report OSM only for frozen-weight networks; once parameters are allowed to adapt, the models preserve informative variance.

Together, these observations suggest that OSM is not an inevitable consequence of message-passing but rather a contingent outcome that depends on training dynamics and architectural design choices.

Empirical Example We show a simple training scenario where we see how OSM is *not* a property of all DGNs and how different elements make it difficult to draw clear conclusions. We train several DGNs under two propagation variants: the vanilla AXW update and the rescaled $AX(2W)$, inspired by Roth and Liebig [336, Figure 1]. We measure OSM with 2 different metrics: Dirichlet Energy (DE) and its norm-normalized version, the Rayleigh Coefficient (RQ), which was also previously used in some works [88, 128, 284, 336]. Figure 57 shows three key facts: (i) some architectures never collapse, (ii) a minor rescaling can reverse the trend, and (iii) DE and RQ often disagree. Hence OSM is neither universal nor straightforward to diagnose.

First, OSM, as measured by Dirichlet Energy (DE), is not universal: GIN's DE explodes instead of collapsing in the vanilla setting (a), which shows that changing the aggregation-function may lead to an opposite behavior of OSM effect.

Second, a minor scaling in the feature transformation ($2W$ instead of W) flips the behavior of several models: curves that decayed in (a) now grow or stabilize, and vice-versa. Similar small tweaks (normalization layers, self-loops, alternative aggregators) can likewise create or remove DE collapse, which has been leveraged by prior work to propose OSM mitigation approaches, such as the ones based on feature normalization [423, 427].

Finally, whether OSM is observed depends on the measure of choice. DE reflects raw smoothness, whereas the RQ normalizes by the feature norm. As a result, the same model can exhibit mutually contradictory trends for different metrics [419]. First, GCN's DE collapses under the vanilla aggretation (a), explodes with a simple rescaling (b), yet RQ remains essentially flat in both normalised plots (c–d), indicating that GCN embeddings are being rescaled, not necessarily oversmoothed. On the contrary, GAT and SAGE, which had similar behavior as GCN in (a) and (b), decay using RQ (c-d) at a (roughly) linear rate, highlighting how different architectures respond differently to the use of RQ instead of DE. Furthermore, subfigure (d) shows that $AX(2W)$ can stabilize RQ for certain methods, suggesting that normalizations or small parameter adjustments do not affect all models uniformly. Therefore, conclusions about OSM depend heavily on metric and model, an observation that we rarely found in the literature.

In conclusion, OSM is neither inevitable nor uniquely defined: its observation hinges on the architecture, on seemingly innocuous hyper-parameters, and on which stability metric (DE vs. RQ) one adopts. Therefore, a natural question arises: does OSM actually limit the models' predictive accuracy? We investigate this question in the next section.

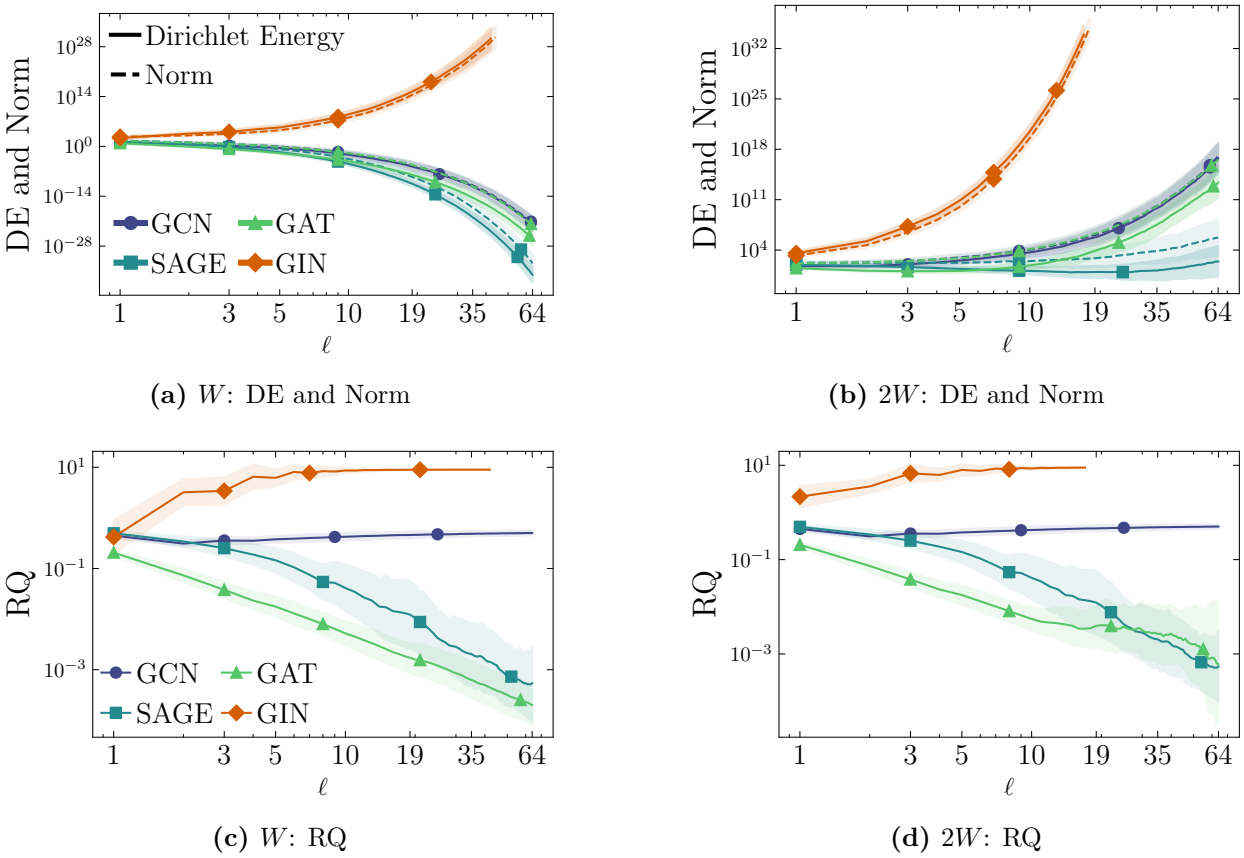
E.3.2 Belief: OSM is the Cause of Performance Degradation.

Part of the literature focuses on the narrative that OSM is the cause of lower test accuracy in DGNs. The hypothesis is that if embeddings collapse to a non-meaningful space, then the separability of the nodes will become challenging and accuracy decreases.

However, this hypothesis ignores some critical aspects, such as *i)* the separability of node embeddings with respect to the nodes' labels, and *ii)* how such separability evolves in the intermediate OSM phase (*if* it happens, as we discussed before).

Regarding the first statement, although it is true that if there is total collapse to the same value then the embeddings will not be informative at all, the main problem remains the node embeddings' separability. As shown in the previous section, some changes in the architectures can avoid OSM, but they might have no impact on the overall accuracy. For instance, multiplying by two the weight matrix leads to a general increase in DE for all architectures, however, the accuracy will remain the same since the embeddings have been simply scaled up and the embeddings' separability is not affected negatively. In addition, avoiding an embedding collapse does not necessarily lead to an improvement in generalization accuracy. For instance, comparing a GCN with and without bias, both versions show a decrease in performance as the number of layers increases, whereas only GCN shows a collapse in DE [337, Figure 3].

On the other hand, and related to the second statement, embedding collapse will not always lead to a decrease in accuracy. OSM happens faster in some subspaces than in others and this effect will be beneficial if labels are correlated with those subspaces [242]. For instance, if we classify points into two classes, and all nodes of distinct classes collapse into different points, the OSM metric will detect such collapse. However, the separability of the



(a) W: DE and Norm

(b) $2W$: DE and Norm

(c) W: RQ

(d) $2W$: RQ

Figure 57. **(a-b)**: We depict the evolution, with increasing number of layers, of the DE $= \mathrm{Tr}(X^T \Delta X)$ and the feature norm $\|X\|_F$, using W and $2W$ feature transformations for different architectures. **(c-d)**: Evolution of the RQ $= \mathrm{Tr}(X^T \Delta X)/\|X\|^2$ for W and $2W$ as before.

node embeddings will remain possible, illustrating also the limits of wide-spread used OSM metrics with respect to label information.

This intuitive behavior has been identified in the literature as a form of "beneficial" smoothing phase [242, 336, 405]). In this phase, the nodes of each class first collapse into a class-dependent point before the *potential* second stage, at which point all nodes converge to the same representation. Finally, although overall pairwise distances might shrink in deep layers, *within-class* distances might contract more than *between-class* ones, so class separability improves despite the global collapse, as discussed by Cong, Ramezani, and Mahdavi [112].

In conclusion, low accuracy in DGNs cannot be attributed to OSM alone. The separability of node embeddings plays a major role, where other training problems such as vanishing gradients or over-fitting also arise when using a big number of layers [26, 112, 411, 423].

Message of the Section

i) OSM is not a property of all DGNs

ii) OSM is not necessarily the cause of performance degradation. The performance is related to node embeddings' separability, which can be also affected by many other elements, such as vanishing gradients.

iii) Therefore, to study the performance of DGNs, it might be better to study how they achieve separability of node embeddings, and how the OSM relates to node separability.

E.4 Homophily-Heterophily and the Role of the Task

In the context of node classification, the term *homophily* (resp. *heterophily*) generally refers to some form of similarity (resp. dissimilarity) between a node and its neighbors [286]. This (dis)similarity can be measured with respect to class labels, node features, or both; the vast majority of works in the literature opt for the first, *but this choice is often implicit and taken for granted*, making some statements hard to interpret when one is aware of the other ways to measure it.

E.4.1 Belief: Homophily is Good, Heterophily is Bad

A recurrent narrative in the literature is that the message-passing mechanism of DGNs is particularly suited for homophilic graphs, whereas it is unfit for heterophilic graphs. The apparent motivation is that, in homophilic graphs, all you need to do to solve a node classification task is to look at similar neighbors, and a local message-passing strategy implements just the right inductive bias. This is in contrast to a class-heterophilic graph, where there exist neighboring nodes of a different class that might make it harder for message-passing to isolate the "relevant information", intended as neighbors of the same class. Such a belief is supported by empirical evidence on a rather restricted set of benchmarks [301, 315, 346] with varying levels of homophily/heterophily.

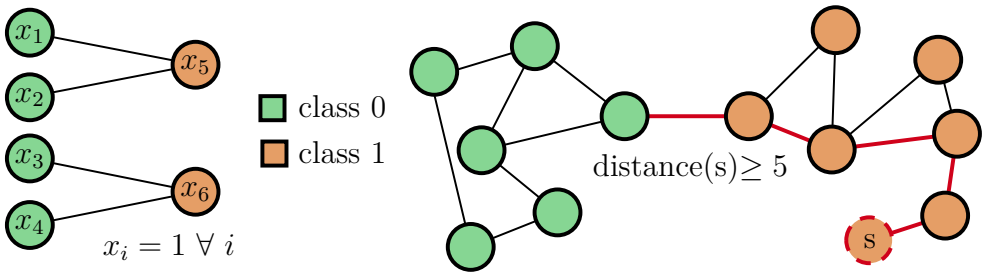


Figure 58. *Left:* a fully heterophilic graph inspired by [280] where a 1-layer, sum-based DGN can perfectly classify the nodes due to a difference in the node degree. *Right:* a highly homophilic graph where the task is to predict if a node is at a distance greater than five from a specific node. Here, the performances of a DGN will be poor unless information from nodes of another community – from the perspective of a class-0 node – is captured.

Researchers already tried to challenge these considerations in the past [142, 277, 280, 319, 399]. Consider Figure 58 (left), where a bipartite graph with identical node features is fully class-heterophilic. If we apply a single sum-based graph convolution, the nodes can be perfectly classified as the resulting embeddings depend on the incoming degree. Therefore, **there exist heterophilic graphs where a DGN can achieve perfect classification**. On the contrary, Figure 58 (right) depicts a highly class-homophilic graph, where nodes belong to one of two classes if they are at a distance greater or lower than five from a given node s. In this case, the information on the nodes does not even matter; if we were to follow the above belief, we would be encouraged to use a few layers of message-passing, and as a result **we could never solve the task perfectly**.

E.4.2 Belief: Different Classes Imply Different Features

In the previous section, we deliberately ignored the interplay between a node's features and its class label, which is induced by the task at hand. The reason is that we wanted to clarify the distinction with another, more subtle, and problematic belief: nodes belonging from different classes should have different (to be read as separable) feature distributions. Under this assumption, it is also implicit that class homophily will imply feature homophily.

Such an assumption is often key in arguments supporting the belief of Appendix E.4.1. Indeed, if nodes of different classes have different feature distributions, then applying a local message-passing iteration to a highly class-homophilic graph should "preserve the distance" between node embeddings of different classes. On the contrary, in a heterophilic setting, a graph convolution would mix information coming from different feature distributions, which may be detrimental to performances.

The logic is not incorrect per-se, but our key counterargument is the following: if different classes imply different feature distributions, why would one need to apply a DGN rather than a simple MLP? In other words, **either there is a very strong assumption** that the task does not depend on the topological information, or the feature-class distributions induced by the task allow us to somehow take a shortcut in terms of learned function, neglecting the role that the topology might have.

It appears therefore necessary to consider less trivial and more fine-grained scenarios, where the feature distributions of different classes partially or totally overlap, the topology has a key role in the task definition, and topological properties induce a positive/negative effect on the performance of message-passing models as done, for instance, in recent works [95, 424].

E.4.3 Belief: Long-range Propagation is Evaluated on Heterophilic Graphs

The common beliefs of Appendices E.4.1 and E.4.2 have been used to support yet another argument, namely that we should evaluate the ability of DGNs to propagate long-range information on heterophilic graphs. The rationale seems to be that, in order for DGNs to perform well, nodes of a given class should focus on information of similar (w.r.t. class and/or features) nodes; therefore, in a heterophilic graph, it may be necessary to capture information far away (*i.e.*, long-range) from the immediate neighborhood.

Once more, what is really important is to **distinguish the task**, *e.g.*, one that depends

on long-range propagation, **from the class labels the task induces on the nodes**. As a matter of fact, the "long-range" task of Figure 58 (right) induces a highly homophilic graph, while the heterophilic graph of Figure 58 (left) is not associated with a long-range propagation task. Therefore, we cannot draw a generic relation between long-range tasks and heterophily without making further assumptions.

Message of the Section

i) Generic claims about the performance of DGNs under homophily and heterophily do not hold, nor does their relation with long-range problems.

ii) Homophily/heterophily is a function of the task, but the converse is not true.

iii) We should move past the coarse homophily-heterophily dichotomy and **focus more on the task** and the interplay it induces between features, structure, and class labels.

E.5 The Many Facets of "Oversquashing" and their Negative Implications

The term oversquashing originated from Alon and Yahav [14] and referred to an "*exponentially growing information into fixed-size vector*" by repeated application of message-passing. In other words, oversquashing was associated with the **computational tree** (Figure 59) induced by message-passing layers on each node of the graph. Later, oversquashing was connected by [383] to the existence of **topological bottlenecks**: "*edges with high negative curvature are those causing the graph bottleneck and thus leading to the oversquashing phenomenon*". Since then, researchers have adopted one or even both definitions of oversquashing at the same time, contributing to an apparent understanding that these definitions subsume the same problem.

In this section, we argue that **this is not the case** and that, as a community, **we should clearly separate the term "oversquashing"** into *(at least)* two separate terms:

Computational Bottlenecks and **Topological Bottlenecks**

Computational bottlenecks, defined in Definition E.1, are inevitably related to the *message-passing architecture*, for which the graph is the computational medium, whereas topological bottlenecks refer to the *graph connectivity*. These two problems are clearly intertwined, but in the following we show that they do not always coexist, hence it makes sense to treat them as **fundamentally distinct problems**.

E.5.1 Belief: Oversquashing as Synonym of Topological Bottleneck

The prolific line of work that associates "oversquashing" with topological bottlenecks seems to have gained popularity with Topping et al. [383]. In that paper, edges with negative curvature are first associated with (topological) bottlenecks, then a theorem puts in relation message-passing on a graph containing a bottleneck with the Jacobian sensitivity of node

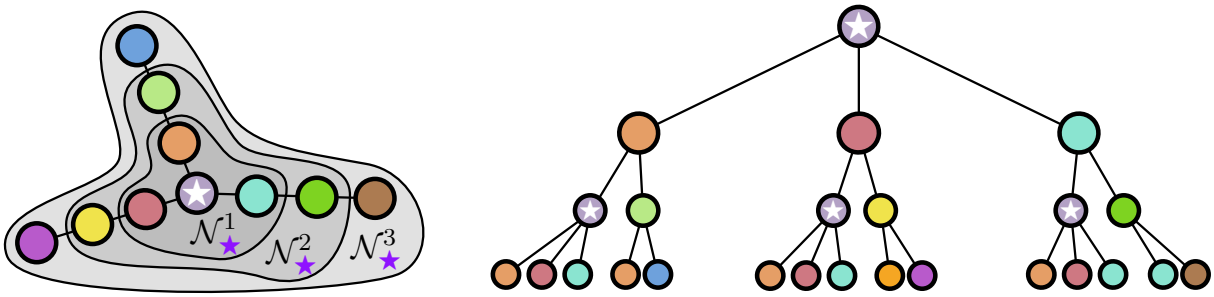


Figure 59. We intuitively visualize what happens when we repeatedly aggregate the neighborhood of the star node using the message-passing paradigm. We define in Appendix E.2 a computational bottleneck as the size of this computational tree (24 computational nodes *vs* nine 3-hop neighbors).

representations, typically defined by $I(u, v) = \|\partial \mathbf{h}_u^K / \partial \mathbf{h}_v^0\|$ and denoting how much the final representation of a node u after K layers is influenced by the initial representation of a node v [410]. Put simply, a topological bottleneck may imply low sensitivity. **However, the converse is not necessarily true:** we can have low sensitivity on a graph where there are no bottlenecks. Figure 60 (left) shows a grid graph where there are no topological bottlenecks. The repeated application of message passing will, however, quickly generate a computational bottleneck. Therefore, saying that there are no topological bottlenecks does not imply that there are no computational bottlenecks.

To improve on topological bottlenecks, a widely investigated approach is graph rewiring [29], which was also the subject of scrutiny recently [384, 385]. Rewiring is based on the intuition that improving topological bottlenecks metrics should improve the performance of DGNs [23, 35, 124, 234], by reducing the distance between nodes that should communicate. At the same time, it should become clear now that, under the DGN paradigm, *rewiring might worsen the computational bottleneck* – as long as the same number of message-passing layers is used – while improving the topological one. This perspective was also put forward by Errica et al. [143], with a theoretical analysis on how message filtering, as shown in Figure 58 (middle), reduces both the computational bottleneck and sensitivity yet improves performances while leaving the graph structure unaltered.

Another, slightly more technical way to see why low Jacobian sensitivity does not imply the presence of any topological bottlenecks is to follow the chain of upper bounds that link the metrics used to measure the computational and topological bottlenecks [23, 59, 128, 234]. Several recent works [59, 127] have shown that sensitivity is bounded by above by a term that includes the effective resistance, a purely topological distance metric that quantifies the expected commute time of a random walk between nodes u and v [246]. In particular, the bound subtracts the lower bound on the maximum effective resistance, which depends on the inverse of the spectral gap (a proxy for topological bottlenecks) [99, 274]. This connection provides a useful intuition: as the topological bottleneck gets worse, the lower bound on the maximum effective resistance increases, which in turn reduces the upper bounds on the sensitivity of Black et al. [59]. **However, the converse does not hold.** There are graphs where the maximum effective resistance between distant nodes is large despite the absence of topological bottlenecks. For instance, consider the example of a grid graph in Figure 60 (left) where there is no topological bottleneck, yet the effective resistance between diagonally opposite corners grows linearly with the grid size. As a result, sensitivity between those nodes

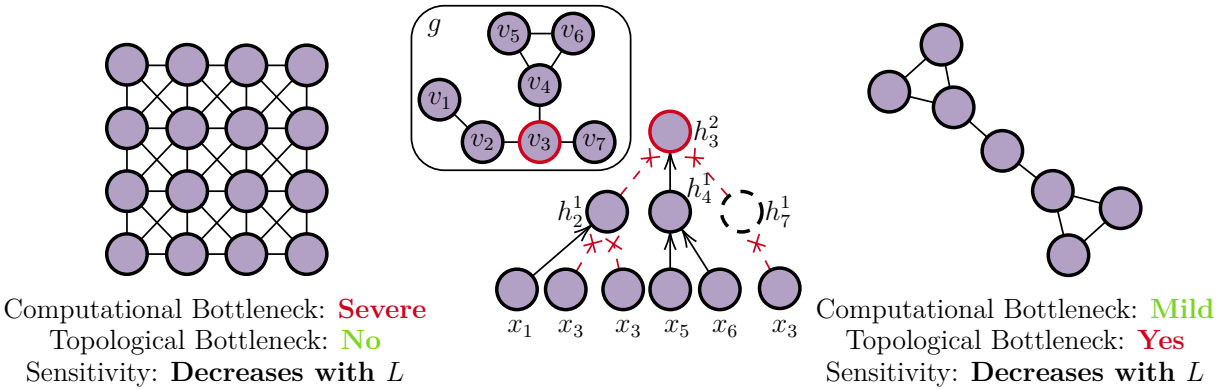


Figure 60. *Left:* in a grid graph, the computational bottleneck grows very quickly but there is no topological bottleneck. *Middle:* A visualization of the computational graph rooted at node v_3 for two message passing layers, highlighting how pruning messages reduces the computational bottleneck. *Right*: In this graph, there is a topological bottleneck and a mild computational bottleneck. As with the grid graph (Appendix E.5.1), the sensitivity decreases with the number of message-passing layers.

decays with depth, even though the graph has no identifiable topological bottlenecks. This illustrates that computational bottlenecks can arise independently of topological ones, **and that low sensitivity does not necessarily imply the presence of either of them**.

To show that low sensitivity does not necessarily imply a topological bottleneck, Figure 61 analyzes sensitivity's decreasing trend when the number of message passing layers L increases on the grid graph of Figure 60 of size 10×10. Increasing the size of the embedding space postpones the collapse of the sensibility.

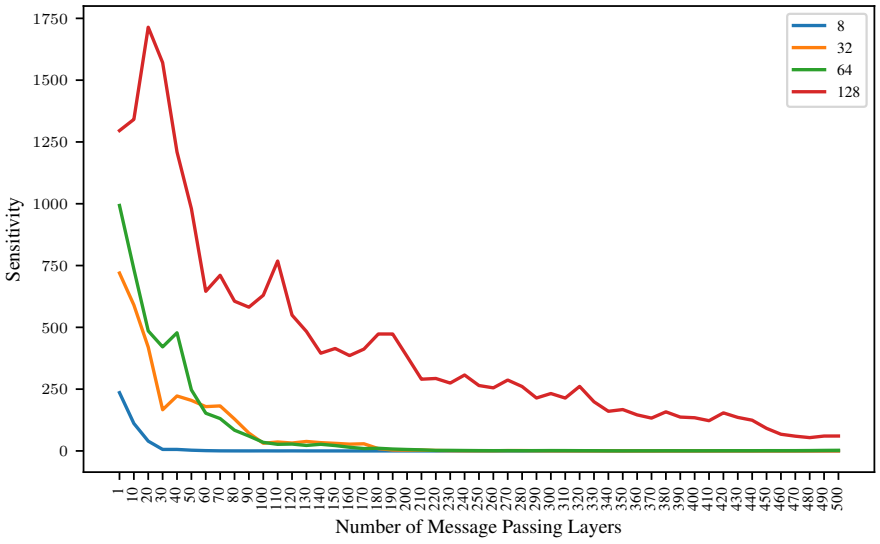


Figure 61. Sensitivity Decreases on a Grid Graph without Topological Bottlenecks. We plot the sensitivity of the grid graph of Figure 60 for the Graph Convolutional Network [245] model for different node embedding sizes.

E.5.2 Belief: Oversquashing as Synonym of Computational Bottleneck

We briefly complement the previous section with a discussion on "oversquashing" as a computational bottleneck, which was introduced by Alon and Yahav [14] and has been the (often implicit) study subject of works that prune, to some extent, the computational tree induced on every node by the iterative message-passing process [143, 334]. Also in this case, there exist cases where reducing the computational bottleneck may be harmful: Figure 60 (right) shows a graph where there is a topological bottleneck but no severe computational bottleneck (for a limited number of layers). In this case, excessive pruning of the computational tree might cause distant nodes to interrupt all communications. Therefore, computational and topological bottlenecks are problems that should be tackled separately.

E.5.3 Belief: Oversquashing is Problematic for Long-range Tasks

Since its definition by Alon and Yahav [14], oversquashing has often been considered a problem in long-range tasks. The reason stems from its relation to the exponentially growing computational tree as the number of message-passing layers increases: whenever a node has to receive information from another node at distance d, classical (synchronous) message-passing architectures need to apply at least d layers to capture that information. As a result, the relevant information may get lost due to the exponentially large computational bottleneck. Importantly, topological bottlenecks can only make the problem worse, by forcing the information of a group of messages to be squeezed through an edge – please refer to the next section for a discussion about long-range tasks and topological bottlenecks.

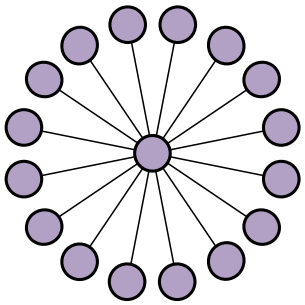


Figure 62. Hubs can exhibit computational bottlenecks.

The main message here is that long-range tasks "force" classical message-passing architecture to create a computational bottleneck to propagate the necessary information, **but one can observe computational bottlenecks even in short-range tasks**. An obvious example is the high-degree node of Figure 62, where, after *just one* layer, the center node receives a high number of messages, effectively creating a computational bottleneck in terms of information to be squashed into a fixed-size vector. Therefore, while the task of long-range is related to computational bottlenecks under classical DGNs, computational bottlenecks are not a prerogative of long-range tasks.

E.5.4 Belief: Topological Bottlenecks Associated with Long-range Problems

The last belief we discuss is that topological bottlenecks are the primary obstacle to solving long-range tasks. This intuition stems from the fact that narrow cuts impede information flow between distant parts of the graph. It is indeed true that a topological bottleneck can worsen communication between distant nodes, especially if the bottleneck lies along a path that connects them. However, this perspective is limited in two important ways. First, a

topological bottleneck is only harmful if it lies on the information paths between nodes that are supposed to communicate. A topological bottleneck may exist without affecting task-relevant dependencies. Second, the graph topology can worsen the long-range communication even in the absence of any identifiable topological bottleneck, by inducing computational bottlenecks. As we discussed in Appendix E.5.1, the grid graph is an illustrative case: despite the lack of topological bottlenecks, to connect the opposite corner nodes we need, at least, as many message-passing layers as the distance between them, thus leading to a huge computational bottleneck that will likely hamper the ability to process long-range dependencies.

In addition, some of the techniques that reduce topological bottlenecks rely on introducing more edges or nodes into the graph, with the aim of reducing the distance between far-away nodes that should communicate. It is important to note that, although these approaches might be beneficial for the task at hand, they also worsen the computational bottleneck by adding more branches to the computational tree.

In conclusion, this highlights a deeper issue: long-range problems are not solely caused by topological bottlenecks, rather they can be understood as a form of information attenuation caused by computational bottlenecks in the message-passing mechanism, which can potentially be exacerbated by topological bottlenecks. Thus, solving a topological bottleneck is neither necessary nor sufficient to solve all long-range problems.

Message of the Section

i) "Oversquashing" is an ambiguous term that led to unclear research statements. Talking about ***computational* and *topological* bottlenecks**, instead, better defines the research scope of a contribution, since **they are two fundamentally distinct problems**.

ii) There can be computational bottlenecks but no topological ones, and vice-versa. Hence, each of these two bottlenecks, though intertwined, deserves a dedicated research effort. This also means **creating ad-hoc benchmarks for each type of bottleneck rather than relying solely on real-world tasks**, where it is not as easy to distinguish the combined effect of the two bottlenecks.

iii) Computational bottlenecks can happen in short-range as well as long-range tasks.

iv) Performance issues in long-range tasks are not solely caused by topological bottlenecks; computational bottlenecks play a role as well.

E.6 Conclusions

This chapter posits that the fast pace of advances in the graph machine learning field has generated several commonly accepted beliefs and hypotheses, rooted in the notions of over-smoothing, "oversquashing", long-range tasks, and heterophily, which are the cause of misunderstandings between researchers. Our contribution was to highlight such beliefs in plain sight, provide an explanation for their emergence, and demystify them when necessary with simple counterexamples. First, we argued that OSM may not be an actual problem and that node embeddings' separability should be preferred when looking for the root causes of

performance degradation. Then, we showed how talking about computational and topological bottlenecks resolves most, if not all, inconsistencies generated by the inflated use of the "oversquashing" term. Finally, we highlighted the role of the task in statements involving homophily, heterophily, and long-range dependencies. By providing much-needed clarifications around these aspects, we hope to foster further advancements in the graph machine learning field.

Appendix F

Tutorials on Challenges of Graph Neural Networks

F.1 Graph Rewiring: from Theory to Applications in Fairness

[24] Adrian Arnaiz-Rodriguez, Ahmed Begga, Francisco Escolano, Nuria Oliver, and Edwin Hancock. "Graph Rewiring: From Theory to Applications in Fairness". In: *Proceedings of the Learning on Graphs Conference (LoG 2022)*. **Tutorial**. Virtual Event, Dec. 2022. URL: https://ellisalicante.org/tutorials/GraphRewiring

Venue Learning on Graphs Conference (LoG 2022), virtual edition

Date 11 December 2022

Organisers Adrián Arnaiz-Rodríguez, Ahmed Begga, Francisco Escolano, Nuria Oliver, Edwin Hancock

Panelists Petar Veličković, Marinka Zitnik, Francesco Di Giovanni, Francesco Fabbri

URL https://ellisalicante.org/tutorials/GraphRewiring includes overview, video and slides.

Synopsis. The tutorial motivates graph *rewiring* as a remedy for information-flow pathologies—including over-smoothing and over-squashing—in Graph Neural Networks (GNNs). After introducing effective-resistance theory, it surveys inductive rewiring methods such as DiffWire and shows how targeted edge augmentations simultaneously boost long-range predictive performance and *structural group fairness* in social-network applications, illustrating how rewiring reduces unfairness in graphs, linking directly to the results later formalized in Chapter 3.

F.2 Graph Learning: Principles, Challenges, and Open Directions

[22] Adrian Arnaiz-Rodriguez and Ameya Velingker. "Graph Learning: Principles, Challenges, and Open Directions". In: *41st International Conference on Machine Learning (ICML 2024)*. **Tutorial**. Vienna, Austria, July 2024. URL: https://icml.cc/virtual/2024/tutorial/35233

Venue 41st International Conference on Machine Learning (ICML 2024), Vienna

Date 23 July 2024

Organisers Adrián Arnaiz-Rodríguez, Ameya Velingker

Panelists Michael Bronstein, Mikel Galkin, Bryan Perozzi, Christopher Morris

URL https://icml.cc/virtual/2024/tutorial/35233 and https://icml2024graphs.ameyavelingker.com/, include overview, video and slides.

Synopsis.
This ICML 2024 tutorial provides a comprehensive overview of *graph machine learning*, emphasizing *structural and algorithmic challenges* that emerge in practice. It is structured around three key components:

- **Principles.** The tutorial begins by revisiting the foundational concepts behind graph learning, including message passing, spectral theory, graph convolution, and attention mechanisms. It reflects on the theoretical assumptions underlying popular Graph Neural Network (GNN) models and their implications for real-world applications.
- **Challenges.** The second part delves into major issues affecting graph learning performance, such as over-smoothing, over-squashing, bottlenecks, and poor inductive generalization. It highlights how common architectural choices or data regimes often exacerbate these issues and discusses mitigation strategies and state-of-the-art contributions including rewiring strategies, positional encodings, higher-order GNNs, and geometric deep learning approaches.
- **Open Directions.** Finally, the tutorial lists open challenges on this problems, with special attention to the proper understanding of the GNN challenges and their metrics, and how tradeoffs arise at their intersection.

Throughout, the tutorial combines theory with concrete applications, offering both *formal insights* and *empirical results*, and invites a broader discussion on the future of graph learning beyond current architectural paradigms.

Parte VI

Resumen en Español

Implementación Eficaz de la IA Fiable

Chapter summary and context

Esta tesis ofrece un enfoque sociotécnico integral hacia la Inteligencia Artificial fiable (TAI), integrando perspectivas algorítmicas, centradas en el ser humano y jurídicas. Defendemos que la TAI no puede lograrse únicamente mediante mejores algoritmos, sino que también debe considerar las interacciones humanas y las restricciones legales. Nuestros hallazgos contribuyen al discurso más amplio sobre la TAI, orientando a investigadores en IA y responsables políticos sobre cómo desarrollar sistemas que alineen los principios técnicos con las necesidades sociales, garantizando que las decisiones impulsadas por IA sean responsables y acordes con los estándares éticos.

F.3 IA en la Toma de Decisiones de Alto Riesgo

La Inteligencia Artificial (IA) está transformando profundamente la forma en que se toman las decisiones. Los algoritmos de IA permiten analizar grandes cantidades de datos de manera rápida y eficiente, ofreciendo soluciones optimizadas a problemas complejos. Esta capacidad ha convertido a la IA en una herramienta clave para organizaciones públicas y privadas que buscan mejorar la eficiencia, reducir costes y personalizar servicios.

La capacidad de los algoritmos de IA para encontrar patrones complejos y gestionar grandes volúmenes de datos hace que los sistemas basados en IA sean una opción ideal para abordar problemas del mundo real, incluso en escenarios críticos y de alto riesgo social, donde el impacto de las decisiones sobre la vida de las personas puede ser profundo, como en sanidad, empleo, justicia, educación, finanzas, seguridad, inmigración o la exposición a la información en redes sociales y medios de comunicación.

Estos casos de uso críticos no solo han sido identificados por académicos [41] o actores relevantes [202, 292], sino que también han sido reconocidos en normativas europeas recientemente adoptadas — como la *Ley de Inteligencia Artificial* (UE AI Act) [162] y la *Ley de Servicios Digitales* (UE DSA) [158].

En escenarios de alto riesgo, el diseño, implementación, despliegue, evaluación y auditoría de los sistemas de IA deben realizarse con cautela para **minimizar los *daños*** y las posibles consecuencias negativas de su uso, con el objetivo último de alcanzar sistemas de **Inteligencia Artificial fiable** [52, 209, 307].

F.4 Inteligencia Artificial Fiable

Las preocupaciones sobre los riesgos éticos asociados a los sistemas de IA han dado lugar al desarrollo de enfoques técnicos, marcos regulatorios y directrices para garantizar su uso ético y responsable. Desde las primeras iniciativas promovidas por organizaciones internacionales [149, 388] y empresas tecnológicas [194, 292], hasta la adopción de estrategias nacionales [52, 241] y supranacionales [209], el concepto de **Inteligencia Artificial fiable** (TAI, por sus siglas en inglés) se ha establecido como el estándar para garantizar que los sistemas de IA respeten los derechos fundamentales, eviten riesgos y daños sistémicos y tengan un impacto social positivo.

En Europa, este esfuerzo se materializó en 2019 con la publicación de las Directrices sobre IA fiable del Grupo de Expertos de Alto Nivel en IA de la Comisión Europea (HLEG) [209]. Este marco sentó las bases para las regulaciones europeas posteriores sobre el uso responsable de la IA, como el AI Act de la UE y la DSA.

La Ley de Servicios Digitales (UE DSA) fue adoptada en 2022 en Europa con el objetivo de mejorar la transparencia, la rendición de cuentas y la protección de los derechos fundamentales en entornos digitales, regulando las plataformas en línea y los servicios intermediarios.

La Regulación de IA de la UE (UE AI Act) entró en vigor en 2024 como la primera regulación transversal sobre el uso de la IA a nivel mundial. Establece normas para el desarrollo, despliegue y uso de sistemas de IA basadas en sus niveles de riesgo asociados.

Actualmente, los requisitos legales reflejados en las regulaciones están siendo traducidos en prácticas aplicables. Los Códigos de Conducta para los modelos de IA de Propósito General [152] y las normativas ISO sobre IA fiable [219, 220] ayudan a cumplir con los marcos normativos proporcionando criterios técnicos para su implementación eficaz [355].

Componentes y Principios Éticos de la TAI. El Grupo de Expertos de Alto Nivel en Inteligencia Artificial de la Comisión Europea definió tres componentes fundamentales de los sistemas de IA fiable [209]:

(i) Los sistemas de IA deben ser **lícitos**, cumpliendo plenamente con toda la legislación y regulaciones pertinentes.

(ii) Los sistemas de IA deben ser **éticos**, exigiendo la adhesión a principios y valores establecidos.

(iii) Los sistemas de IA deben ser **robustos**, tanto desde una perspectiva social como técnica, dado que pueden causar daños no intencionados.

Estos componentes de la TAI se definen para preservar los derechos fundamentales mediante **principios éticos**, "que deben respetarse para garantizar que los sistemas de IA se desarrollen, desplieguen y utilicen de manera fiable" [209, p. 11]. Estos principios éticos instan a los desarrolladores a construir, desplegar y utilizar la IA de formas que (i) respeten la autonomía humana, (ii) prevengan el daño, (iii) promuevan la equidad y (iv) sean explicables.

Entre los principios éticos que sustentan la Inteligencia Artificial fiable, dos desempeñan un papel fundamental en esta tesis: la **promoción de la equidad** y la **prevención del daño**.

El principio de *prevención del daño* requiere que los sistemas de IA sean auditables y exige el desarrollo de técnicas para evitar efectos adversos, como decisiones sesgadas o automatización insegura. Específicamente, el principio de *promoción de la equidad* exige que los sistemas no perpetúen ni amplifiquen las desigualdades existentes; la discriminación basada en atributos protegidos está prohibida por la ley y se aborda mediante normas técnicas para la evaluación del sesgo [220].[42]

Requisitos y directrices clave para una implementación eficaz de la TAI. Para operacionalizar los principios de la Inteligencia Artificial fiable a lo largo del ciclo de vida de los sistemas de IA, el HLEG propuso siete **requisitos concretos** para evitar daños relacionados con la IA, a saber:

i. Supervisión humana.

ii. Robustez técnica y seguridad.

iii. Privacidad y gobernanza de los datos.

iv. Transparencia.

v. Diversidad y no discriminación.

vi. Bienestar social y medioambiental.

vii. Rendición de cuentas.

Para desarrollar eficazmente IA fiable y cumplir con los principios éticos y requisitos anteriores, HLEG propone varias **directrices clave**, incluidas la necesidad de salvaguardias tanto *técnicas*, como métricas y auditorías; como *no técnicas*, basadas en regulación o la gobernanza. Además, el informe defiende prestar especial atención a situaciones que involucren grupos vulnerables, especialmente aquellos en riesgo de exclusión, y a las asimetrías de poder o de información a la hora de desarrollar, desplegar y utilizar sistemas de IA que se adhieran a los principios éticos. Finalmente, el group de expertos recomienda implicar a las partes interesadas durante todo el ciclo de vida del sistema de IA para comunicar de forma clara y proactiva información sobre sus capacidades y limitaciones técnicas y sociales.

En conclusión, los *tres* componentes, *cuatro* principios, *siete* requisitos y directrices clave para la Inteligencia Artificial fiable, tal como los define High-Level Expert Group on AI [209], convergen en un imperativo operativo: la identificación, medición y mitigación sistemáticas de los *daños* inducidos en las esferas interrelacionadas del diseño, uso y gobernanza de los sistemas de IA.

F.5 Daños Algorítmicos

Dado que los principios éticos de la IA fiable buscan minimizar el daño potencial causado por el despliegue de sistemas de IA, es crucial comprender la naturaleza de estos daños.

[42] Por ejemplo, el Título VII de la Ley de Derechos Civiles de EE. UU. de 1964 [387], la Ley General de Igualdad de Trato de Alemania de 2006 (AGG) [228], o el Artículo 17 del Estatuto de los Trabajadores español [72] prohíben la discriminación basada en características socialmente relevantes o *atributos protegidos*, como el origen étnico o racial, el género, la orientación sexual, la religión, la discapacidad o la edad.

Basándonos en trabajos previos [84, 282, 330], agrupamos los daños sociales derivados del mal funcionamiento algorítmico en seis patrones recurrentes.

(i) Los *daños de rendimiento* surgen cuando la precisión predictiva varía entre grupos, como se observa en el ejemplo clásico del rendimiento dispar de los sistemas de reconocimiento facial según el sexo y la etnia [84].

(ii) Los *daños de asignación* aparecen cuando el algoritmo asigna incorrectamente oportunidades escasas; por ejemplo, herramientas de calificación crediticia que subestiman a mujeres o minorías étnicas, restringiendo así su acceso a financiación [181].

(iii) El *daño por estereotipia* ocurre cuando las predicciones del sistema están sistemáticamente sesgadas y refuerzan estereotipos culturales, como en modelos de lenguaje que asocian "enfermera" con conceptos femeninos e "ingeniero" con masculinos [74].

(iv) Los *daños de denigración* afectan a individuos cuando las clasificaciones erróneas violan el respeto básico, como en el incidente en el que un servicio de etiquetado de imágenes clasificó erróneamente a un grupo de personas como animales.[43]

(v) Los *daños de representación* implican una sobre o infrarepresentación sistemática, como en el caso de creadores de contenido en línea que son menos recomendados por las plataformas, lo que reduce su alcance e ingresos [103, 138, 347, 402].

(vi) Los *daños procedimentales* se consideran cuando los procesos de toma de decisiones violan normas socialmente aceptadas de justicia procesal, como en el caso de algoritmos que ignoran la posibilidad de apelación o explicación en decisiones automatizadas.

Además, es fundamental reconocer el concepto de *daño* es un constructo social y puede abarcar daños intangibles a entornos sociales, culturales y políticos. Esta naturaleza social de los daños puede dar lugar a nuevas manifestaciones que surgen con cada innovación tecnológica o dinámica social. Por ejemplo, el uso de un algoritmo por parte de humanos puede generar diferentes daños, como el exceso de confianza en las decisiones de los modelos de IA o la reticencia a utilizar un sistema de IA si no se explica con claridad la justificación de sus decisiones [192].

De forma más general, un daño resultante de esta la interacción humano-IA puede definirse como situaciones en las que una herramienta de apoyo a la toma de decisiones provoca que un experto humano tome una decisión peor de la que habría tomado de forma independiente. La disminución del rendimiento del sistema infringe directamente el componente y requisito de robustez técnica del informe HLEG TAI [209], que está estrechamente vinculado al principio TAI de prevención de daños. En este contexto, el informe enfatiza que una alta precisión y confiabilidad son cruciales cuando los sistemas de IA afectan directamente las vidas humanas.

En conclusión, los daños sociales derivados de los sistemas de IA pueden provenir de diversas fuentes, incluyendo fallos algorítmicos, resultados inesperados y un uso problemático por parte del humano. Las manifestaciones específicas de estos daños dependen de la tarea y el contexto social concretos.

Comprender las diversas manifestaciones del daño algorítmico es esencial para desarrollar una IA fiable, pero el verdadero desafío radica en hacer de la IA fiable una realidad en la

[43] https://www.bbc.com/news/technology-33347866 (Consultado en mayo de 2025)

práctica [209, Cap. 2], es decir, prevenir y mitigar eficazmente estos daños en el complejo panorama sociotécnico en el que se implementan los sistemas de IA.

Sin embargo, la implementación eficaz de sistemas de IA fiables que puedan evitar resultados perjudiciales, especialmente en contextos de alto riesgo, sigue siendo un desafío. Esto se debe a, por un lado, las limitaciones técnicas de los algoritmos y, por otro lado, a la continua alineación entre los principios éticos, las decisiones de diseño del sistema, el comportamiento humano y las restricciones legales.

F.6 Desafíos y Preguntas de Investigación

¿Alguna persona (o grupo de personas) se vería perjudicada si se desplegara un sistema de IA?

A pesar del auge de propuestas académicas y nuevas legislaciones, una implementación fiable de los sistemas de IA sigue siendo la excepción más que la norma en los despliegues reales [27, 173, 287]. El principal obstáculo radica en la naturaleza *sociotécnica* de la IA: los daños no intencionados pueden surgir en múltiples niveles interrelacionados. Siguiendo los componentes, principios y requisitos clave de la IA fiable propuestos por el HLEG (Apéndice F.4), tratamos un sistema de IA como un *sistema sociotécnico* que opera en **tres esferas**:

i. **Diseño técnico**, relativo a los algoritmos y los datos;

ii. **Uso**, relacionado con la interacción humano-IA y el uso general de los algoritmos;

iii. **Gobernanza**, que abarca la regulación, las normas y la rendición de cuentas de los sistemas de IA.

Estas tres esferas deben funcionar conjuntamente para implementar la IA fiable de forma práctica. La Figura 63 ilustra las tres esferas. Cada una plantea desafíos internos, y también surgen tensiones en sus intersecciones.

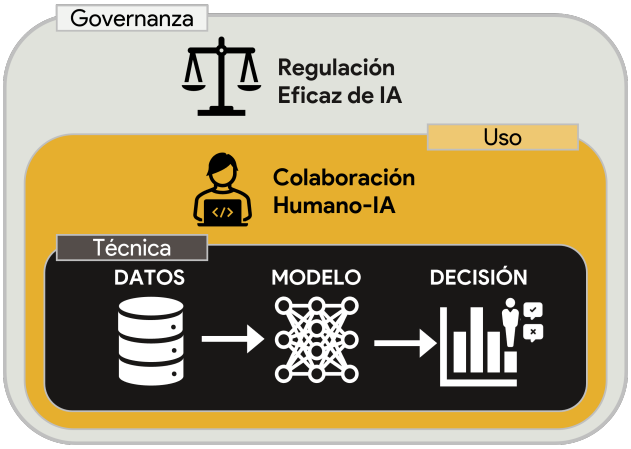


Figura 63. Mitigación de daños desde tres perspectivas: **Técnica**, **Uso** y **Gobernanza**.

Dentro de la esfera de *diseño técnico*, los daños suelen manifestarse como resultados discriminatorios derivados de sesgos algorítmicos [294]. En la esfera de *uso*, pueden surgir consecuencias no deseadas adicionales, como que un sistema colaborativo rinda peor que sus

componentes humanos o algorítmicos por separado, generando daños derivados de la automatización total [366]. En la esfera de *gobernanza*, las nuevas regulaciones pueden contradecir normas existentes o ser difíciles de armonizar [3, 18].

Además, surgen tensiones entre las esferas técnica y jurídica debido a la dificultad de comprender y definir conceptos técnicos en el ámbito legal. Por ejemplo, la definición legal de un "sistema de IA" no es trivial [168], o los conceptos de "explicabilidad" e "interpretabilidad" tienen significados distintos en ambos dominios [312]. Por último, las capacidades de la IA pueden superar los marcos legales existentes, y las definiciones regulatorias pueden no alinearse con las construcciones técnicas [321].

Por lo tanto, cerrar la brecha entre principios y práctica en el desarrollo de sistemas de IA fiable presenta desafíos interconectados que abarcan todo su ciclo de vida sociotécnico: desde un diseño técnico sólido hasta procesos de gobernanza inclusivos y normativos.

Objetivo de la Tesis. Esta tesis realiza varias contribuciones en el contexto de la IA fiable con un objetivo común:

Mitigar los daños provocados por la IA y hacer que la IA confiable sea aplicable en las tres esferas sociotécnicas.

Esta tesis propone estrategias de mitigación para algunos de los daños descritos previamente en las tres esferas. Desde una perspectiva algorítmica, nos centramos en mitigar el daño causado por el sesgo y la discriminación. En la literatura sobre aprendizaje automático, se han propuesto métodos de preprocesamiento, procesamiento interno y posprocesamiento para reducir la discriminación algorítmica. Desde una perspectiva sociotécnica, estos métodos también deben ser técnicamente robustos, interpretables por humanos y estar alineados con la normativa vigente.

Desde una perspectiva de colaboración entre humanos e IA, contribuimos proponiendo un método novedoso diseñado específicamente para la complementariedad entre humanos e IA, que crea un sólido proceso de toma de decisiones, a la vez que se alinea con la normativa vigente.

Finalmente, desde la perspectiva de la gobernanza, nos centramos en la regulación. Una regulación eficaz de la IA requiere la colaboración interdisciplinaria entre expertos legales y técnicos para desarrollar una comprensión compartida de cómo las regulaciones se alinean con la naturaleza técnica de los sistemas de IA y para esclarecer cómo las regulaciones específicas de IA pueden coexistir con las regulaciones existentes.

El resto de este capítulo examina las estrategias de mitigación en cada ámbito y las relaciona con las cuatro preguntas de investigación de la tesis y sus contribuciones correspondientes.

F.6.1 Esfera técnica: Justicia Algorítmica

Challenge

Desarrollar métricas técnicas y algoritmos que detecten y mitiguen diferentes tipos de sesgo y riesgos sistémicos, manteniéndose al mismo tiempo robustos, interpretables y alineados con los valores sociales.
Q1: ¿Cómo puede mejorarse la justicia algorítmica de los modelos de aprendizaje automático en la toma de decisiones de alto riesgo revelando cómo influyen los datos individuales en las métricas de justicia algoírtimca de disparidad entre grupos?
Q2: ¿Cómo podemos medir y mitigar la discriminación estructural en redes sociales?

Sesgos Algorítmicos

Un principio y requisito clave de la IA fiable es la equidad y la no discriminación, conceptos que han sido ampliamente estudiados en la toma de decisiones humanas [215]. Durante décadas, numerosos estudios han corroborado empíricamente que las decisiones humanas están sesgadas [229]. Un *sesgo* es una desviación sistemática en la toma de decisiones que favorece o perjudica injustificadamente a ciertos grupos, ya sea por factores cognitivos, sociales o estructurales.

En consecuencia, la mayoría de los países del mundo occidental han implementado leyes contra la discriminación y de igualdad de oportunidades para evitar decisiones discriminatorias. Por ejemplo, el Título VII de la Ley de Derechos Civiles de EE. UU. de 1964 [387], la Ley General de Igualdad de Trato de Alemania de 2006 (AGG) [228], o el Artículo 17 del Estatuto de los Trabajadores en España [72] prohíben la discriminación basada en características socialmente relevantes o *atributos protegidos*, como el origen étnico o racial, el género, la orientación sexual, la religión, la discapacidad o la edad. Un atributo protegido es, por tanto, una característica personal que no puede ser utilizada como base para decisiones discriminatorias en contextos de alto impacto, incluyendo el empleo, la educación, la vivienda o el acceso a bienes y servicios.

Surgen nuevos desafíos a medida que las decisiones en escenarios de alto riesgo se delegan cada vez más en algoritmos basados en IA. Numerosos estudios demuestran cómo los modelos de IA pueden perpetuar o amplificar sesgos sociales existentes, impactando de manera desproporcionada a ciertos grupos sociales, particularmente a los más vulnerables [41, 212]. Este fenómeno, conocido como *sesgo algorítmico*, consiste en la introducción o exacerbación de desigualdades en decisiones automatizadas debido a diversas causas, como sesgos históricos en los datos de entrenamiento, una elección inapropiada del modelo o algoritmo para el problema, o una interpretación sesgada de los resultados [27, 48, 288, 428].

Como se mencionó en el Apéndice F.5, el sesgo algorítmico puede generar diferentes daños sociales dependiendo del caso de uso. Estos daños incluyen la reproducción de desigualdades estructurales y la limitación del acceso de ciertos grupos a oportunidades y recursos cruciales, como la educación, el empleo, el crédito o la información relevante [238].

Aunque la literatura recoge una amplia variedad de definiciones de justicia algorítmica basadas en distintas nociones éticas y matemáticas, así como en diferentes contextos de aplicación [93, 96, 215, 288, 393], nuestro enfoque se centra en las *métricas estadísticas de justicia algorítmica grupal* [94, 393], tanto para la toma de decisiones de alto riesgo [41] como para el sesgo estructural en redes sociales [132]. La razón principal para usar medidas

estadísticas basadas en disparidades grupales es que son predominantes en las prácticas de cumplimiento normativo, se alinean con nociones éticas de equifdad y justicia [94], se alinean con las leyes nacionales contra la discriminación [72, 387], y sustentan las disposiciones sobre alto riesgo de la Regulación Europea de IA [162]. Su adopción se ve incluso consolidada por estándares internacionales [219, 220].

En consecuencia, esta tesis busca profundizar y operacionalizar estas métricas en dos contextos regulatorios:

i. **Toma de decisiones de alto riesgo (UE AI Act).** Casos canónicos incluyen herramientas de calificación crediticia que penalizan a grupos desfavorecidos [181], como el modelo de reincidencia COMPAS que sobreestima el riesgo para personas negras [16] o sistemas automatizados de cribado de candidaturas laborales que degradan a las solicitantes mujeres. Estos ejemplos violarían actualmente las leyes nacionales contra la discriminación o el mandato de no discriminación del AI Act de la UE para sistemas de alto riesgo. Aunque existen numerosas métricas de justicia algorítmica para estos procesos de toma de decisiones, la comprensión sobre cómo los datos individuales de entrenamiento influyen en las disparidades entre grupos es limitada. Este efecto limita nuestra capacidad de entender el origen de la discriminación y de obtener conclusiones que puedan llevarse a la práctica. Abordamos esta brecha mediante un enfoque de valorización de datos que calcula la influencia de cada dato de entrenamiento en las métricas de justicia algorítmica grupal de toma de decisiones (**Q1**; véase el Capítulo 2).

ii. **Plataformas sociales en línea (UE DSA).** En redes sociales, el uso de sistemas de recomendación de contenido y conexiones puede conducir a efectos como el refuerzo de popularidad, cámaras de eco, polarización política, marginación de minorías y exposición desigual [44, 103, 138, 163-165, 211, 402]. Estos riesgos se enmarcan dentro del ámbito de la DSA de la UE, que tiene como objetivo la implementación de monitorizar y mitigar de los "riesgos sistémicos" en las plataformas en línea. Estas obligaciones ya están vigentes para las grandes plataformas en línea.[44] Sin embargo, no existen métricas estadísticas estandarizadas para algunos de estos riesgos sistémicos, como el acceso y la exposición desigual a la información. Llenamos este vacío proponiendo medidas de injusticia grupal estructural (*Structural Group Unfairness*) y una estrategia de modificación de la red que reduce las disparidades en el acceso y la exposición a la información (**Q2**; véase el Capítulo 3).

Al centrarse en estos dos ámbitos definidos legalmente, esta tesis vincula los avances técnicos directamente con los entornos regulatorios donde actualmente estos avances técnicos se requieren con urgencia.

Mitigación de Sesgos Algorítmicos

Desde una perspectiva algorítmica, la mitigación del sesgo en modelos de aprendizaje automático puede abordarse en tres niveles: preprocesamiento, modificando los datos de entrada antes del aprendizaje; procesamiento interno, ajustando la función de optimización del modelo para incluir un término de justicia algorítmica; y postprocesamiento, corrigiendo las

[44] https://digital-strategy.ec.europa.eu/en/policies/dsa-enforcement (consultado en mayo de 2025)

predicciones tras la inferencia para mitigar sesgos en ellas. Dependiendo de las restricciones del sistema, puede elegirse o combinarse alguno de estos enfoques para mejorar la justicia en la toma de decisiones [206, 231, 417].

Antes del entrenamiento del modelo, las técnicas de preprocesamiento buscan transformar la distribución de los datos para reducir la discriminación subyacente. Este enfoque es aplicable cuando los datos son accesibles y pueden modificarse sin restricciones. Entre las estrategias comunes se encuentran el remuestreo para equilibrar la distribución de los grupos protegidos [231], el reponderado de instancias en la función de pérdida [21], y el aprendizaje de representaciones justas [417], que transforma las características originales en un espacio latente donde se minimiza el atributo protegido, preservando la información relevante para la tarea.

Los métodos de procesamiento interno mitigan los sesgos durante el proceso de optimización del modelo, ajustando la función de coste o imponiendo restricciones que reduzcan disparidades en las predicciones. Este enfoque es útil cuando se tiene control sobre el entrenamiento del modelo, pero no sobre los datos de entrada. Entre las técnicas más utilizadas se encuentra la optimización adversarial [254], en la que un modelo auxiliar intenta generar predicciones indistinguibles con respecto al atributo protegido, reduciendo así la información sensible que el modelo puede explotar para sesgar decisiones. Otra estrategia común es la inclusión de regularizadores de justicia algorítmica [232], que modifican la función de pérdida para minimizar las disparidades en métricas de justicia durante el entrenamiento, penalizando las diferencias en error o en la distribución de predicciones entre distintos grupos.

Las técnicas de postprocesamiento actúan sobre la salida de un modelo ya entrenado, aplicando ajustes a las predicciones sin modificar los datos ni optimizar el modelo. Estos métodos son útiles cuando el modelo debe tratarse como una caja negra, sin control sobre su entrenamiento. La técnica de postprocesamiento más común consiste en asignar umbrales de predicción diferentes entre grupos para asegurar tasas de error equivalentes [114, 206].

Nuestras contribuciones en los Capítulos 2 y 3 introducen métricas y métodos para auditar y mitigar los sesgos en procesos de toma de decisiones de alto riesgo y en el acceso y exposición a la información en redes sociales.

Primero, mejoramos la comprensión de métricas de justicia algorítmica en la toma de decisiones ampliamente utilizadas en escenarios de alto riesgo, proponiendo `FairShap`, un enfoque de valorización de datos que cuantifica la influencia de cada dato de entrenamiento en las métricas de justicia grupal (**Q1**; Capítulo 2). También proporciona recomendaciones accionables para reponderar o depurar datos con el fin de mitigar decisiones sesgadas.

En segundo lugar, introducimos nuevas métricas y algoritmos que miden y mitigan los daños de asignación y representación derivados del acceso y exposición desigual a la información en redes sociales (**Q2**; Capítulo 3).

F.6.2 Esfera de uso: Colaboración Humano-IA

Challenge

Dado que los algoritmos nunca operan en el vacío, los sistemas deben aprovechar el conocimiento contextual de los responsables humanos y evitar los riesgos de la automatización total.
Q3: ¿Cómo deben diseñarse los algoritmos para permitir una colaboración eficaz entre humanos e IA en tareas de asignación de recursos?

A pesar de la percepción de que los algoritmos de IA operan de forma completamente autónoma, en muchas aplicaciones, especialmente las de alto riesgo, las normativas exigen que las decisiones finales dependan de la intervención humana [153]. En estos casos, la IA actúa como un sistema de apoyo a la decisión, asistiendo sin sustituir completamente el control humano. Como se señaló en el Apéndice F.5, esta interacción no está exenta de daños no intencionados que pueden surgir durante la toma de decisiones conjunta.

Un número creciente de trabajos investiga cómo diseñar algoritmos robustos que tengan en cuenta explícitamente el comportamiento humano, la colaboración y la supervisión.[45] Los marcos propuestos van desde una supervisión humana ligera hasta una toma de decisiones completamente conjunta [261, 302], y abordan objetivos que van desde la explicabilidad [278] hasta el diseño de interfaces que fomentan la confianza del usuario [92].

De entre el amplio abanico de enfoques en esta esfera, nos centramos en los sistemas de *complementariedad humano-IA* [37, 364]. Estos sistemas están diseñados para lograr un rendimiento del sistema superior al de humanos o IA trabajando por separado, evitando así el riesgo de que el sistema combinado rinda peor que cualquiera de sus componentes por separado. Esto se alinea con el principio del HLEG de prevenir el daño cumpliendo con el requisito de robustez técnica [209]. En concreto, nos centramos en la complementariedad por diseño [113, 123, 365, 367], donde el sistema está diseñado *teóricamente* para garantizar que las decisiones combinadas sean siempre al menos tan precisas como las de cualquiera de los agentes por separado. Por ejemplo, para clasificar imágenes, un algoritmo puede presentar al usuario una lista de opciones cuya longitud se adapta a su nivel de habilidad [367].

En el Capítulo 4, presentamos un sistema automatizado para tareas de asignación de recursos que considera de forma inherente la complementariedad humano-IA. Esto previene consecuencias no deseadas y permite aprovechar el conocimiento contextual adicional que poseen los humanos (**Q3**).

[45] Para una revisión reciente sobre IA centrada en las personas, véase Capel y Brereton [92]

F.6.3 Esfera de gobernanza: Regulación Eficaz de la IA

Challenge

Los incentivos y salvaguardas legales fracasarán si las normas no se ajustan a las realidades técnicas de la IA. La gobernanza debe, por tanto, coevolucionar con la práctica algorítmica, y viceversa.

Q4: ¿Cuál es el grado de alineación entre las regulaciones existentes y los requisitos y realidades técnicas de una IA fiable en el entorno laboral?

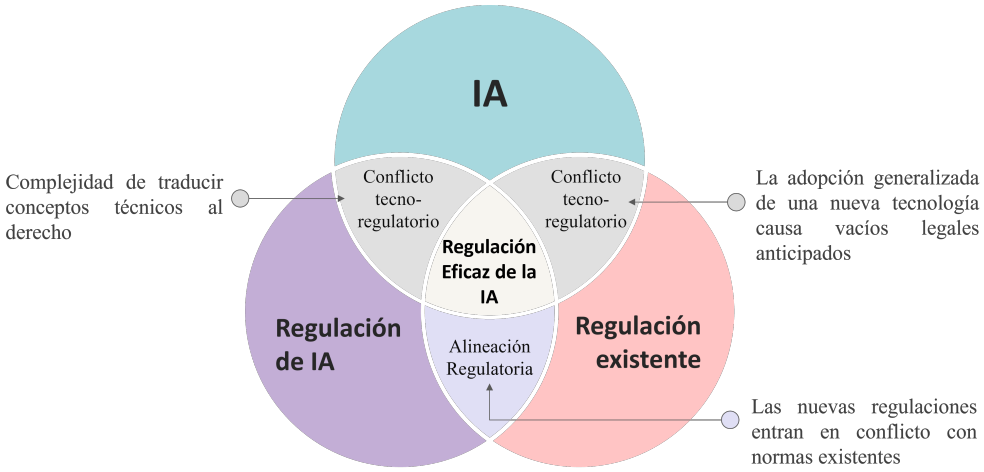


Figura 64. Una gobernanza eficaz requiere alinear los marcos regulatorios, evitar conflictos tecno-regulatorios y traducir conceptos técnicos en términos legales.

Del mismo modo que los algoritmos no operan de forma autónoma y están sujetos a la supervisión humana, su aplicación tiene lugar dentro de un marco social y regulatorio, lo que exige considerar su impacto sobre la sociedad. Prevenir los riesgos y daños asociados a la IA no puede limitarse a soluciones técnicas, sino que debe abordarse de forma holística, incluyendo regulación, auditorías y mecanismos de rendición de cuentas.

Regulaciones y estándares como el AI Act de la UE [162] buscan establecer criterios claros para garantizar que el despliegue de sistemas de IA no entre en conflicto con valores y derechos fundamentales. En paralelo, los procesos de estandarización y auditoría permiten traducir los principios regulatorios en procedimientos técnicos concretos, evaluando de forma sistemática el cumplimiento normativo de los modelos [219, 220, 355]. Por último, la rendición de cuentas es esencial para garantizar que las empresas y organizaciones asuman la responsabilidad por los impactos negativos de sus modelos sobre la sociedad y adopten medidas correctivas cuando sea necesario [150].

No obstante, aunque los marcos previos y las regulaciones existentes han abordado la mitigación de daños a través de la gobernanza, la implementación eficaz de estas soluciones regulatorias no está exenta de desafíos (véase el Figura 64). Pueden surgir inconsistencias entre los métodos técnicos y el marco legal vigente [17], contradicciones entre regulaciones específicas sobre IA y otras normativas legales existentes [18], o incluso divergencias entre conceptos técnicos de la IA y sus correspondientes regulaciones [168, 312].

Un ejemplo de este conflicto técnico-legal se da en España, donde el Estatuto de los Trabajadores [72] exige que ciertas decisiones, como los despidos, estén causalmente justificadas, mientras que otras, como la contratación, deben evitar la discriminación. Sin embargo, la mayoría de los sistemas de IA basados en aprendizaje automático operan mediante correlaciones, lo que dificulta la identificación causal, y además no están libres de sesgos que pueden desembocar en decisiones discriminatorias.

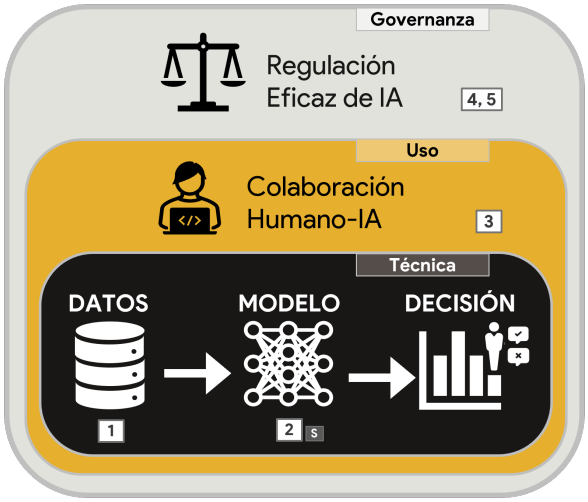
En el Capítulo 5, exploramos cómo se solapan los requisitos de la IA fiable con la regulación laboral española, y cómo los desafíos técnicos de la Inteligencia Artificial interactúan con esta regulación (**Q4**). En concreto, identificamos el marco jurídico laboral español y europeo aplicable a los sistemas de IA y analizamos las conexiones entre los requisitos de IA fiable, el AI Act de la UE y el derecho laboral español. Para completar el estudio de la intersección, nos centramos en un desalineamiento específico entre la naturaleza técnica de los sistemas de IA y el derecho laboral, analizando las implicaciones del dilema técnico entre correlación y causalidad en el contexto laboral. Finalmente, proponemos orientaciones técnicas y regulatorias para alinear las esferas técnica y normativa con el fin de mitigar este problema.

F.7 Conclusión

Esta tesis demuestra que lograr una IA fiable es inseparable del abordaje de tres esferas interconectadas: el diseño técnico, el uso humano y la gobernanza. Mediante la introducción de métodos de justicia algorítmica centrados en los datos y en los grafos, la propuesta de un sistema de complementariedad humano-IA para tareas de asignación de recursos, y el análisis del alineamiento entre la IA fiable y la legislación laboral vigente, la tesis traduce los principios éticos abstractos de justicia y prevención del daño en herramientas y directrices concretas. En conjunto, estas contribuciones demuestran que la fiabilidad de los sistemas de IA solo puede mantenerse cuando las técnicas computacionales y los requisitos legales para los sistemas de IA se desarrollan en armonía. El resultado es una hoja de ruta práctica e interdisciplinar para sistemas de IA que sean tanto efectivos como alineados con los valores sociales.

La Figura 65 sitúa los trabajos presentados en esta tesis dentro del marco holístico propuesto de IA fiable, distinguiendo las tres esferas propuestas.

Guía de lectura. Esta introducción propone un marco sociotécnico para la implementación de la IA fiable: conectando principios éticos, avances técnicos y mandatos regulatorios. Dado que los capítulos de esta tesis abordan tres esferas sociotécnicas y se apoyan en dominios técnicos diversos, cada capítulo incluye su propio contexto, notación y trabajos relacionados de forma concisa. Esta estructura autocontenida permite al lector abordar cualquier capítulo de forma independiente, mientras que esta introducción proporciona la visión unificadora a través de la cual se integran todas las contribuciones.



Trabajos Principales

[1] **DMLR @ ICLR24.** Towards Algorithmic Fairness by means of Instance-level Data Re-weighting based on Shapley Values.
A. Arnaiz-Rodriguez, N. Oliver.

[2] **ICWSM 2025.** Structural Group Unfairness: Measurement and Mitigation by means of the Effective Resistance.
A. Arnaiz-Rodriguez, Curto. G, N. Oliver.

[3] Towards Human-AI Complementarity in Matching Tasks.
A. Arnaiz-Rodriguez, N. Corvelo, S. Thejaswi, N. Oliver, M. Gomez-Rodriguez. Under Review.

[4] **RGDTSS.** The Intersection of Trustworthy AI and Labor Law. A Legal and Technical Study from a Tripartite Taxonomy.
A. Arnaiz-Rodriguez, J. Losada

[5] **RRLDE.** Studying Causality in Algorithmic Decision Making: the Impact of IA in the Business Domain
A. Arnaiz-Rodriguez, J. Losada

Trabajo complementario

[S] **LoG 2022.** DiffWire: Inductive Graph Rewiring via the Lovász Bound.
A. Arnaiz-Rodriguez, A. Begga, F. Escolano, N. Oliver

[S] **arXiv. 2025.** Oversmoothing, "Oversquashing", Heterophily, Long-Range, and more: Demystifying Common Beliefs in Graph ML.
A. Arnaiz-Rodríguez and F. Errica

[S] **ICML 2024**, ***Tutorial.*** Graph Learning: Principles, Challenges, and Open Directions.
A. Arnaiz-Rodríguez and A. Velingker

[S] **LoG 2022**, ***Tutorial.*** Graph Rewiring: from theory to Fairness.
A. Arnaiz-Rodriguez, A. Begga, F. Escolano, E. Hancock, N. Oliver

Figura 65. Trabajos relacionados con esta tesis situados en las tres esferas del sistema sociotécnico.

Bibliography

[1] Emmanuel Abbe. "Community Detection and Stochastic Block Models: Recent Developments". In: *Journal of Machine Learning Research* 18.177 (2018), pp. 1–86. URL: http://jmlr.org/papers/v18/16-480.html (cit. on pp. 150, 163).

[2] Ralph Abboud, Radoslav Dimitrov, and Ismail Ilkan Ceylan. "Shortest path networks for graph property prediction". In: *The 1st Learning on Graphs Conference (LoG)*. 2022 (cit. on p. 177).

[3] Mateo Aboy, Timo Minssen, and Effy Vayena. "Navigating the EU AI Act: implications for regulated digital medical products". In: *npj Digital Medicine* 7.1 (2024), p. 237 (cit. on pp. 7, 204).

[4] Sami Abu-El-Haija et al. "Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing". In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. 2019 (cit. on p. 177).

[5] Shipra Agrawal et al. "MNL-bandit: A dynamic learning approach to assortment selection". In: *Operations Research* 67.5 (2019), pp. 1453–1485 (cit. on p. 71).

[6] Narges Ahani et al. "Dynamic Placement in Refugee Resettlement". In: *Operations Research* (Sept. 2023) (cit. on p. 70).

[7] Narges Ahani et al. "Placement optimization in refugee resettlement". In: *Operations Research* 69.5 (2021), pp. 1468–1486 (cit. on p. 70).

[8] AI, Algorithmic and Automation Incident Repository (AIAAIC). *Xsolla employee monitoring terminations*. AIAAIC Repository, accessed 2025. 2021. URL: https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/xsolla-employee-monitoring-terminations (cit. on p. 100).

[9] Singh Akansha. "Over-squashing in graph neural networks: A comprehensive survey". In: *arXiv preprint arXiv:2308.15568* (2023) (cit. on p. 177).

[10] Morteza Alamgir and Ulrike Luxburg. "Phase transition in the family of p-resistances". In: *Advances in Neural Information Processing Systems*. 2011. URL: https://proceedings.neurips.cc/paper/2011/file/07cdfd23373b17c6b337251c22b7ea57-Paper.pdf (cit. on p. 163).

[11] Emanuele Albini et al. "Counterfactual shapley additive explanations". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 1054–1070 (cit. on pp. 27, 42).

[12] Vedat Levi Alev et al. “Graph Clustering using Effective Resistance”. In: *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Vol. 94. 2018, pp. 1–16. DOI: 10.4230/LIPIcs.ITCS.2018.41. URL: http://drops.dagstuhl.de/opus/volltexte/2018/8369 (cit. on pp. 150, 159, 167).

[13] Junaid Ali, Preethi Lahoti, and Krishna P Gummadi. “Accounting for model uncertainty in algorithmic discrimination”. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 336–345 (cit. on p. 27).

[14] Uri Alon and Eran Yahav. “On the Bottleneck of Graph Neural Networks and its Practical Implications”. In: *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. 2021. URL: https://openreview.net/forum?id=i80OPhOCVH2 (cit. on pp. 143, 144, 175, 177, 181, 188, 191).

[15] Rohan Alur et al. “Auditing for human expertise”. In: *Advances in Neural Information Processing Systems* 36 (2024) (cit. on pp. 69, 71).

[16] Julia Angwin et al. “Machine bias”. In: *Ethics of data and analytics*. Auerbach Publications, 2022, pp. 254–264 (cit. on pp. 9, 35, 118, 119, 206).

[17] Adrian Arnaiz Rodriguez and Julio Losada Carreño. “Estudio de la causalidad en la toma de decisiones algorítmicas: el impacto de la IA en el ámbito empresarial.” Spanish. In: *Revista Internacional y Comparada de Relaciones Laborales y Derecho del Empleo* 12.3 (Dec. 2024). EN: Studying Causality in Algorithmic Decision Making: the Impact of IA in the Business Domain. ISSN: 2282-2313. URL: https://ejcls.adapt.it/index.php/rlde_adapt/issue/view/105 (cit. on pp. xviii, 12, 17, 85, 209).

[18] Adrian Arnaiz Rodriguez and Julio Losada Carreño. “La intersección de la IA fiable y el Derecho del Trabajo. Un estudio jurídico y técnico desde una taxonomía tripartita”. Spanish. In: *Revista General de Derecho del Trabajo y de la Seguridad Social* 69 (2024). EN: The Intersection of Trustworthy AI and Labour Law. A Legal and Technical Study from a Tripartite Taxonomy, p. 2. ISSN: 1969-9626. URL: https://www.iustel.com/v2/revistas/detalle_revista.asp?id_noticia=427491 (cit. on pp. xviii, 7, 12, 17, 85, 90, 204, 209).

[19] Adrian Arnaiz-Rodriguez, Georgina Curto Rex, and Nuria Oliver. “Structural Group Unfairness: Measurement and Mitigation by Means of the Effective Resistance”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 19. 1. Also presented at IC2S2 2024 and TrustLOG @ WWW 2024. June 2025, pp. 83–106. DOI: 10.1609/icwsm.v19i1.35805. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/35805 (cit. on pp. xvii, 14, 15, 49).

[20] **Adrian Arnaiz-Rodriguez** and Federico Errica. “Oversmoothing, “Oversquashing”, Heterophily, Long-Range, and more: Demystifying Common Beliefs in Graph Machine Learning”. In: *22nd International Workshop on Mining and Learning with Graphs (MLG 2025) at ECML-PKDD 2025*. July 2025. URL: https://arxiv.org/abs/2505.15547 (cit. on pp. xviii, 15, 49, 108, 175).

[21] Adrian Arnaiz-Rodriguez and Nuria Oliver. “Towards Algorithmic Fairness by means of Instance-level Data Re-weighting based on Shapley Values”. In: *ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR)*. Apr. 2024. URL: https://openreview.net/forum?id=ivf1QaxEGQ (cit. on pp. xvii, 10, 13, 23, 207).

[22] Adrian Arnaiz-Rodriguez and Ameya Velingker. "Graph Learning: Principles, Challenges, and Open Directions". In: *41st International Conference on Machine Learning (ICML 2024)*. **Tutorial**. Vienna, Austria, July 2024. URL: https://icml.cc/virtual/2024/tutorial/35233 (cit. on pp. xviii, 15, 49, 183, 196).

[23] Adrian Arnaiz-Rodriguez et al. "DiffWire: Inductive Graph Rewiring via the Lovász Bound". In: *Proceedings of the First Learning on Graphs Conference*. Vol. 198. Proceedings of Machine Learning Research. PMLR, Dec. 2022, 15:1–15:27. URL: https://proceedings.mlr.press/v198/arnaiz-rodri-guez22a.html (cit. on pp. xviii, 14, 15, 49, 56, 122, 124, 143, 177, 180, 189).

[24] Adrian Arnaiz-Rodriguez et al. "Graph Rewiring: From Theory to Applications in Fairness". In: *Proceedings of the Learning on Graphs Conference (LoG 2022)*. **Tutorial**. Virtual Event, Dec. 2022. URL: https://ellisalicante.org/tutorials/GraphRewiring (cit. on pp. xviii, 15, 49, 195).

[25] **Adrian Arnaiz-Rodriguez** et al. "Towards Human-AI Complementarity in Matching Tasks". In: *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases - The Third Workshop on Hybrid Human-Machine Learning and Decision Making*. Sept. 2025. URL: https://arxiv.org/abs/2508.13285 (cit. on pp. xvii, 16, 69).

[26] Álvaro Arroyo et al. "On vanishing gradients, over-smoothing, and over-squashing in gnns: Bridging recurrent and graph learning". In: *arXiv preprint arXiv:2502.10818* (2025) (cit. on pp. 177, 185).

[27] Maria De-Arteaga, Stefan Feuerriegel, and Maytal Saar-Tsechansky. "Algorithmic fairness in business analytics: Directions for research and practice". In: *Production and Operations Management* 31.10 (2022), pp. 3749–3770 (cit. on pp. 7, 9, 203, 205).

[28] Hugo Attali, Davide Buscaldi, and Nathalie Pernelle. "Delaunay graph: Addressing over-squashing and over-smoothing using delaunay triangulation". In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*. 2024 (cit. on p. 177).

[29] Hugo Attali, Davide Buscaldi, and Nathalie Pernelle. "Rewiring Techniques to Mitigate Oversquashing and Oversmoothing in GNNs: A Survey". In: *arXiv preprint arXiv:2411.17429* (2024) (cit. on pp. 177, 189).

[30] Haris Aziz et al. "Optimal kidney exchange with immunosuppressants". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. AAAI Press, 2021, pp. 21–29 (cit. on p. 70).

[31] Davide Bacciu et al. "A Gentle Introduction to Deep Learning for Graphs". In: *Neural Networks* 129 (2020) (cit. on p. 175).

[32] Vahid Balazadeh Meresht et al. "Learning to switch among agents in a team". In: *Transactions on Machine Learning Research* 2022.7 (2022), pp. 1–30 (cit. on p. 71).

[33] Brian Ball and Mark EJ Newman. "Friendship networks and social status". In: *Network Science* 1.1 (2013), pp. 16–30 (cit. on p. 52).

[34] Julia Balla. "Over-squashing in Riemannian Graph Neural Networks". In: *Proceedings of the 2nd Learning on Graphs Conference (LoG)*. 2023 (cit. on p. 177).

[35] Pradeep Kr Banerjee et al. “Oversquashing in gnns through the lens of information contraction and graph expansion”. In: *Proceedings of the 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 2022 (cit. on pp. 177, 180, 189).

[36] Kirk Bansak et al. “Improving refugee integration through data-driven algorithmic assignment”. In: *Science* 359.6373 (2018), pp. 325–329 (cit. on p. 70).

[37] Gagan Bansal et al. “Is the most accurate ai the best teammate? optimizing ai for teamwork”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 13. 2021, pp. 11405–11414 (cit. on pp. 11, 69, 208).

[38] Albert-László Barabási. “Network science”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371.1987 (2013), p. 20120375 (cit. on p. 61).

[39] Federico Barbero et al. “Locality-Aware Graph Rewiring in GNNs”. In: *Proceedings of the 12th International Conference on Learning Representations (ICLR)*. 2024 (cit. on p. 177).

[40] Pablo Barceló et al. “The Logical Expressiveness of Graph Neural Networks”. In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=r1lZ7AEKvB (cit. on p. 144).

[41] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. The MIT Press, 2019 (cit. on pp. 3, 9, 13, 23, 25, 30, 199, 205).

[42] Solon Barocas and Andrew D Selbst. “Big data’s disparate impact”. In: *California law review* (2016), pp. 671–732 (cit. on pp. 25, 66, 101).

[43] Ainhize Barrainkua et al. “Uncertainty matters: stable conclusions under unstable assessment of fairness results”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 1198–1206 (cit. on p. 38).

[44] Ashkan Bashardoust et al. “Reducing Access Disparities in Networks using Edge Augmentation”. In: *Proceedings of the 2023 FAccT*. ACM, 2023 (cit. on pp. 9, 50, 52, 53, 58, 59, 67, 122, 206).

[45] Samyadeep Basu, Phil Pope, and Soheil Feizi. “Influence Functions in Deep Learning Are Fragile”. In: *International Conference on Learning Representations*. 2021 (cit. on p. 27).

[46] Joshua Batson et al. “Spectral Sparsification of Graphs: Theory and Algorithms”. In: *Commun. ACM* 56.8 (Aug. 2013), pp. 87–94. ISSN: 0001-0782 (cit. on p. 157).

[47] Peter W Battaglia et al. “Relational inductive biases, deep learning, and graph networks”. In: *arXiv preprint arXiv:1806.01261* (2018). URL: https://arxiv.org/abs/1806.01261 (cit. on p. 145).

[48] Elias Baumann and Josef Lorenz Rumberger. “State of the art in fair ML: from moral philosophy and legislation to fair classifiers”. In: *arXiv preprint arXiv:1811.09539* (2018) (cit. on pp. 9, 205).

[49] Gordon Baxter and Ian Sommerville. "Socio-technical systems: From design methods to systems engineering". In: *Interacting with computers* 23.1 (2011), pp. 4–17 (cit. on p. 66).

[50] Dominique Beaini et al. "Towards Foundational Models for Molecular Learning on Large-Scale Multi-Task Datasets". In: *Proceedings of the 12th International Conference on Learning Representations (ICLR)*. 2024 (cit. on p. 175).

[51] Rachel KE Bellamy et al. "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias". In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 4–1 (cit. on p. 118).

[52] Yoshua Bengio et al. *International AI Safety Report*. Tech. rep. DSIT 2025/001. 2025. URL: https://www.gov.uk/government/publications/international-ai-safety-report-2025 (cit. on pp. 3, 4, 97, 109, 199, 200).

[53] Richard Berk et al. "Fairness in criminal justice risk assessments: The state of the art". In: *Sociological Methods & Research* 50.1 (2021), pp. 3–44 (cit. on p. 33).

[54] Gregory Berkolaiko et al. "Edge connectivity and the spectral gap of combinatorial and quantum graphs". In: *Journal of Physics A: Mathematical and Theoretical* 50.36 (2017), p. 365201. URL: https://doi.org/10.1088/1751-8121/aa8125 (cit. on p. 163).

[55] Wendong Bi et al. "Make heterophilic graphs better fit gnn: A graph rewiring approach". In: *IEEE Transactions on Knowledge and Data Engineering* (2024) (cit. on p. 177).

[56] Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. "Spectral Clustering with Graph Neural Networks for Graph Pooling". In: *Proceedings of the 37th International Conference on Machine Learning*. 2020. URL: https://proceedings.mlr.press/v119/bianchi20a.html (cit. on pp. 146, 152).

[57] Ruta Binkyte et al. "Causality Is Key to Understand and Balance Multiple Goals in Trustworthy ML and Foundation Models". In: *arXiv preprint arXiv:2502.21123* (2025). URL: https://arxiv.org/abs/2502.21123 (cit. on pp. 101, 108).

[58] Emily Black and Matt Fredrikson. "Leave-One-out Unfairness". In: *ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 285–295 (cit. on p. 27).

[59] Mitchell Black et al. "Understanding oversquashing in gnns through the lens of effective resistance". In: *Proceedings of the 40th International Conference on Machine Learning (ICML)*. 2023 (cit. on pp. 54, 55, 59, 123, 124, 129, 177, 179, 181, 189).

[60] Cristian Bodnar et al. "Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in gnns". In: *Proceedings of the 36th Conference on Advances in Neural Information Processing Systems (NeurIPS)*. 2022 (cit. on p. 177).

[61] Miranda Bogen and Aaron Rieke. "Help wanted: An examination of hiring algorithms, equity, and bias". In: (2018) (cit. on p. 70).

[62] Boletín Oficial del Estado. *Constitución Española, de 27 de diciembre de 1978*. BOE-A-1978-31229. Spanish Constitution establishing the fundamental legal and political framework of Spain. Dec. 1978. URL: https://www.boe.es/buscar/act.php?id=BOE-A-1978-31229 (cit. on pp. 94, 98).

[63] Boletín Oficial del Estado. *Ley 3/2023, de 28 de febrero, de Empleo*. BOE-A-2023-5363. Spanish Employment Law (EL). Feb. 2023. URL: https://www.boe.es/eli/es/l/2023/02/28/3 (cit. on pp. 89, 90, 104).

[64] Boletín Oficial del Estado. *Ley 31/1995, de 8 de noviembre, de Prevención de Riesgos Laborales*. BOE-A-1995-24292. Spanish law on occupational health and safety (LPRL). Nov. 1995. URL: https://www.boe.es/eli/es/l/1995/11/08/31 (cit. on pp. 89, 90, 104).

[65] Boletín Oficial del Estado. *Real Decreto 1215/1997, de 18 de julio, por el que se establecen las disposiciones mínimas de seguridad y salud para la utilización por los trabajadores de los equipos de trabajo*. BOE-A-1997-17824. Royal Decree 1215/1997 on minimum safety and health requirements for the use of work equipment by workers. July 1997. URL: https://www.boe.es/buscar/act.php?id=BOE-A-1997-17824 (cit. on pp. 93, 104).

[66] Boletín Oficial del Estado. *Real Decreto 1483/2012, de 29 de octubre, por el que se aprueba el Reglamento de los procedimientos de despido colectivo y de suspensión de contratos y reducción de jornada*. BOE-A-2012-13419. Royal Decree 1483/2012 approving the regulation on collective dismissal procedures and on contract suspension and working time reduction. Oct. 2012. URL: https://www.boe.es/buscar/act.php?id=BOE-A-2012-13419 (cit. on pp. 100, 104).

[67] Boletín Oficial del Estado. *Real Decreto 1561/1995, de 21 de septiembre, sobre jornadas especiales de trabajo*. BOE-A-1995-21158. Spanish regulation on special working hours. Sept. 1995. URL: https://www.boe.es/eli/es/rd/1995/09/21/1561 (cit. on pp. 90, 104).

[68] Boletín Oficial del Estado. *Real Decreto 2001/1983, de 28 de julio, sobre regulación de la jornada de trabajo, jornadas especiales y descansos*. BOE-A-1983-20994. Spanish regulation on working time, special shifts, and rest periods. July 1983. URL: https://www.boe.es/eli/es/rd/1983/07/28/2001 (cit. on pp. 90, 104).

[69] Boletín Oficial del Estado. *Real Decreto 614/2001, de 8 de junio, sobre disposiciones mínimas para la protección de la salud y la seguridad de los trabajadores frente al riesgo eléctrico*. BOE-A-2001-12103. Royal Decree 614/2001 on minimum health and safety provisions regarding electrical risk at the workplace. June 2001. URL: https://www.boe.es/buscar/doc.php?id=BOE-A-2001-12103 (cit. on pp. 93, 104).

[70] Boletín Oficial del Estado. *Real Decreto 902/2020, de 13 de octubre, de igualdad retributiva entre mujeres y hombres*. BOE-A-2020-12166. Spanish regulation on equal pay between men and women. Oct. 2020. URL: https://www.boe.es/eli/es/rd/2020/10/13/902 (cit. on pp. 90, 94, 104).

[71] Boletín Oficial del Estado. *Real Decreto Legislativo 1/2007, de 16 de noviembre, por el que se aprueba el texto refundido de la Ley General para la Defensa de los Consumidores y Usuarios y otras leyes complementarias*. BOE-A-2007-20555. Spanish consolidated text of consumer and user protection law. Nov. 2007. URL: https://www.boe.es/eli/es/rdlg/2007/11/16/1 (cit. on pp. 90, 104).

[72] Boletín Oficial del Estado. *Real Decreto Legislativo 2/2015, de 23 de octubre, por el que se aprueba el texto refundido de la Ley del Estatuto de los Trabajadores*. Spanish. Texto refundido de la Ley del Estatuto de los Trabajadores. Accessed: 2024-11-18. Oct. 2015. URL: https://www.boe.es/buscar/act.php?id=BOE-A-2015-11430 (cit. on pp. 5, 8, 9, 12, 17, 86, 89, 90, 96, 98, 104, 201, 205, 206, 210).

[73] Boletín Oficial del Estado. *Real Decreto Legislativo 5/2000, de 4 de agosto, por el que se aprueba el texto refundido de la Ley sobre Infracciones y Sanciones en el Orden Social (TRLISOS)*. BOE-A-2000-15060. Spanish law on labor infractions and sanctions (TRLISOS). Aug. 2000. URL: https://www.boe.es/eli/es/rdlg/2000/08/04/5 (cit. on pp. 89, 99, 104).

[74] Tolga Bolukbasi et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings". In: *Advances in neural information processing systems* 29 (2016) (cit. on pp. 6, 202).

[75] Stephen P Borgatti. "Centrality and network flow". In: *Social networks* 27.1 (2005), pp. 55–71 (cit. on pp. 51, 53).

[76] Stephen P Borgatti, Candace Jones, and Martin G Everett. "Network measures of social capital". In: *Connections* 21.2 (1998), pp. 27–36 (cit. on pp. 51, 55).

[77] Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. "Survey on applications of multi-armed and contextual bandits". In: *2020 IEEE Congress on Evolutionary Computation (CEC)*. IEEE. 2020, pp. 1–8 (cit. on p. 71).

[78] Enrico Bozzo and Massimo Franceschet. "Resistance distance, closeness, and betweenness". In: *Social Networks* 35.3 (2013), pp. 460–469 (cit. on pp. 52, 54, 122, 124, 126).

[79] Ulrik Brandes and Daniel Fleischer. "Centrality measures based on current flow". In: *Annual symposium on theoretical aspects of computer science*. Springer. 2005, pp. 533–544 (cit. on pp. 52, 54, 124).

[80] Shaked Brody, Uri Alon, and Eran Yahav. "How Attentive are Graph Attention Networks?" In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=F72ximsx7C1 (cit. on p. 143).

[81] Jonathan Brophy. "Exit through the training data: A look into instance-attribution explanations and efficient data deletion in machine learning". In: *Technical report Oregon University* (2020) (cit. on p. 27).

[82] Simona Brunnerová et al. *Collective bargaining practices on Ai and algorithmic management in european services sectors*. 2024 (cit. on p. 94).

[83] Thomas Bühler and Matthias Hein. "Spectral Clustering Based on the Graph p-Laplacian". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. Montreal, Quebec, Canada: Association for Computing Machinery, 2009, pp. 81–88. ISBN: 9781605585161. DOI: 10.1145/1553374.1553385. URL: https://doi.org/10.1145/1553374.1553385 (cit. on pp. 151, 163).

[84] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 77–91 (cit. on pp. 5, 33, 40, 202).

[85] Ronald S Burt. "Structural holes and good ideas". In: *American journal of sociology* 110.2 (2004), pp. 349–399 (cit. on pp. 49, 51, 55, 58).

[86] Ronald S Burt. "The network structure of social capital". In: *Research in organizational behavior* 22 (2000), pp. 345–423 (cit. on pp. 50, 52).

[87] Ronald S Burt. "The social capital of opinion leaders". In: *The Annals of the American Academy of Political and Social Science* 566.1 (1999), pp. 37–54 (cit. on p. 58).

[88] Chen Cai and Yusu Wang. "A note on over-smoothing for graph neural networks". In: *Graph Representation Learning and Beyond Workshop, 37th International Conference on Machine Learning (ICML)*. 2020 (cit. on pp. 177–179, 182, 183).

[89] Chen Cai et al. "On the connection between mpnn and graph transformer". In: *Proceedings of the 40th International Conference on Machine Learning (ICML)*. 2023, pp. 3408–3430 (cit. on p. 177).

[90] Flavio Calmon et al. "Optimized Pre-Processing for Discrimination Prevention". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017 (cit. on pp. 28, 36, 46, 118).

[91] Shaosheng Cao, Wei Lu, and Qiongkai Xu. "Deep neural networks for learning graph representations". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. 2016. URL: https://ojs.aaai.org/index.php/AAAI/article/view/10179 (cit. on p. 143).

[92] Tara Capel and Margot Brereton. "What is human-centered about human-centered AI? A map of the research landscape". In: *Proceedings of the 2023 CHI conference on human factors in computing systems*. 2023, pp. 1–23 (cit. on pp. 11, 101, 208).

[93] Alycia N Carey and Xintao Wu. "The causal fairness field guide: Perspectives from social and formal sciences". In: *Frontiers in big Data* 5 (2022), p. 892837 (cit. on pp. 9, 205).

[94] Alycia N Carey and Xintao Wu. "The statistical fairness field guide: perspectives from social and formal sciences". In: *AI and Ethics* 3.1 (2023), pp. 1–23. URL: https://doi.org/10.1007/s43681-022-00183-3 (cit. on pp. 9, 23, 109, 205, 206).

[95] Daniele Castellana and Federico Errica. "Investigating the Interplay between Features and Structures in Graph Learning". In: *MLG Workshop at ECML PKDD*. 2023 (cit. on p. 187).

[96] Alessandro Castelnovo et al. "A clarification of the nuances in the fairness metrics landscape". In: *Scientific Reports* 12.1 (2022), p. 4209 (cit. on pp. 9, 30, 205).

[97] Simon Caton and Christian Haas. "Fairness in Machine Learning: A Survey". In: *ACM Comput. Surv.* (Aug. 2023) (cit. on p. 27).

[98] Junyi Chai and Xiaoqian Wang. "Fairness with Adaptive Weights". In: *International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 2853–2866 (cit. on pp. 26–28, 35).

[99] Ashok K Chandra et al. "The electrical resistance of a graph captures its commute and cover times". In: *Proceedings of the twenty-first annual ACM symposium on Theory of computing*. 1989, pp. 574–586 (cit. on pp. 54, 55, 122, 124, 180, 189).

[100] Mohammad-Amin Charusaie et al. “Sample efficient learning of predictors that complement humans”. In: *International Conference on Machine Learning*. 2022 (cit. on p. 71).

[101] Deli Chen et al. “Measuring and relieving the over-smoothing problem for graph neural networks from the topological view”. In: *Proceedings of the 34th AAAI conference on artificial intelligence (AAAI)*. 2020 (cit. on pp. 145, 177, 179, 183).

[102] Jianfei Chen, Jun Zhu, and Le Song. “Stochastic Training of Graph Convolutional Networks with Variance Reduction”. In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 2018 (cit. on p. 181).

[103] Jiawei Chen et al. “Bias and debias in recommender system: A survey and future directions”. In: *ACM Transactions on Information Systems* 41.3 (2023), pp. 1–39 (cit. on pp. 6, 9, 202, 206).

[104] Ming Chen et al. “Simple and deep graph convolutional networks”. In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. 2020 (cit. on p. 177).

[105] Tianlong Chen et al. “Bag of tricks for training deeper graph neural networks: A comprehensive benchmark study”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.3 (2022) (cit. on p. 177).

[106] Alexandra Chouldechova. “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”. In: *Big data* 5.2 (2017), pp. 153–163 (cit. on p. 25).

[107] Alexandra Chouldechova and Aaron Roth. “A snapshot of the frontiers of fairness in machine learning”. In: *Communications of the ACM* 63.5 (2020), pp. 82–89 (cit. on p. 24).

[108] Fan RK Chung. *Spectral graph theory*. Vol. 92. American Mathematical Soc., 1997 (cit. on pp. 50, 52, 123, 124, 147, 178, 180).

[109] Mark Coeckelbergh. *The Political Philosophy of AI*. Polity, 2022 (cit. on p. 66).

[110] Ronald R Coifman et al. “Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps”. In: *Proceedings of the NAS of the United States of America* 102.21 (2005), pp. 7426–7431 (cit. on p. 121).

[111] James S Coleman. “Social capital in the creation of human capital”. In: *American journal of sociology* 94 (1988), S95–S120 (cit. on pp. 50, 51).

[112] Weilin Cong, Morteza Ramezani, and Mehrdad Mahdavi. “On provable benefits of depth in training graph convolutional networks”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 9936–9949 (cit. on pp. 183, 185).

[113] Nina Corvelo Benz and Manuel Gomez-Rodriguez. “Human-aligned calibration for ai-assisted decision making”. In: *Advances in Neural Information Processing Systems* 36 (2024) (cit. on pp. 11, 69, 208).

[114] André F Cruz and Moritz Hardt. “Unprocessing Seven Years of Algorithmic Fairness”. In: *The Twelfth International Conference on Learning Representations*. 2024 (cit. on pp. 10, 108, 207).

[115] Sen Cui et al. “Addressing algorithmic disparity and performance inconsistency in federated learning”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 26091–26102 (cit. on p. 108).

[116] Rachel Cummings et al. “On the compatibility of privacy and fairness”. In: *Adjunct publication of the 27th conference on user modeling, adaptation and personalization*. 2019, pp. 309–315 (cit. on p. 108).

[117] Georgina Curto and Flavio Comim. “SAF: Stakeholder’s Agreement on Fairness in the Practice of Machine Learning Developent”. In: *Science and Engineering Ethics* 29 (2023) (cit. on p. 66).

[118] David Danks and Alex John London. “Algorithmic Bias in Autonomous Systems.” In: *Ijcai*. Vol. 17. 2017, pp. 4691–4697 (cit. on p. 66).

[119] Jeffrey Dastin. “Amazon scraps secret AI recruiting tool that showed bias against women”. In: *Ethics of data and analytics*. Auerbach Publications, 2022, pp. 296–299 (cit. on p. 99).

[120] J.L. Davis, A. Williams, and M.W. Yang. “Algorithmic reparation”. In: *Big Data and Society* 8.2 (2021) (cit. on pp. 66, 67).

[121] Abir De et al. “Classification Under Human Assistance”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021 (cit. on p. 71).

[122] Abir De et al. “Regression under human assistance”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020 (cit. on p. 71).

[123] Giovanni De Toni et al. “Towards Human-AI Complementarity with Prediction Sets”. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024 (cit. on pp. 11, 69, 208).

[124] Andreea Deac, Marc Lackenby, and Petar Veličković. “Expander graph propagation”. In: *Proceedings of the 1st Learning on Graphs Conference (LoG)*. 2022 (cit. on pp. 177, 189).

[125] Karel Devriendt and Renaud Lambiotte. “Discrete curvature on graphs from the effective resistance”. In: *Journal of Physics: Complexity* 3.2 (2022), p. 025008 (cit. on pp. 56, 124, 125, 127, 144, 150, 160).

[126] Francesco Di Giovanni et al. “How does over-squashing affect the power of GNNs?” In: *Transactions on Machine Learning Research* (2024) (cit. on p. 177).

[127] Francesco Di Giovanni et al. “On over-squashing in message passing neural networks: The impact of width, depth, and topology”. In: *International conference on machine learning*. PMLR. 2023, pp. 7865–7885 (cit. on pp. 123, 177, 189).

[128] Francesco Di Giovanni et al. “Understanding convolution on graphs via energies”. In: *Transactions on Machine Learning Research* (2023) (cit. on pp. 177–181, 183, 189).

[129] Aafaq Mohi ud din and Shaima Qureshi. “Limits of depth: Over-smoothing and over-squashing in GNNs”. In: *Big Data Mining and Analytics* 7 (2024) (cit. on p. 177).

[130] Frances Ding et al. “Retiring Adult: New Datasets for Fair Machine Learning”. In: *Advances in Neural Information Processing Systems* 34 (2021) (cit. on p. 42).

[131] Kaize Ding, Jundong Li, and Huan Liu. "Interactive anomaly detection on attributed networks". In: *Proceedings of the twelfth ACM international conference on web search and data mining*. 2019, pp. 357–365 (cit. on p. 71).

[132] Yushun Dong et al. "Fairness in graph mining: A survey". In: *IEEE Transactions on Knowledge and Data Engineering* (2023) (cit. on pp. 9, 14, 52, 205).

[133] Peter G Doyle and J Laurie Snell. *Random walks and electric networks*. Vol. 22. American Mathematical Soc., 1984 (cit. on pp. 54, 58, 59).

[134] Audrey Durand et al. "Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis". In: *Machine learning for healthcare conference*. PMLR. 2018, pp. 67–82 (cit. on p. 71).

[135] Vijay Prakash Dwivedi and Xavier Bresson. "A Generalization of Transformer Networks to Graphs". In: *AAAI Workshop on Deep Learning on Graphs: Methods and Applications* (2021). URL: https://arxiv.org/pdf/2012.09699.pdf (cit. on pp. 146, 167).

[136] Vijay Prakash Dwivedi et al. "Long range graph benchmark". In: *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*. 2022 (cit. on pp. 175, 177).

[137] Cynthia Dwork et al. "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012, pp. 214–226 (cit. on pp. 23, 25).

[138] Bora Edizel et al. "FaiRecSys: mitigating algorithmic bias in recommender systems". In: *International Journal of Data Science and Analytics* 9.2 (2020), pp. 197–213 (cit. on pp. 6, 9, 202, 206).

[139] El País. *150 despidos en un segundo: así funcionan los algoritmos que deciden a quién echar del trabajo*. News article in *El País*, accessed 2025. Oct. 2021. URL: https://elpais.com/icon/2021-10-10/150-despidos-en-un-segundo-asi-funcionan-los-algoritmos-que-deciden-a-quien-echar-del-trabajo.html (cit. on p. 100).

[140] Wendy Ellens et al. "Effective graph resistance". In: *Linear algebra and its applications* 435.10 (2011), pp. 2491–2506 (cit. on pp. 54, 59, 123, 124, 126, 127, 181).

[141] Bastian Epping et al. "Graph Neural Networks Do Not Always Oversmooth". In: *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*. 2024 (cit. on p. 177).

[142] Federico Errica. "On class distributions induced by nearest neighbor graphs for node classification of tabular data". In: *Proceedings of the 37th Conference on Advances in Neural Information Processing Systems (NeurIPS)*. 2023 (cit. on p. 187).

[143] Federico Errica et al. "Adaptive message passing: A general framework to mitigate oversmoothing, oversquashing, and underreaching". In: *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. 2025 (cit. on pp. 177, 189, 191).

[144] M. Estevez Almenzar et al. *Glossary of Human-Centric Artificial Intelligence*. Tech. rep. JRC129614. Luxembourg: Publications Office of the European Union, 2022. DOI: 10.2760/860665 (cit. on p. 97).

[145] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018 (cit. on p. 6).

[146] European Agency for Safety and Health at Work (EU-OSHA). *Artificial Intelligence for Worker Management: Implications for Occupational Safety and Health*. Tech. rep. EU-OSHA, 2023. DOI: 10.2802/76354. URL: https://osha.europa.eu/sites/default/files/artificial-intelligence-worker-management_en.pdf (cit. on p. 88).

[147] European Agency for Safety and Health at Work (EU-OSHA). *Worker Participation and Representation: The Impact on Risk Prevention of AI Worker Management Systems*. Tech. rep. TE-01-24-007-EN-N. EU-OSHA, 2024. DOI: 10.2802/7488542. URL: https://osha.europa.eu/sites/default/files/documents/Worker-participation-representation-impact-risk-prevention-AI-worker-management_EN.pdf (cit. on p. 88).

[148] European Commission. *Commission Staff Working Document: Impact Assessment Report. Accompanying the Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence*. Tech. rep. Brussels: European Commission, Sept. 2022, pp. 2, 8, 27 (cit. on p. 97).

[149] European Commission. *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679*. Article 29 Data Protection Working Party. 2018 (cit. on pp. 4, 200).

[150] European Commission. *Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)*. COM(2022) 496 final. Sept. 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496 (cit. on pp. 12, 90, 104, 209).

[151] European Commission. *Proposal for a Directive on liability for defective products (Revised Product Liability Directive)*. COM(2022) 495 final. Sept. 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0495 (cit. on pp. 90, 104).

[152] European Commission, EU AI Office. *General-Purpose AI Code of Practice (Third Draft)*. Published by the European Commission, DG CONNECT. Facilitated by the EU AI Office under Article 56 of the AI Act. European Union, 2025. URL: https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-published-written-independent-experts (cit. on pp. 4, 103, 109, 110, 200).

[153] European Parliament and Council of the European Union. *Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. OJ L 119, 4.5.2016, p. 1–88, May 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj (visited on 11/18/2024) (cit. on pp. 11, 16, 90, 99, 104, 208).

[154] European Parliament and Council of the European Union. *Regulation (EU) 2023/2854 on harmonised rules on fair access to and use of data (Data Act)*. Known as the Data Act. European Union, Dec. 2023. URL: https://eur-lex.europa.eu/eli/reg/2023/2854/oj (cit. on pp. 90, 104).

[155] European Parliament and Council of the European Union. *Directive (EU) 2019/1937 of the European Parliament and of the Council of 23 October 2019 on the protection of persons who report breaches of Union law*. European Union, Oct. 2019. URL: https://eur-lex.europa.eu/eli/dir/2019/1937/oj/eng (cit. on p. 94).

[156] European Parliament and Council of the European Union. *Regulation (EU) 2018/1807 on a framework for the free flow of non-personal data in the European Union*. European Union, Nov. 2018. URL: https://eur-lex.europa.eu/eli/reg/2018/1807/oj (cit. on p. 90).

[157] European Parliament and Council of the European Union. *Regulation (EU) 2022/1925 on contestable and fair markets in the digital sector (Digital Markets Act)*. Known as the Digital Markets Act (DMA). European Union, Sept. 2022. URL: https://eur-lex.europa.eu/eli/reg/2022/1925/oj (cit. on p. 90).

[158] European Parliament and Council of the European Union. *Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act)*. OJ L 277, 27.10.2022, p. 1–102. Accessed: 2024-11-18. European Union, Oct. 2022. URL: http://data.europa.eu/eli/reg/2022/2065/oj (cit. on pp. 3, 14, 67, 90, 104, 199).

[159] European Parliament and Council of the European Union. *Regulation (EU) 2022/868 on European data governance (Data Governance Act)*. Known as the Data Governance Act (DGA). European Union, May 2022. URL: https://eur-lex.europa.eu/eli/reg/2022/868/oj (cit. on pp. 90, 93, 104).

[160] European Parliament and Council of the European Union. *Regulation (EU) 2023/1230 on machinery products*. Known as the Machinery Regulation. European Union, 2023. URL: https://eur-lex.europa.eu/eli/reg/2023/1230/oj (cit. on pp. 90, 104).

[161] European Parliament and Council of the European Union. *Regulation (EU) 2023/988 on general product safety*. Known as the General Product Safety Regulation (GPSR). European Union, May 2023. URL: https://eur-lex.europa.eu/eli/reg/2023/988/oj (cit. on pp. 90, 104).

[162] European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 on Artificial Intelligence (Artificial Intelligence Act)*. OJ L 1689, 2024. Accessed: 2024-11-18. European Union, Oct. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj (cit. on pp. 3, 9, 12, 86, 90, 91, 97, 104, 199, 206, 209).

[163] Francesco Fabbri et al. "Exposure inequality in people recommender systems: The long-term effects". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16. 2022, pp. 194–204 (cit. on pp. 9, 14, 206).

[164] Francesco Fabbri et al. "Rewiring what-to-watch-next recommendations to reduce radicalization pathways". In: *Proceedings of the ACM Web Conference 2022*. 2022, pp. 2719–2728 (cit. on pp. 9, 14, 206).

[165] Francesco Fabbri et al. "The effect of homophily on disparate visibility of minorities in people recommender systems". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 165–175 (cit. on pp. 9, 14, 206).

[166] Michael Feldman et al. "Certifying and removing disparate impact". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, pp. 259–268 (cit. on pp. 24, 27, 28).

[167] Xiaoli Fern and Quintin Pope. "Text counterfactuals via latent optimization and shapley-guided search". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 5578–5593 (cit. on p. 27).

[168] David Fernández-Llorca et al. "An interdisciplinary account of the terminological choices by EU policymakers ahead of the final agreement on the AI Act: AI system, general purpose AI system, foundation model, and generative AI". In: *Artificial Intelligence and Law* (2024), pp. 1–14 (cit. on pp. 7, 12, 90, 204, 209).

[169] Lukas Fesser and Melanie Weber. "Mitigating over-smoothing and over-squashing using augmentations of forman-ricci curvature". In: *Proceedings of the 4th Learning on Graphs Conference (LoG)*. 2024 (cit. on p. 177).

[170] Matthias Fey and Jan E. Lenssen. "Fast Graph Representation Learning with PyTorch Geometric". In: *ICLR Workshop on Representation Learning on Graphs and Manifolds*. 2019 (cit. on p. 153).

[171] Benjamin Fish et al. "Gaps in information access in social networks?" In: *The World Wide Web Conference*. 2019, pp. 480–490 (cit. on pp. 50, 55).

[172] Luciano Floridi. "Soft ethics, the governance of the digital and the General Data Protection Regulation". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2133 (2018), p. 20180081 (cit. on p. 87).

[173] Luciano Floridi et al. "AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations". In: *Minds and Machines* 28 (2018), pp. 689–707. DOI: 10.1007/s11023-018-9482-5. URL: https://doi.org/10.1007/s11023-018-9482-5 (cit. on pp. 7, 203).

[174] Francois Fouss et al. "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation". In: *IEEE Transactions on knowledge and data engineering* 19.3 (2007), pp. 355–369 (cit. on pp. 52, 121, 122).

[175] Fabrizio Frasca et al. "SIGN: Scalable Inception Graph Neural Networks". In: *ICML 2020 Workshop on Graph Representation Learning and Beyond*. 2020. URL: https://grlplus.github.io/papers/77.pdf (cit. on p. 145).

[176] Nancy Fraser and Axel Honneth. *Redistribution or recognition? A political -philosophical exchange*. Verso Books, 2003 (cit. on p. 67).

[177] Linton C Freeman. "A set of measures of centrality based on betweenness". In: *Sociometry* (1977), pp. 35–41 (cit. on p. 55).

[178] Daniel Freund et al. "Group fairness in dynamic refugee assignment". In: *Proceedings of the ACM Conference on Economics and Computation*. ACM, 2023, p. 701 (cit. on p. 70).

[179] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232 (cit. on p. 35).

[180] Andrew Fuchs, Andrea Passarella, and Marco Conti. "Optimizing delegation between human and AI collaborative agents". In: *arXiv preprint arXiv:2309.14718* (2023) (cit. on p. 71).

[181] Andreas Fuster et al. “Predictably unequal? The effects of machine learning on credit markets”. In: *The Journal of Finance* 77.1 (2022), pp. 5–47 (cit. on pp. 6, 9, 202, 206).

[182] Rickard Brüel Gabrielsson, Mikhail Yurochkin, and Justin Solomon. “Rewiring with positional encodings for graph neural networks”. In: *Transactions on Machine Learning Research* (2023) (cit. on p. 177).

[183] Kiran Garimella et al. “Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship”. In: *Proceedings of the 2018 WWW conference*. 2018, pp. 913–922 (cit. on p. 50).

[184] Stacia Sherman Garr and Carole Jackson. “Diversity & inclusion technology: The rise of a transformative market”. In: *Red Thread Research and Mercer* (2019) (cit. on p. 70).

[185] Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. “Predict then Propagate: Graph Neural Networks meet Personalized PageRank”. In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. 2019 (cit. on p. 177).

[186] Timnit Gebru et al. “Datasheets for datasets”. In: *Communications of the ACM* 64.12 (2021), pp. 86–92 (cit. on p. 24).

[187] Amirata Ghorbani and James Zou. “Data shapley: Equitable valuation of data for machine learning”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2242–2251 (cit. on pp. 24, 27, 29, 33, 36, 41, 42).

[188] Arpita Ghosh, Stephen Boyd, and Amin Saberi. “Minimizing effective resistance of a graph”. In: *SIAM review* 50.1 (2008), pp. 37–66 (cit. on pp. 58, 59, 122, 124).

[189] Donald B Gillies. “Solutions to general non-zero-sum games”. In: *Contributions to the Theory of Games* 4 (1959), pp. 47–85 (cit. on p. 27).

[190] Justin Gilmer et al. “Neural Message Passing for Quantum Chemistry”. In: *Proceedings of the 34th International Conference on Machine Learning*. ICML. 2017, pp. 1263–1272 (cit. on pp. 143, 175).

[191] Jhony H Giraldo et al. “On the trade-off between over-smoothing and over-squashing in deep graph neural networks”. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (ICKM)*. 2023 (cit. on pp. 177, 180, 183).

[192] Ella Glikson and Anita Williams Woolley. “Human trust in artificial intelligence: Review of empirical research”. In: *Academy of management annals* 14.2 (2020), pp. 627–660 (cit. on pp. 6, 202).

[193] Chenghua Gong et al. “A Survey on Learning from Graphs with Heterophily: Recent Advances and Future Directions”. In: *arXiv preprint arXiv:2401.09769* (2024) (cit. on p. 177).

[194] Google PAIR. *People + AI Guidebook*. May 2019 (cit. on pp. 4, 200).

[195] Marco Gori, Gabriele Monfardini, and Franco Scarselli. “A new model for learning in graph domains”. In: *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*. 2005 (cit. on pp. 143, 175, 177).

[196] Mark Granovetter. "The strength of weak ties: A network theory revisited". In: *Sociological theory* (1983), pp. 201–233 (cit. on pp. 50, 51, 53, 58).

[197] Alessio Gravina et al. "On Oversquashing in Graph Neural Networks Through the Lens of Dynamical Systems". In: *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*. 2025 (cit. on p. 177).

[198] Limor Gultchin et al. "Beyond Impossibility: Balancing Sufficiency, Separation and Accuracy". In: *In NeurIPS Workshop on Algorithmic Fairness through the Lens of Causality and Privacy* (2022) (cit. on p. 30).

[199] Didem Gündoğdu et al. "The bridging and bonding structures of place-centric networks: Evidence from a developing country". In: *PloS one* 14.9 (2019), e0221148 (cit. on p. 50).

[200] Benjamin Gutteridge et al. "Drew: Dynamically rewired message passing with delay". In: *Proceedings of the 40th International Conference on Machine Learning (ICML)*. 2023 (cit. on p. 177).

[201] Philipp Hacker and Jan-Hendrik Passoth. "Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond". In: *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer. 2022, pp. 343–373 (cit. on pp. 24, 27).

[202] Thilo Hagendorff. "The ethics of AI ethics: An evaluation of guidelines". In: *Minds and Machines* 30.1 (2020), pp. 99–120 (cit. on pp. 3, 24, 199).

[203] Will Hamilton, Zhitao Ying, and Jure Leskovec. "Inductive Representation Learning on Large Graphs". In: *Advances in Neural Information Processing Systems*. 2017 (cit. on pp. 143, 145).

[204] William L Hamilton. *Graph representation learning*. Morgan & Claypool Publishers, 2020 (cit. on p. 177).

[205] Zayd Hammoudeh and Daniel Lowd. "Training data influence analysis and estimation: A survey". In: *Machine Learning* 113.5 (2024), pp. 2351–2403 (cit. on pp. 27, 28, 31).

[206] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning". In: *Advances in Neural Information Processing Systems* 29 (2016) (cit. on pp. 10, 23–25, 30, 36, 46, 117, 207).

[207] Arman Hasanzadeh et al. "Bayesian graph neural networks with adaptive connection sampling". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. 2020 (cit. on p. 177).

[208] Patrick Hemmer et al. "Complementarity in Human-AI Collaboration: Concept, Sources, and Evidence". In: *arXiv preprint arXiv:2404.00029* (2024) (cit. on p. 69).

[209] High-Level Expert Group on AI. *Ethics Guidelines for Trustworthy AI*. European Commission. 2019. URL: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 (cit. on pp. 3–6, 11, 86, 87, 199–203, 208).

[210] NT Hoang, Takanori Maehara, and Tsuyoshi Murata. "Revisiting graph neural networks: Graph filtering perspective". In: *25th International Conference on Pattern Recognition (ICPR)*. 2021, pp. 8376–8383. URL: https://ieeexplore.ieee.org/document/9412278 (cit. on pp. 144, 151, 162).

[211] Marilena Hohmann, Karel Devriendt, and Michele Coscia. "Quantifying ideological polarization on a network using generalized Euclidean distance". In: *Science Advances* 9.9 (2023), eabq2044. DOI: 10.1126/sciadv.abq2044. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.abq2044. URL: https://www.science.org/doi/abs/10.1126/sciadv.abq2044 (cit. on pp. 9, 14, 52, 53, 206).

[212] Benedikt Höltgen and Nuria Oliver. "Reconsidering Fairness Through Unawareness from the Perspective of Model Multiplicity". In: *arXiv preprint arXiv:2505.16638* (2025) (cit. on pp. 9, 205).

[213] Keke Huang et al. "How Universal Polynomial Bases Enhance Spectral Graph Neural Networks: Heterophily, Over-smoothing, and Over-squashing". In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*. 2024 (cit. on pp. 177, 183).

[214] Wenbing Huang et al. "Tackling over-smoothing for general graph convolutional networks". In: *arXiv preprint arXiv:2008.09864* (2020) (cit. on pp. 177, 179, 182).

[215] Ben Hutchinson and Margaret Mitchell. "50 years of test (un) fairness: Lessons for machine learning". In: *Proceedings of the conference on fairness, accountability, and transparency*. Atlanta, 2019, pp. 49–58 (cit. on pp. 8, 9, 205).

[216] EunJeong Hwang et al. "An Analysis of Virtual Nodes in Graph Neural Networks for Link Prediction (Extended Abstract)". In: *Proceedings of the 1st Learning on Graphs Conference (LoG)*. 2022 (cit. on p. 177).

[217] Kori Inkpen et al. "Advancing human-AI complementarity: The impact of user expertise and algorithmic tuning on joint decision making". In: *ACM Transactions on Computer-Human Interaction* 30.5 (2023), pp. 1–29 (cit. on p. 69).

[218] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC). *IT – AI – Artificial intelligence concepts and terminology*. Tech. rep. ISO/IEC 22989:2022. International Standard. ISO/IEC, 2022. URL: https://www.iso.org/standard/74296.html (cit. on p. 91).

[219] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC). *IT – AI – Overview of Trustworthiness in AI*. Tech. rep. ISO/IEC TR 24028:2020. Technical Report. ISO/IEC, 2020. URL: https://www.iso.org/standard/77608.html (cit. on pp. 4, 9, 12, 103, 200, 206, 209).

[220] ISO. *IT — AI — Bias in AI systems and AI aided decision making*. Tech. rep. ISO/IEC TR 24027:2021. Technical Report. International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), 2021. URL: https://www.iso.org/standard/77607.html (cit. on pp. 4, 5, 9, 12, 13, 30, 103, 108, 200, 201, 206, 209).

[221] Matthew O Jackson. *The human network: How your social position determines your power, beliefs, and behaviors*. Vintage, 2019 (cit. on pp. 51, 58).

[222] Adarsh Jamadandi, Celia Rubio-Madrigal, and Rebekka Burkholz. "Spectral Graph Pruning Against Over-Squashing and Over-Smoothing". In: *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*. 2024 (cit. on p. 177).

[223] Ruoxi Jia et al. “Efficient Task-Specific Data Valuation for Nearest Neighbor Algorithms”. In: *Proc. VLDB Endow.* 12.11 (2019), pp. 1610–1623. ISSN: 2150-8097 (cit. on pp. 31, 114, 115).

[224] Ruoxi Jia et al. “Towards efficient data valuation based on the shapley value”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 1167–1176 (cit. on p. 27).

[225] Heinrich Jiang and Ofir Nachum. “Identifying and correcting label bias in machine learning”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 702–712 (cit. on pp. 25, 27, 28, 36, 46).

[226] Kevin Fu Jiang et al. “OpenDataVal: a Unified Benchmark for Data Valuation”. In: *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*. 2023 (cit. on pp. 24, 31).

[227] Sangwon Jung et al. “Re-weighting Based Group Fairness Regularization via Classwise Robust Optimization”. In: *The Eleventh International Conference on Learning Representations*. 2023 (cit. on pp. 26, 27).

[228] Bundesministerium der Justiz und für Verbraucherschutz. *Allgemeines Gleichbehandlungsgesetz (AGG)*. Consultado el 2024-11-18. 2006. URL: https://www.gesetze-im-internet.de/agg/ (cit. on pp. 5, 8, 201, 205).

[229] D. Kahneman, O. Sibony, and C.R. Sunstein. *Noise: A Flaw in Human Judgment*. Little, Brown, 2021. ISBN: 9780316451383. URL: https://books.google.es/books?id=fhIBEAAAQBAJ (cit. on pp. 8, 205).

[230] Faisal Kamiran and Toon Calders. “Classifying without discriminating”. In: *2009 2nd international conference on computer, control and communication*. IEEE. 2009, pp. 1–6 (cit. on pp. 35, 118).

[231] Faisal Kamiran and Toon Calders. “Data preprocessing techniques for classification without discrimination”. In: *Knowledge and information systems* 33.1 (2012), pp. 1–33 (cit. on pp. 10, 24, 25, 28, 35, 46, 207).

[232] Toshihiro Kamishima et al. “Fairness-aware classifier with prejudice remover regularizer”. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*. Springer. 2012, pp. 35–50 (cit. on pp. 10, 24, 207).

[233] Jian Kang and Hanghang Tong. “N2N: Network Derivative Mining”. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. CIKM ’19. Beijing, China: Association for Computing Machinery, 2019, pp. 861–870. ISBN: 9781450369763. DOI: 10.1145/3357384.3357910. URL: https://doi.org/10.1145/3357384.3357910 (cit. on pp. 151, 161).

[234] Kedar Karhadkar, Pradeep Kr. Banerjee, and Guido Montufar. “FoSR: First-order spectral rewiring for addressing oversquashing in GNNs”. In: *Proceedings of the 11th International Conference on Learning Representations (ICLR)*. 2023 (cit. on pp. 52, 62, 124, 177, 180, 189).

[235] Kimmo Karkkainen and Jungseock Joo. "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 1548–1558 (cit. on pp. 33, 40).

[236] Leo Katz. "A new status index derived from sociometric analysis". In: *Psychometrika* 18.1 (1953), pp. 39–43 (cit. on pp. 51, 121).

[237] Anees Kazi et al. "Differentiable Graph Module (DGM) for Graph Convolutional Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), pp. 1–1. URL: https://ieeexplore.ieee.org/document/9763421 (cit. on pp. 144, 145).

[238] Michael Kearns and Aaron Roth. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, 2019 (cit. on pp. 9, 205).

[239] Katherine C Kellogg, Melissa A Valentine, and Angele Christin. "Algorithms at work: The new contested terrain of control". In: *Academy of management annals* 14.1 (2020), pp. 366–410 (cit. on p. 88).

[240] David Kempe, Jon Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network". In: *Proceedings of the ninth ACM SIGKDD*. 2003, pp. 137–146 (cit. on pp. 52, 59, 121).

[241] Erin Kenneally and David Dittrich. "The menlo report: Ethical principles guiding information and communication technology research". In: *Available at SSRN 2445102* (2012) (cit. on pp. 4, 200).

[242] Nicolas Keriven. "Not too little, not too much: a theoretical analysis of graph (over) smoothing". In: *Proceedings of the 36th Conference on Advances in Neural Information Processing Systems (NeurIPS)*. 2022 (cit. on pp. 177, 184, 185).

[243] J. King and C. Meinhardt. *Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World*. Stanford University, Human-Centered Artificial Intelligence. Feb. 2024 (cit. on p. 97).

[244] Thomas N Kipf and Max Welling. "Variational Graph Auto-Encoders". In: *In NeurIPS Workshop on Bayesian Deep Learning* (2016). URL: http://bayesiandeeplearning.org/2016/papers/BDL_16.pdf (cit. on p. 143).

[245] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *International Conference on Learning Representations (ICLR)*. 2017. URL: https://openreview.net/forum?id=SJU4ayYgl (cit. on pp. 143, 178, 190).

[246] Douglas J Klein and Milan Randić. "Resistance distance". In: *Journal of mathematical chemistry* 12 (1993), pp. 81–95 (cit. on pp. 52, 54, 56, 121, 122, 167, 180, 189).

[247] Johannes Klicpera, Stefan Weißenberger, and Stephan Günnemann. "Diffusion Improves Graph Learning". In: *Advances in Neural Information Processing Systems*. 2019. URL: https://proceedings.neurips.cc/paper/2019/file/23c894276a2c5a16470e6a3
Paper.pdf (cit. on pp. 145, 152, 166).

[248] Pang Wei Koh and Percy Liang. "Understanding black-box predictions via influence functions". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1885–1894 (cit. on p. 27).

[249] Ron Kohavi et al. "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid." In: *Kdd*. Vol. 96. 1996, pp. 202–207 (cit. on pp. 35, 118, 119).

[250] Emmanouil Krasanakis et al. "Adaptive sensitive reweighting to mitigate bias in fairness-aware classification". In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 853–862 (cit. on pp. 25, 27, 28).

[251] Anoop Krishnan, Ali Almadan, and Ajita Rattani. "Understanding fairness of gender classification algorithms across gender-race groups". In: *2020 19th IEEE international conference on machine learning and applications (ICMLA)*. IEEE. 2020, pp. 1028–1035 (cit. on p. 33).

[252] Yongchan Kwon and James Zou. "Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, May 2022, pp. 8780–8802 (cit. on p. 27).

[253] Yongchan Kwon and James Zou. "Data-oob: Out-of-bag estimate as a simple and efficient data value". In: *International conference on machine learning*. PMLR. 2023, pp. 18135–18152 (cit. on p. 42).

[254] Preethi Lahoti et al. "Fairness without Demographics through Adversarially Reweighted Learning". In: *Advances in Neural Information Processing Systems*. 2020 (cit. on pp. 10, 26–28, 108, 207).

[255] Vivian Lai et al. "Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023, pp. 1369–1385 (cit. on p. 70).

[256] Lap Chi Lau, Ramamoorthi Ravi, and Mohit Singh. *Iterative Methods in Combinatorial Optimization*. Vol. 46. Cambridge University Press, 2011 (cit. on pp. 70, 73).

[257] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444 (cit. on p. 96).

[258] Seungeon Lee et al. "Matchings, Predictions and Counterfactual Harm in Refugee Resettlement Processes". In: *arXiv preprint arXiv:2407.13052* (2024) (cit. on p. 70).

[259] Jure Leskovec and Julian Mcauley. "Learning to discover social circles in ego networks". In: *Advances in neural information processing systems* 25 (2012) (cit. on p. 60).

[260] Guohao Li et al. "DeepGCNs: Can GCNs go as deep as CNNs?" In: *Proceedings of the 17th IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019 (cit. on p. 177).

[261] Jingshu Li et al. "As Confidence Aligns: Exploring the Effect of AI Confidence on Human Self-confidence in Human-AI Decision Making". In: *arXiv preprint arXiv:2501.12868* (2025) (cit. on pp. 11, 208).

[262] Pan Li et al. "Distance Encoding: Design Provably More Powerful Neural Networks for Graph Representation Learning". In: *Advances in Neural Information Processing Systems* 33 (2020). URL: https://proceedings.neurips.cc/paper/2020/file/2f73168bf3656f697507752ec592c437-Paper.pdf (cit. on p. 146).

[263] Peizhao Li and Hongfu Liu. "Achieving Fairness at No Utility Cost via Data Reweighing with Influence". In: *International Conference on Machine Learning*. Vol. 162. PMLR, July 2022, pp. 12917–12930 (cit. on pp. 27, 28, 36, 46).

[264] Qimai Li, Zhichao Han, and Xiao-Ming Wu. "Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. 2018. URL: https://ojs.aaai.org/index.php/AAAI/article/view/11604 (cit. on p. 143).

[265] Qimai Li, Zhichao Han, and Xiao-Ming Wu. "Deeper insights into graph convolutional networks for semi-supervised learning". In: *Proceedings of the 32nd AAAI conference on artificial intelligence (AAAI)*. 2018 (cit. on pp. 177, 182).

[266] Derek Lim et al. "Large Scale Learning on Non-Homophilous Graphs: New Benchmarks and Strong Simple Methods". In: *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*. 2021 (cit. on p. 177).

[267] Derek Lim et al. "Sign and Basis Invariant Networks for Spectral Graph Representation Learning". In: *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*. 2022. URL: https://openreview.net/forum?id=Bl M64by6gc (cit. on p. 146).

[268] Yaron Lipman, Raif M Rustamov, and Thomas A Funkhouser. "Biharmonic distance". In: *ACM Transactions on Graphics (TOG)* 29.3 (2010), pp. 1–11 (cit. on p. 122).

[269] Meng Liu, Hongyang Gao, and Shuiwang Ji. "Towards deeper graph neural networks". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*. 2020, pp. 338–348 (cit. on pp. 177, 179).

[270] Yang Liu et al. "Curvdrop: A ricci curvature based approach to prevent graph neural networks from over-smoothing and over-squashing". In: *Proceedings of the ACM International World Wide Web Conference (WWWW)*. 2023 (cit. on pp. 177, 180).

[271] Yixin Liu et al. "Beyond smoothing: Unsupervised graph representation learning with edge heterophily discriminating". In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2023 (cit. on p. 177).

[272] Ziwei Liu et al. "Deep learning face attributes in the wild". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 3730–3738 (cit. on p. 40).

[273] Antonio Longa et al. "Graph Neural Networks for Temporal Graphs: State of the Art, Open Challenges, and Opportunities". In: *Transactions on Machine Learning Research* (2023) (cit. on p. 175).

[274] László Lovász. "Random walks on graphs". In: *Combinatorics, Paul erdos is eighty* 2.1-46 (1993), p. 4. URL: https://web.cs.elte.hu/~lovasz/erdos.pdf (cit. on pp. 122, 123, 144, 146, 180, 189).

[275] Sitao Luan et al. "Is heterophily a real nightmare for graph neural networks to do node classification?" In: *arXiv preprint arXiv:2109.05641* (2021) (cit. on p. 177).

[276] Sitao Luan et al. "Revisiting Heterophily For Graph Neural Networks". In: *Proceedings of the 36th Conference on Advances in Neural Information Processing Systems (NeurIPS)*. 2022 (cit. on p. 177).

[277] Sitao Luan et al. "When do graph neural networks help with node classification? investigating the homophily principle on node distinguishability". In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*. 2023 (cit. on p. 187).

[278] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems* 30 (2017) (cit. on pp. 11, 27, 208).

[279] Ulrike von Luxburg, Agnes Radl, and Matthias Hein. "Hitting and Commute Times in Large Random Neighborhood Graphs". In: *Journal of Machine Learning Research* 15.52 (2014), pp. 1751–1798. URL: http://jmlr.org/papers/v15/vonluxburg14a.html (cit. on pp. 147, 163).

[280] Yao Ma et al. "Is Homophily a Necessity for Graph Neural Networks?" In: *10th International Conference on Learning Representations (ICLR)*. 2022 (cit. on pp. 177, 186, 187).

[281] Michael Madaio et al. "Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support". In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW1 (2022), pp. 1–26 (cit. on p. 24).

[282] Michael A Madaio et al. "Co-designing checklists to understand organizational challenges and opportunities around fairness in AI". In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–14 (cit. on pp. 5, 202).

[283] Masayoshi Mase, Art B Owen, and Benjamin B Seiler. "Cohort Shapley value for algorithmic fairness". In: *arXiv preprint arXiv:2105.07168* (2021) (cit. on p. 27).

[284] Sohir Maskey et al. "A fractional graph laplacian approach to oversmoothing". In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*. 2023 (cit. on pp. 177, 178, 183).

[285] Duncan C. McElfresh et al. "Matching algorithms for blood donation". In: *Nature Machine Intelligence* 5.10 (Oct. 2023), pp. 1108–1118 (cit. on p. 70).

[286] Miller McPherson, Lynn Smith-Lovin, and James M Cook. "Birds of a feather: Homophily in social networks". In: *Annual review of sociology* 27 (2001) (cit. on p. 186).

[287] Kristof Meding. "It's complicated. The relationship of algorithmic fairness and non-discrimination regulations in the EU AI Act". In: *arXiv preprint arXiv:2501.12962* (2025) (cit. on pp. 7, 203).

[288] Ninareh Mehrabi et al. "A survey on bias and fairness in machine learning". In: *ACM computing surveys (CSUR)* 54.6 (2021), pp. 1–35 (cit. on pp. 9, 23, 25, 30, 108, 205).

[289] Diego Mesquita, Amauri Souza, and Samuel Kaski. "Rethinking pooling in graph neural networks". In: *Advances in Neural Information Processing Systems*. 2020 (cit. on p. 146).

[290] Alessio Micheli. "Neural network for graphs: A contextual constructive approach". In: *IEEE Transactions on Neural Networks* 20 (2009) (cit. on pp. 175, 177, 182).

[291] Alessio Micheli and Antonio Sestito. "A new neural network model for contextual processing of graphs". In: *Proceedings of the Italian Workshop on Neural Networks (WIRN)*. 2005 (cit. on pp. 175, 177, 178).

[292] Microsoft AI. *Microsoft Responsible AI Standard, v2*. Online; Accessed Dec. 2022. June 2022 (cit. on pp. 3, 4, 199, 200).

[293] Kanishka Misra, Eric M Schwartz, and Jacob Abernethy. "Dynamic online pricing with incomplete information using multiarmed bandit experiments". In: *Marketing Science* 38.2 (2019), pp. 226–252 (cit. on p. 71).

[294] Shira Mitchell et al. "Algorithmic Fairness: Choices, Assumptions, and Definitions". In: *Annual Review of Statistics and its Application* 8 (Mar. 2021), pp. 141–163. ISSN: 2326831X. DOI: https://doi.org/10.1146/annurev-statistics-042720-125902 (cit. on pp. 7, 203).

[295] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020 (cit. on p. 101).

[296] Christopher Morris et al. "Weisfeiler and leman go machine learning: The story so far". In: *Journal of Machine Learning Research* 24 (2023) (cit. on p. 175).

[297] Hussein Mozannar and David Sontag. "Consistent estimators for learning to defer to an expert". In: *International Conference on Machine Learning*. 2020, pp. 7076–7087 (cit. on p. 71).

[298] Hussein Mozannar et al. "Who should predict? exact algorithms for learning to defer to humans". In: *International conference on artificial intelligence and statistics*. 2023 (cit. on p. 71).

[299] Jonas W Mueller, Vasilis Syrgkanis, and Matt Taddy. "Low-rank bandit methods for high-dimensional dynamic pricing". In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 71).

[300] Janine Nahapiet and Sumantra Ghoshal. "Social capital, intellectual capital, and the organizational advantage". In: *Academy of management review* 23.2 (1998), pp. 242–266 (cit. on pp. 51, 54).

[301] Galileo Mark Namata et al. "Query-driven active surveying for collective classification". In: *Proceedings of the Workshop on Mining and Learning with Graphs (MLG)*. 2012 (cit. on p. 186).

[302] Sriraam Natarajan et al. "Human-in-the-loop or AI-in-the-loop? Automate or Collaborate?" In: *arXiv preprint arXiv:2412.14232* (2024) (cit. on pp. 11, 208).

[303] Mark EJ Newman. "A measure of betweenness centrality based on random walks". In: *Social networks* 27.1 (2005), pp. 39–54 (cit. on pp. 52, 56, 124, 125).

[304] Mark EJ Newman. "Mixing patterns in networks". In: *Physical review E* 67.2 (2003), p. 026126 (cit. on p. 51).

[305] Khang Nguyen et al. "Revisiting over-smoothing and over-squashing using ollivier-ricci curvature". In: *Proceedings of the 40th International Conference on Machine Learning (ICML)*. 2023 (cit. on pp. 177, 180).

[306] Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. "Differentiable Learning Under Triage". In: *Advances in Neural Information Processing Systems*. 2021 (cit. on p. 71).

[307] Nuria Oliver. "Artificial intelligence for social good - The way forward". In: *Science, Research and Innovation performance of the EU 2022 report*. European Commission, 2022. Chap. 11, pp. 604–707 (cit. on pp. 3, 23, 88, 199).

[308] Kenta Oono and Taiji Suzuki. "Graph Neural Networks Exponentially Lose Expressive Power for Node Classification". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=S1ldO2EFPr (cit. on pp. 144, 179, 182, 183).

[309] Organisation for Economic Co-operation and Development. *OECD AI Principles overview*. 2024. URL: https://oecd.ai/en/ai-principles (cit. on pp. 91, 92).

[310] Organisation for Economic Co-operation and Development. *Using AI in the Workplace: Opportunities and Risks for Workers*. Tech. rep. OECD Publishing, Mar. 2024. URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/using-ai-in-the-workplace_02d6890a/73d417f9-en.pdf (cit. on pp. 91, 94, 97).

[311] Lawrence Page et al. *The pagerank citation ranking: Bring order to the web*. Tech. rep. technical report, Stanford University, 1998 (cit. on pp. 51, 121).

[312] Cecilia Panigutti et al. "The role of explainable AI in the context of the AI Act". In: *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. 2023, pp. 1139–1150 (cit. on pp. 7, 12, 90, 96, 204, 209).

[313] Pál András Papp et al. "DropGNN: Random Dropouts Increase the Expressiveness of Graph Neural Networks". In: *Advances in Neural Information Processing Systems*. 2021. URL: https://openreview.net/forum?id=fpQojkIV5q8 (cit. on p. 145).

[314] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. "Deep Learning on a Data Diet: Finding Important Examples Early in Training". In: *Advances in Neural Information Processing Systems*. 2021 (cit. on p. 27).

[315] Hongbin Pei et al. "Geom-GCN: Geometric Graph Convolutional Networks". In: *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. 2020 (cit. on pp. 177, 186).

[316] Guilherme Dean Pelegrina, Miguel Couceiro, and Leonardo Tomazeli Duarte. "A preprocessing Shapley value-based approach to detect relevant and disparity prone features in machine learning". In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2024, pp. 279–289 (cit. on p. 27).

[317] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations". In: *Proceedings of the 20th ACM SIGKDD conference*. 2014, pp. 701–710 (cit. on pp. 54, 59, 62).

[318] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA: MIT Press, 2017 (cit. on pp. 96, 101).

[319] Oleg Platonov et al. "A critical look at the evaluation of GNNs under heterophily: Are we really making progress?" In: *Proceedings of the 1th International Conference on Learning Representations (ICLR)*. 2023 (cit. on p. 187).

[320] Oleg Platonov et al. “Characterizing graph datasets for node classification: Homophily-heterophily dichotomy and beyond”. In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*. 2023 (cit. on p. 177).

[321] Consejo General del Poder Judicial. *Inteligencia artificial y justicia*. Ed. by Alfonso Peralta Gutierrez and Jerónimo Pedrosa del Pino. Consejo General del Poder Judicial, 2024. DOI: FA2400201 (cit. on pp. 7, 204).

[322] Franco P Preparata and Michael I Shamos. *Computational geometry: an introduction*. Springer Science & Business Media, 2012. URL: http://www.cs.kent.edu/~dragan/CG/CG-Book.pdf (cit. on p. 152).

[323] Garima Pruthi et al. “Estimating Training Data Influence by Tracing Gradient Descent”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 19920–19930 (cit. on p. 27).

[324] Robert D Putnam. “Bowling alone: America’s declining social capital”. In: *The city reader*. Routledge, 2015, pp. 188–196 (cit. on p. 51).

[325] Huaijun Qiu and Edwin R Hancock. “Clustering and embedding using commute times”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.11 (2007), pp. 1873–1890 (cit. on pp. 124, 149, 180).

[326] Maithra Raghu et al. “The algorithmic automation problem: Prediction, triage, and human effort”. In: *arXiv preprint arXiv:1903.12220* (2019) (cit. on p. 71).

[327] Karthik Rajkumar et al. “A causal test of the strength of weak ties”. eng. In: *Science (American Association for the Advancement of Science)* 377.6612 (2022), pp. 1304–1310. ISSN: 0036-8075. DOI: 10.1126/science.abl4476 (cit. on p. 67).

[328] Ladislav Rampášek et al. “Recipe for a general, powerful, scalable graph transformer”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 14501–14515 (cit. on pp. 52, 146).

[329] Lyle Ramshaw and Robert E Tarjan. “On minimum-cost assignments in unbalanced bipartite graphs”. In: *HP Labs, Palo Alto, CA, USA, Tech. Rep. HPL-2012-40R1* 20 (2012), p. 14 (cit. on pp. 70, 73, 74).

[330] Ahmed Rashed, Abdelkrim Kallich, and Mohamed Eltayeb. “Analyzing Fairness of Computer Vision and Natural Language Processing Models”. In: *Information* 16.3 (2025), p. 182 (cit. on pp. 5, 202).

[331] J Rawls. *A Theory of Justice*. Oxford: Oxford University Press, 1971 (cit. on p. 66).

[332] Veronica Red et al. “Comparing community structure to characteristics in online collegiate social networks”. In: *SIAM review* 53.3 (2011), pp. 526–543 (cit. on p. 60).

[333] M. L. Rodríguez Fernández. “Inteligencia artificial, género y trabajo”. Spanish. In: *Temas laborales: Revista andaluza de trabajo y bienestar social* 171 (2024). In Spanish: explores gender implications of AI in the workplace, pp. 11–39. ISSN: 0213-0750. URL: https://dialnet.unirioja.es/servlet/articulo?codigo=9539778 (cit. on p. 88).

[334] Yu Rong et al. “DropEdge: Towards Deep Graph Convolutional Networks on Node Classification”. In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=Hkx1qkrKPr (cit. on pp. 144, 145, 177, 191).

[335] Alex Rosenblat et al. “Discriminating tastes: Uber’s customer ratings as vehicles for workplace discrimination”. In: *Policy & Internet* 9.3 (2017), pp. 256–279 (cit. on p. 100).

[336] Andreas Roth and Thomas Liebig. “Rank collapse causes over-smoothing and over-correlation in graph neural networks”. In: *Proceedings of the 3rd Learning on Graphs Conference (LoG)*. 2024 (cit. on pp. 177–179, 183, 185).

[337] T Konstantin Rusch, Michael M Bronstein, and Siddhartha Mishra. “A survey on oversmoothing in graph neural networks”. In: *arXiv preprint arXiv:2303.10993* (2023) (cit. on pp. 175, 177, 184).

[338] T Konstantin Rusch et al. “Graph-coupled oscillator networks”. In: *Proceedings of the 39th International Conference on Machine Learning (ICML)*. 2022 (cit. on p. 177).

[339] Haya Salah and Sharan Srinivas. “Predict, then schedule: Prescriptive analytics approach for machine learning-enabled sequential clinical scheduling”. In: *Computers & Industrial Engineering* 169 (2022), p. 108270 (cit. on p. 70).

[340] Fernando P Santos, Yphtach Lelkes, and Simon A Levin. “Link recommendation algorithms and dynamics of polarization in online social networks”. In: *Proceedings of the National Academy of Sciences* 118.50 (2021), e2102141118 (cit. on p. 53).

[341] Akrati Saxena, George Fletcher, and Mykola Pechenizkiy. “Fairsna: Algorithmic fairness in social network analysis”. In: *ACM Computing Surveys* 56.8 (2024), pp. 1–45 (cit. on pp. 52, 62).

[342] Franco Scarselli et al. “The graph neural network model”. In: *IEEE Transactions on Neural Networks* 20 (2009) (cit. on pp. 143, 175, 177).

[343] Stephanie Schoch, Haifeng Xu, and Yangfeng Ji. “CS-Shapley: Class-wise Shapley Values for Data Valuation in Classification”. In: *Advances in Neural Information Processing Systems*. 2022 (cit. on p. 27).

[344] A Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998, pp. 266–272 (cit. on p. 137).

[345] Kajetan Schweighofer et al. “The Disparate Benefits of Deep Ensembles”. In: *Proceedings of the 42nd International Conference on Machine Learning*. Vol. 143. Proceedings of Machine Learning Research. June 2025. URL: https://arxiv.org/abs/2410.13831 (cit. on p. xix).

[346] Prithviraj Sen et al. “Collective classification in network data”. In: *AI magazine* 29.3 (2008), pp. 93–93 (cit. on p. 186).

[347] Dougal Shakespeare et al. “Exploring Artist Gender Bias in Music Recommendation”. In: *The 2nd Workshop on the Impact of Recommender Systems with ACM RecSys*. 2020. URL: https://arxiv.org/pdf/2009.01715 (cit. on pp. 6, 202).

[348] Zhiqi Shao et al. “Unifying over-smoothing and over-squashing in graph neural networks: A physics informed approach and beyond”. In: *arXiv preprint arXiv:2309.02769* (2023) (cit. on p. 177).

[349] Lloyd S. Shapley. “A value for n-person games”. In: *Contributions to the Theory of Games* 2 (1953), pp. 307–317 (cit. on pp. 24, 27).

[350] Dai Shi et al. “Exposition on over-squashing problem on GNNs: Current methods, benchmarks and challenges”. In: *arXiv preprint arXiv:2311.07073* (2023) (cit. on p. 177).

[351] Antonio Todolí Signes. *Algoritmos productivos y extractivos. Cómo regular la digitalización para mejorar el empleo e incentivar la innovación*. Productive and extractive algorithms: How to regulate digitalization to improve employment and encourage innovation. Cizur Menor: Aranzadi, 2023. ISBN: 9788411631327. URL: https://dialnet.unirioja.es/servlet/libro?codigo=931336 (cit. on pp. 99, 100).

[352] Rachael Hwee Ling Sim, Xinyi Xu, and Bryan Kian Hsiang Low. “Data valuation in machine learning:“ingredients”, strategies, and open challenges”. In: *Proc. IJCAI*. 2022, pp. 5607–5614 (cit. on pp. 27, 28).

[353] Aleksandrs Slivkins. “Introduction to multi-armed bandits”. In: *Foundations and Trends® in Machine Learning* 12.1-2 (2019), pp. 1–286 (cit. on pp. 70, 74).

[354] Nathalie Smuha. “Ethics guidelines for trustworthy AI”. In: *AI & Ethics, Date: 2019/05/28-2019/05/28, Brussels, Belgium*. European Commission, 2019 (cit. on p. 23).

[355] Josep Soler Garrido et al. *AI watch: Artificial intelligence standardisation landscape update*. Tech. rep. Joint Research Centre (Seville site), 2023 (cit. on pp. 4, 12, 103, 200, 209).

[356] Joshua Southern et al. “Understanding virtual nodes: Oversquashing and node heterogeneity”. In: *International Conference on Learning Representations (ICLR)*. 2025 (cit. on p. 177).

[357] Alessandro Sperduti and Antonina Starita. “Supervised neural networks for the classification of structures”. In: *IEEE Transactions on Neural Networks* 8.3 (1997) (cit. on p. 175).

[358] Daniel A. Spielman and Nikhil Srivastava. “Graph Sparsification by Effective Resistances”. In: *SIAM Journal on Computing* 40.6 (2011), pp. 1913–1926. DOI: 10.1137/080734029. eprint: https://doi.org/10.1137/080734029. URL: https://doi.org/10.1137/080734029 (cit. on p. 148).

[359] Nick Srnicek. *Platform capitalism*. Wiley, 2016 (cit. on p. 66).

[360] Zoran Stanić. “Graphs with small spectral gap”. In: *Electronic Journal of Linear Algebra* 26 (2013), p. 28. URL: https://journals.uwyo.edu/index.php/ela/article/view/1259 (cit. on p. 163).

[361] Stevan Stanovic, Benoit Gaüzère, and Luc Brun. “Graph Neural Networks with maximal independent set-based pooling: Mitigating over-smoothing and over-squashing”. In: *Pattern Recognition Letters* 187 (2025) (cit. on p. 177).

[362] Ryan Steed and Aylin Caliskan. “Image representations learned with unsupervised pre-training contain human-like biases”. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 701–713 (cit. on p. 33).

[363] Karen Stephenson and Marvin Zelen. “Rethinking centrality: Methods and examples”. In: *Social networks* 11.1 (1989), pp. 1–37 (cit. on pp. 51, 52, 54).

[364] Mark Steyvers et al. “Bayesian modeling of human–AI complementarity”. In: *Proceedings of the National Academy of Sciences* 119.11 (2022), e2111547119 (cit. on pp. 11, 69, 208).

[365] Eleni Straitouri and Manuel Gomez-Rodriguez. “Designing decision support systems using counterfactual prediction sets”. In: *Proceedings of the 41st International Conference on Machine Learning*. 2024 (cit. on pp. 11, 69, 71, 208).

[366] Eleni Straitouri, Suhas Thejaswi, and Manuel Rodriguez. “Controlling counterfactual harm in decision support systems based on prediction sets”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 129443–129479 (cit. on pp. 7, 204).

[367] Eleni Straitouri et al. “Improving expert predictions with conformal prediction”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 32633–32653 (cit. on pp. 11, 69, 208).

[368] Eleni Straitouri et al. “Reinforcement learning under algorithmic triage”. In: *arXiv preprint arXiv:2109.11328* (2021) (cit. on p. 71).

[369] Jessica Su, Aneesh Sharma, and Sharad Goel. “The effect of recommendations on network structure”. In: *Proceedings of the International WWW Conference*. 2016, pp. 1157–1167 (cit. on p. 50).

[370] Qingyun Sun et al. “Position-aware structure learning for graph topology-imbalance by relieving under-reaching and over-squashing”. In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (ICKM)*. 2022, pp. 1848–1857 (cit. on p. 177).

[371] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic Attribution for Deep Networks”. In: *International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 3319–3328 (cit. on p. 27).

[372] Harini Suresh, Natalie Lao, and Ilaria Liccardi. “Misplaced trust: Measuring the interference of machine learning in human decision-making”. In: *Proceedings of the 12th ACM Conference on Web Science*. 2020 (cit. on p. 70).

[373] Christian Szegedy et al. “Inception-v4, inception-resnet and the impact of residual connections on learning”. In: *Thirty-first AAAI conference on artificial intelligence*. 2017 (cit. on p. 40).

[374] Simon Szreter and Michael Woolcock. “Health by association? Social capital, social theory, and the political economy of public health”. In: *International journal of epidemiology* 33.4 (2004), pp. 650–667 (cit. on p. 51).

[375] Prasanna Tambe, Peter Cappelli, and Valery Yakubovich. “Artificial intelligence in human resources management: Challenges and a path forward”. In: *California Management Review* 61.4 (2019), pp. 15–42 (cit. on p. 70).

[376] Steven L Tanimoto, Alon Itai, and Michael Rodeh. “Some matching problems for bipartite graphs”. In: *Journal of the ACM* 25.4 (1978), pp. 517–525 (cit. on pp. 70, 73).

[377] Sotharith Tauch, William Liu, and Russel Pears. “Measuring cascade effects in interdependent networks by using effective graph resistance”. In: *2015 INFOCOM Workshops*. IEEE. 2015, pp. 683–688 (cit. on p. 52).

[378] Prasad Tetali. "Random walks and the effective resistance of networks". In: *Journal of Theoretical Probability* 4 (1991), pp. 101–109 (cit. on p. 54).

[379] Fei Tian et al. "Learning deep representations for graph clustering". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2014. URL: https://ojs.aaai.org/index.php/AAAI/article/view/8916 (cit. on p. 143).

[380] Ali Tizghadam and Alberto Leon-Garcia. "A graph theoretical approach to traffic engineering and network control problem". In: *2009 21st International Teletraffic Congress*. 2009, pp. 1–8 (cit. on pp. 51, 122, 124).

[381] Ali Tizghadam and Alberto Leon-Garcia. "Betweenness centrality and resistance distance in communication networks". In: *IEEE Network* 24.6 (2010), pp. 10–16. DOI: 10.1109/MNET.2010.5634437 (cit. on pp. 52, 124).

[382] Hanghang Tong et al. "Gelling, and melting, large graphs by edge manipulation". In: *Proceedings of the 21st ACM CIKM*. 2012, pp. 245–254 (cit. on pp. 53, 58).

[383] Jake Topping et al. "Understanding over-squashing and bottlenecks on graphs via curvature". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=7UmjRGzp-A (cit. on pp. 62, 123–125, 144, 145, 150, 152, 160, 166, 175, 177, 180, 188).

[384] Floriano Tori, Vincent Holst, and Vincent Ginis. "The Effectiveness of Curvature-Based Rewiring and the Role of Hyperparameters in GNNs Revisited". In: *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. 2025 (cit. on pp. 177, 189).

[385] Domenico Tortorella and Alessio Micheli. "Leave Graphs Alone: Addressing Over-Squashing without Rewiring". In: *The 1st Learning on Graphs Conference (LoG)*. 2022 (cit. on pp. 177, 189).

[386] Stratis Tsirtsis, Manuel Gomez Rodriguez, and Tobias Gerstenberg. "Responsibility judgments in sequential human-AI collaboration". In: *Proceedings of the Annual Conference of the Cognitive Science Society*. 2024 (cit. on p. 71).

[387] U.S. Congress. *Civil Rights Act of 1964, Title VII*. Consultado el 2024-12-19. 1964. URL: https://www.govinfo.gov/content/pkg/STATUTE-78/pdf/STATUTE-78-Pg241.pdf (cit. on pp. 5, 8, 9, 201, 205, 206).

[388] UNESCO. *Recommendations on the Ethics of Artificial Intelligence*. United Nations Educational, Scientific and Cultural Organization, SHS/BIO/REC-AIETHICS/2021, Paris. 2021 (cit. on pp. 4, 200).

[389] United Nations. *International Covenant on Civil and Political Rights*. 1966 (cit. on p. 50).

[390] Petar Veličković. "Message passing all the way up". In: *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*. 2022. URL: https://openreview.net/forum?id=Bc8GiEZkTe5 (cit. on p. 144).

[391] Petar Veličković et al. "Graph Attention Networks". In: *International Conference on Learning Representations* (2018). URL: https://openreview.net/forum?id=rJXMpikCZ (cit. on p. 143).

[392] Ameya Velingker et al. “Affinity-aware graph networks”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 67847–67865 (cit. on pp. 146, 167, 169).

[393] Sahil Verma and Julia Rubin. “Fairness definitions explained”. In: *Proceedings of the international workshop on software fairness*. 2018, pp. 1–7 (cit. on pp. 9, 205).

[394] Tyler Vigen. *Spurious Correlations. Correlation is not Causation*. May 2015. URL: https://www.tylervigen.com/spurious-correlations (cit. on p. 96).

[395] Guan Wang, Charlie Xiaoqian Dang, and Ziye Zhou. “Measure contribution of participants in federated learning”. In: *2019 IEEE international conference on big data (Big Data)*. IEEE. 2019, pp. 2597–2604 (cit. on p. 27).

[396] Hongwei Wang and Jure Leskovec. “Combining graph convolutional neural networks and label propagation”. In: *ACM Transactions on Information Systems* 40.4 (2021) (cit. on p. 177).

[397] Jiachen T Wang and Ruoxi Jia. “Data banzhaf: A robust data valuation framework for machine learning”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 6388–6421 (cit. on p. 27).

[398] Jialu Wang, Xin Eric Wang, and Yang Liu. “Understanding Instance-Level Impact of Fairness Constraints”. In: *International Conference on Machine Learning*. Vol. 162. PMLR, July 2022, pp. 23114–23130 (cit. on pp. 27, 28, 30).

[399] Junfu Wang et al. “Understanding Heterophily for Graph Neural Networks”. In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*. 2024 (cit. on pp. 175, 177, 187).

[400] Keqin Wang et al. “Understanding Oversmoothing in GNNs as Consensus in Opinion Dynamics”. In: *arXiv preprint arXiv:2501.19089* (2025) (cit. on p. 177).

[401] Xindi Wang, Onur Varol, and Tina Eliassi-Rad. “Information access equality on generative models of complex networks”. In: *Applied Network Science* 7.1 (2022), pp. 1–20 (cit. on p. 52).

[402] Yifan Wang et al. “A Survey on the Fairness of Recommender Systems”. In: *ACM Trans. Inf. Syst.* 41.3 (Feb. 2023). ISSN: 1046-8188. DOI: 10.1145/3547333. URL: https://doi.org/10.1145/3547333 (cit. on pp. 6, 9, 202, 206).

[403] Michael Woolcock et al. “The place of social capital in understanding social and economic outcomes”. In: *Canadian journal of policy research* 2.1 (2001), pp. 11–17 (cit. on p. 51).

[404] Shiwen Wu et al. “Graph neural networks in recommender systems: a survey”. In: *ACM Computing Surveys* 55.5 (2022), pp. 1–37 (cit. on pp. 59, 62).

[405] Xinyi Wu et al. “A Non-Asymptotic Analysis of Oversmoothing in Graph Neural Networks”. In: *Proceedings of the 11th International Conference on Learning Representations (ICLR)*. 2023 (cit. on pp. 177, 185).

[406] Xinyi Wu et al. “Demystifying Oversmoothing in Attention-Based Graph Neural Networks”. In: *Proceedings of the 37th Conference on Advances in Neural Information Processing Systems (NeurIPS)*. 2023 (cit. on pp. 177, 183).

[407] Zhaoxuan Wu, Yao Shu, and Bryan Kian Hsiang Low. "DAVINZ: Data Valuation using Deep Neural Networks at Initialization". In: *International Conference on Machine Learning*. 2022 (cit. on p. 27).

[408] Zonghan Wu et al. "A Comprehensive Survey on Graph Neural Networks". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (2021), pp. 4–24. URL: https://ieeexplore.ieee.org/document/9046288 (cit. on pp. 143, 144).

[409] Keyulu Xu et al. "How powerful are graph neural networks?" In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. 2019 (cit. on pp. 143, 175).

[410] Keyulu Xu et al. "Representation learning on graphs with jumping knowledge networks". In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 2018 (cit. on pp. 177, 179, 189).

[411] Chaoqi Yang et al. "Revisiting over-smoothing in deep GCNs". In: *arXiv preprint arXiv:2003.13663* (2020) (cit. on pp. 178, 185).

[412] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. "Understanding the effect of accuracy on trust in machine learning models". In: *Proceedings of the 2019 chi conference on human factors in computing systems*. 2019 (cit. on p. 70).

[413] Zhitao Ying et al. "Hierarchical Graph Representation Learning with Differentiable Pooling". In: *Advances in Neural Information Processing Systems*. 2018. URL: https://proceedings.neurips.cc/paper/2018/file/e77dbaf6759253c7c6d0efc5690369c7-Paper.pdf (cit. on p. 146).

[414] Wenhang Yu et al. "Graph structure reforming framework enhanced by commute time distance for graph classification". In: *Neural Networks* 168 (2023) (cit. on p. 177).

[415] Muhammad Bilal Zafar et al. "Fairness Beyond Disparate Treatment and Disparate Impact: Learning Classification without Disparate Mistreatment". In: *International Conference on World Wide Web*. 2017, pp. 1171–1180 (cit. on pp. 23, 25, 34).

[416] Edward N. Zalta and Uri Nodelman. *Stanford Encyclopedia of Philosophy*. 2011. URL: https://plato.stanford.edu/entries/discrimination/ (cit. on p. 67).

[417] Rich Zemel et al. "Learning fair representations". In: *International Conference on Machine Learning*. PMLR. 2013, pp. 325–333 (cit. on pp. 10, 24, 207).

[418] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 335–340 (cit. on p. 24).

[419] Kaicheng Zhang et al. "Rethinking Oversmoothing in Graph Neural Networks: A Rank-Based Perspective". In: *arXiv preprint arXiv:2502.04591* (2025) (cit. on pp. 183, 184).

[420] Saijun Zhang, Steven G Anderson, and Min Zhan. "The differentiated impact of bridging and bonding social capital on economic well-being: An individual level perspective". In: *J. Soc. & Soc. Welfare* 38 (2011), p. 119 (cit. on p. 53).

[421] Xuan Zhang et al. "Artificial intelligence for science in quantum, atomistic, and continuum systems". In: *arXiv preprint arXiv:2307.08423* (2023) (cit. on p. 175).

[422] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020 (cit. on p. 70).

[423] Lingxiao Zhao and Leman Akoglu. "PairNorm: Tackling Oversmoothing in GNNs". In: *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. 2020 (cit. on pp. 177, 184, 185).

[424] Yilun Zheng, Sitao Luan, and Lihui Chen. "What is missing in homophily? disentangling graph homophily for graph neural networks". In: *arXiv preprint arXiv:2406.18854* (2024) (cit. on pp. 177, 187).

[425] Jie Zhou et al. "Graph Neural Networks: A Review of Methods and Applications". In: *CoRR* abs/1812.08434 (2018). arXiv: 1812.08434. URL: http://arxiv.org/abs/1812.08434 (cit. on p. 144).

[426] Kaixiong Zhou et al. "Dirichlet energy constrained learning for deep graph neural networks". In: *Proceedings of the 35th Conference on Advances in Neural Information Processing Systems (NeurIPS)*. 2021 (cit. on pp. 177, 179).

[427] Kaixiong Zhou et al. "Towards Deeper Graph Neural Networks with Differentiable Group Normalization". In: *Proceedings of the 34th Conference on Advances in Neural Information Processing Systems (NeurIPS)*. 2020 (cit. on pp. 177, 184).

[428] Nengfeng Zhou et al. "Bias, fairness and accountability with artificial intelligence and machine learning algorithms". In: *International Statistical Review* 90.3 (2022), pp. 468–480 (cit. on pp. 9, 205).

[429] Jiong Zhu et al. "Beyond homophily in graph neural networks: Current limitations and effective designs". In: *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*. 2020 (cit. on pp. 60, 177).

[430] Yanqiao Zhu et al. "A Survey on Graph Structure Learning: Progress and Opportunities". In: *arXiv PrePrint* (2021). URL: https://arxiv.org/abs/2103.03036 (cit. on p. 146).