Edge Formation and its Influence in Machine Learning

by

Lisette Elizabeth Munz
née Espín Noboa

Approved Dissertation thesis for the partial fulfilment of the requirements for a
Doctor of Natural Sciences (Dr. rer. nat.)
Fachbereich 4: Informatik
Universität Koblenz-Landau

Chair of PhD Board:	Prof. Dr. Ralf Lämmel
Chair of PhD Commission:	Jun.-Prof. Dr. Dennis Riehle
Examiner and Supervisor:	Prof. Dr. Matthias Thimm
Examiner and Supervisor:	Prof. Dr. Claudia Wagner
Examiner and Supervisor:	Prof. Dr. Markus Strohmaier

Date of the doctoral viva: March 25, 2022

Abstract

Social networks are ubiquitous structures that we generate and enrich every-day while connecting with people through social media platforms, emails, and any other type of interaction. While these structures are intangible to us, they carry important information. For instance, the political leaning of our friends can be a proxy to identify our own political preferences. Similarly, the credit score of our friends can be decisive in the approval or rejection of our own loans. This explanatory power is being leveraged in public policy, business decision-making and scientific research because it helps machine learning techniques to make accurate predictions. However, these generalizations often benefit the majority of people who shape the general structure of the network, and put in disadvantage under-represented groups by limiting their resources and opportunities. Therefore it is crucial to first understand how social networks form to then verify to what extent their mechanisms of edge formation contribute to reinforce social inequalities in machine learning algorithms.

To this end, in the first part of this thesis, I propose HopRank and Janus two methods to characterize the mechanisms of edge formation in real-world undirected social networks. HopRank is a model of information foraging on networks. Its key component is a biased random walker based on transition probabilities between k-hop neighborhoods. Janus is a Bayesian framework that allows to identify and rank plausible hypotheses of edge formation in cases where nodes possess additional information. In the second part of this thesis, I investigate the implications of these mechanisms—that explain edge formation in social networks—on machine learning. Specifically, I study the influence of homophily, preferential attachment, edge density, fraction of minorities, and the directionality of links on both performance and bias of collective classification, and on the visibility of minorities in top-k ranks. My findings demonstrate a strong correlation between network structure and machine learning outcomes. This suggests that systematic discrimination against certain people can be: (i) anticipated by the type of network, and (ii) mitigated by connecting strategically in the network.

Zusammenfassung

Soziale Netzwerke sind allgegenwärtige Strukturen, die wir jeden Tag generieren und bereichern, während wir uns über Plattformen der sozialen Medien, E-Mails und jede andere Art von Interaktion mit Menschen verbinden. Während diese Strukturen für uns nicht greifbar sind, sind sie sehr wichtige Informationsträger. Zum Beispiel kann die politische Neigung unserer Freunde ein Näherungswert sein, um unsere eigenen politischen Präferenzen zu identifizieren. Gleichermaßen kann die Kreditwürdigkeit unserer Freunde entscheidend bei der Gewährung oder Ablehnung unserer eigenen Kredite sein. Diese Erklärungskraft wird bei der Gesetzgebung, bei Unternehmensentscheidungen und in der Forschung genutzt, da sie maschinellen Lerntechniken hilft, genaue Vorhersagen zu treffen. Diese Verallgemeinerungen kommen jedoch häufig nur der Mehrheit der Menschen zugute, welche die allgemeine Struktur des Netzwerks prägen, und benachteiligen unterrepräsentierte Gruppen, indem sie ihre Mittel und Möglichkeiten begrenzen. Daher ist es wichtig zuerst zu verstehen, wie sich soziale Netzwerke bilden, um dann zu überprüfen, inwieweit ihre Mechanismen der Kantenbildung dazu beitragen, soziale Ungleichheiten in Algorithmen des maschinellen Lernens zu verstärken.

Zu diesem Zweck schlage ich im ersten Teil dieser Arbeit HopRank und Janus vor, zwei Methoden um die Mechanismen der Kantenbildung in realen ungerichteten sozialen Netzwerken zu charakterisieren. HopRank ist ein Modell der Daten-Hamsterei in Netzwerken. Sein Schlüsselkonzept ist ein gezinkter zufälliger Wanderer, der auf Übergangswahrscheinlichkeiten zwischen K-Hop-Nachbarschaften basiert. Janus ist ein Bayessches Rahmenwerk, mit dem wir plausible Hypothesen der Kantenbildung in Fällen identifizieren und bewerten können, in denen Knoten zusätzliche Daten enthalten. Im zweiten Teil dieser Arbeit untersuche ich die Auswirkungen dieser Mechanismen—welche die Kantenbildung in sozialen Netzwerken erklären—auf das maschinelle Lernen. Insbesondere untersuche ich den Einfluss von Homophilie, bevorzugter Bindung, Kantendichte, Anteil von Minderheiten und der Richtung von Verbindungen sowohl auf Leistung als auch auf systematische Fehler von kollektiver Klassifizierung und auf die Sichtbarkeit von Minderheiten in Top-K-Rängen. Meine Ergebnisse zeigen eine starke Korrelation zwischen der Netzwerkstruktur und den Ergebnissen des maschinellen Lernens. Dies legt nahe, dass die systematische Diskriminierung spezieller Personen: (i) durch den Netzwerktyp vorweggenommen und (ii) durch strategisches Verbinden im Netzwerk verhindert werden kann.

I dedicate this thesis to my supportive and lovely husband Reinhard, and to all new generation of female scientists yet to come.

Acknowledgments

I cannot close this chapter in my life without thanking all people who have directly and indirectly helped me along this journey. First, thanks to my family in Ecuador, without their support and encouragement I would not have been able to return to Germany to pursue a PhD. after my Masters. A special thanks to my extraordinary parents, Emperatriz and Carlos, who have always been a reference of courage, hard work, honesty, dignity and equality. To my sister Mabel, and my brother Andrés who helped me organize my church wedding in Guayaquil while I was still in Cologne. I would also like to thank my nieces Amanda and Bárbara for helping me realize a reading club with them during the lockdown. To my family in Germany, specially to Sibylle and Heinz two amazing parents in-law who have always cared about my well being and career. They made me feel very welcome since day one. To my brothers in-law Jürgen and Wilfried for helping us moving out from our apartments in Saarbrücken and Cologne. Finally, a big thanks to my husband Reinhard without his support and ability to listen to me in all phases of my PhD., I would not have been able to made it through.

Second, thanks to my supervisors and committee. During the 5 years and 8 months of my PhD. I have had the pleasure and honor to work with incredible scientists. I would like to start thanking Markus Strohmaier who has been a fantastic mentor. Thanks to his guidance and support I applied to internships at Stanford and USC to expand my scientific network and collaborations. I will always be grateful to him for his kindness, positivism, and interest to periodically catch-up with all of us to make sure we were doing well not only at GESIS but also at home. After Markus left the University of Koblenz, Claudia Wagner took over his mentorship and in 2018 we started collaborating together. Since then, I have learned a lot from her. She has taught me how to become better at writing and at defending my ideas in a discussion. I really appreciate her dedication and commitment to every project we have worked together. It has been a great pleasure to work with both Markus and Claudia, two role models of academic leadership: they work on interesting problems that have impact on society, they chair departments with enthusiasm and commitment in an environment of inclusiveness, and most important, they show by example that it is possible and necessary to balance life and work in academia. Last but not least, my infinite gratitude to Matthias Thimm who accepted being my main supervisor at the University of Koblenz after Claudia moved to RWTH Aachen.

Next, I thank Philipp Singer, Florian Lemmerich and Fariba Karimi, three outstanding postdocs whom I had the pleasure to collaborate with. During my first two years of PhD., Philipp taught me all I needed to know about Bayesian inference and HypTrails, and thanks to Florian I can now be more concise on what I write in my articles. I am grateful to Fariba for her great patience, specially on those moments where we would disagree but continue the discussion to deeply understand each others' perspectives.

The gratitude continues with my friends and colleagues from GESIS. Special thanks to Diana Lindner who was my first contact at GESIS and very quickly became one of my friends. I really enjoyed all the lunches and trips we had together. Also, Reinhard and I will always be grateful to her for being the interpreter at our civil wedding in Amorbach. She has always been very supportive and I am glad that we have spent time together. Many thanks to my office mate Marcos Oliveira, who not only joined GESIS in 2018, but also room #409, the Latino office. Although we officially never collaborated together, he has been a mentor to me. He has always had time to listen to my problems, discuss network science related ideas, and choose where to eat (before the pandemic hit). I am grateful to many researchers, interns, and staff for their help, feedback, and care including Julian Kohne, Mohsen Jadidi, Johann (Wanja) Schaible, Lisa Posch, Arnim Bleier, David Brodesser, Nora Kirkiza, Olga Zagovora, Kenan Erdogan, Anna Samoilenko, Mathieu Génois, Haiko Lietz, Sebastian Stier, Lennart Schürmann, Christoph Carl Kling (CCK), Maria Zens, Katrin Weller, Fabian Flöck, Michaela Grein, Heike Bolz, Daniel Friedrich and specially to Mattia Samori, Juhi Kulshrestha, Indira Sen, Roberto Ulloa and Dimitar (Dimi) Dimitrov for always having time to listen to my practice talks and give excellent feedback.

Finally, I would like to thank Kristina Lerman, Mark Musen and Simon Walk for hosting me as a visitor researcher at USC and Stanford. These internships allowed me to grow as a researcher, and besides the publications I will always remember the beauty of California.

Vita

Research Interests

I am interested in the areas of *Computational Social Science*, *Network Science*, and *Machine Learning for Social Good*. My main research focuses on studying the influence of *network structure* in algorithms (i.e., ranking, sampling, and relational classification) and human behavior (e.g., navigation). My passion is to get involved with new algorithms that allow me to discover meaningful patterns and trends behind the interactions of the crowd, and to propose fair solutions that can be implemented in real scenarios.

Education

University of Koblenz-Landau, Germany 2016–2022
PhD. in Computer Science
Thesis: Edge Formation and its Influence in Machine Learning (bit.ly/PhD-LEEN)
Advisors: Prof. Dr. Matthias Thimm, Prof. Dr. Claudia Wagner, Prof. Dr. Markus Strohmaier

University of Saarland, Germany 2011–2014
Master in Computer Science
Thesis: Inferring topical context for content on Twitter (bit.ly/Master-LEEN)
Advisor: Prof. Dr. Krishna Gummadi

Escuela Superior Politécnica del Litoral (ESPOL), Ecuador 2004–2010
Engineering in Computer Science (Technological Systems)
Thesis: Analysis, design and implementation of an academic social network (bit.ly/Bachelor-LEEN)
Advisor: Prof. Dr. Xavier Ochoa Chehab

Research Experience

Complexity Science Hub (CSH) Jan'22–*present*
Post-doctoral Researcher
Project: Understanding and mitigating inequalities produced by network-based algorithms

Central European University (CEU) Jan'21–*present*
Post-doctoral Researcher
Project: Inferring high-resolution poverty maps with multimodal data

GESIS - Leibniz Institute for the Social Sciences May'15–Jan'21
Research Assistant
Project: Methods to understand and explain edge formation, and biases in network inference

University of Southern California (USC-ISI) Jun'18–Aug'18
Visiting Research Assistant
The influence of network structure in relational classification (bit.ly/NetworkBias)

University of Southern California (USC-ISI) Jun'17–Aug'17
Visiting Research Assistant
The influence of sampling in relational classification (bit.ly/SamplingBias)

Stanford Jan'17–Feb'17
Visiting Student Researcher
Project: The influence of semantic structure in navigation on BioPortal (bit.ly/HopRank)

Max Planck Institute for Software Systems (MPI-SWS) Aug'13–Oct'14
Intern and Research Assistant
Project: Inferring topical context on Twitter, and high level categories using DMOZ

Teaching Experience

GESIS, Cologne, Germany - virtual Sep'20
Introduction to Social Network Science with Python, 30 hours (bit.ly/MS-SNS-GitHub)

GESIS, Cologne, Germany Feb'18
Introduction to Machine Learning using Python: Data cleaning, pandas, visualization, prediction task, hyper-parameter tuning, 4 hours (bit.ly/Tutorial-ML1, bit.ly/Tutorial-ML2).

Invited Talks

Corvinus workshop, BCE NETi Lab - virtual 12 Jan'22
The Good, the Bad and the Ugly of inferring poverty maps with multimodal data (bit.ly/Talk-CW22)

Core Data Science, Facebook - virtual 29 Jul'21
Explaining and Improving Machine Learning Outcomes on Networks and Multimodal Data (bit.ly/talk-FBCDS21)

Women in Data Science (WiDS), ESPOL - virtual 2 Oct'20
Edge formation and its influence in Machine Learning (bit.ly/LEEN-WiDS2020)

Professional Experience

NEOBOX S.A., Guayaquil, Ecuador 2007–2020
Co-Founder, President and CTO. Startup for software development.

Information Technologies Center, Guayaquil, Ecuador 2007–2011
Project leader and developer at the e-Learning and Teaching Technologies group.

Research Center and Educational Services, Guayaquil, Ecuador 2006
Web and database manager for the academic census online system.

Selected Projects

TTopic: Tweet, hashtag and trending topic categorization based on user's expertise (DJango, PostgreSQL, Twitter API, Web scraping, JQuery).

TrullyFollowing: Web Search Interface of real/fake users on Twitter (AJAX, HTML, Django).

Reina: E-voting system for beauty queen pageants. Used in *Reina de Guayaquil* 2009-2016 and *Miss World Ecuador* 2014-2016 (AIR, ActionScript, MySQL, Server-client architecture).

Skills

Data Science: Deep learning (image classification and object detection). NLP (topic modeling, clustering, semantic search). Supervised machine learning (collective classification, random forests, logistic, linear, and multiple regression, quadratic assignment procedure). Unsupervised machine learning (Non-negative matrix/tensor factorization). Hypothesis testing and model selection using Bayesian statistics.

Programming Languages (currently using): Python [scripting and object-oriented] (TensorFlow, Pandas, GeoPandas, SciPy, NumPy, scikit-learn, networkX, seaborn, joblib, tqdm, pqdm, graph_tool), Bash, SBash

Programming Languages (used in the past): JAVA, C, C#, Visual Basic

Web Development: HTML, Javascript, CSS, AJAX, PHP, ActionScript, AIR, DJango

IDEs: Jupyter Notebooks, PyCharm, Netbeans, Eclipse

Applications: Overleaf, GitHub

Database Engines: PostgreSQL, MySQL, SQLite

Data Exchange formats: CSV, Pickle, JSON, XML, HDF5, GEOJson

Operating Systems: Linux (Debian, Ubuntu), Mac OS, Windows

Other systems: SLURM for high performance computing (HPC) clusters

Spoken Languages: Spanish (native), English (professional), German (basic)

Honors and Awards

Best Reviewer: ICWSM'21 and TheWebConf'18.

Best Poster: "Biases in Relational Classification" LatinXinAI at ICML 2020.

Honorable mentions: PC member at TheWebConf'22. Project award on methods for analyzing and modeling textual data, CSS Summer School 2018.

Academic Research Grant: Google Cloud credit for $1,648.80$€.

Travel Award: NeurIPS 2018 for $250.00.

Stanford Fellowship 2017: A 2-month fellowship at the Biomedical Informatics Research department (Stanford-USA).

Summer School 2016: Complex networks: theory, methods, and applications (Como-Italy).

IceLab Camp 2015: Interdisciplinary research camp for young scientists (Umeå-Sweden).

Google CodeF 2012: An exclusive career event for female computer scientists. First place in the programming challenge (Munich-Germany).

SENESCYT 2011: An ecuadorian scholarship to obtain a Master's degree in Computer Science in Germany. Acceptance rate 15%.

Hobbies

Online reading club with kids (Inspiring kids through the stories of extraordinary men and women).

Certified Zumba® Fitness Instructor since 2013.

References

Fariba Karimi. Group leader at the Computational Social Science Department, Complexity Science Hub, Vienna.

Márton Karsai. Associate Professor and Director of the PhD Program in Network Science, Central European University, Vienna.

Claudia Wagner. Professor for Applied Computational Social Science, RWTH Aachen University. Head of the Computational Social Science Department, GESIS–Leibniz Institute for the Social Sciences, Cologne. External Faculty Member Complexity Science Hub, Vienna.

Markus Strohmaier. Chair for Data Science in the Economic and Social Sciences Business School, University of Mannheim. Scientific Coordinator for Digital Behavioral Data, GESIS–Leibniz Institute for the Social Sciences, Cologne. External Faculty Member Complexity Science Hub, Vienna.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Motivation

Social networks are everywhere! In this thesis I refer to social networks as the intangible structure that we create when connecting directly or indirectly with people. For instance, all our friends (and friends of our friends) on Facebook, the cascades of retweets of a particular post on Twitter, and the citations among articles from a specific scientific field. Even the transitions that people make when navigating the Web or a city can be considered as social networks. In other words, in this thesis a social network is any form of relation or interaction between people or made by people.

Social networks have been widely studied over the last century in classical sociology [177–179], and more recently across the social, physical, and computer sciences [31, 91, 171, 226, 261, 262]. Their popularity in the scientific field is due to their complex structures. The way people connect to each other provides powerful insights for a plethora of social phenomena. For instance, it has been shown that the structure of a social network can influence *employment rates* between social groups [11, 247], *managerial performance* [213], and *epidemic spread* [136, 235]. In the past, such studies required surveying and even observing people to construct such networks [179]. Today, they can be studied at larger scales, due to the Web and mobile technologies [102, 131, 139, 191]. In fact, everyday we leave fine-grain footprints of our actions and interactions which collectively may reflect characteristics of our preferences and personality [7, 150, 241].

While all these interactions are intangible to us, they are stored in big data centers for statistical purposes, in particular, forecasting. Online platforms such as Facebook and Twitter collect and analyze all our *digital traces* to not only infer information about ourselves (e.g., romantic partnerships [14] and political ideology [126]), but also to recommend us new content [229], new products [277], and even new connections [15, 106, 158]. There is no doubt that these predictions are beneficial to us because they bring us closer together [183], and can potentially increase the diversity of content that we consume [221]. However, at the same time, they can harm society by reinforcing under-representation of minorities [20, 55, 83, 133, 243], denigration [231, 280], polarization in social media [70, 210] and segregation [28, 169, 248].

In particular, these predictions are based on *machine learning* algorithms which rely on our *digital trace data*. Therefore, if we connect with certain people more often than with others, these predictions will likely recommend us content or people from the same circle; a phenomenon known as *filter bubbles* [119, 197] and *echo chambers* [58, 86, 205]. Besides, it has been shown that recommendation algorithms suffer from the *popularity bias* problem[1] [2, 24], where popular items get a lot of exposure while less popular ones are under-represented in the recommendations. This leads to the isolation of unpopular people and therefore restricts them from information and new opportunities. Consequently, by accepting these recommendations we create a feedback loop [166] and reinforce societal issues such as the invisibility syndrome [36, 89, 196] and the glass ceiling effect [11, 53, 243].

Debiasing machine learning outcomes and making them fair, however, is very challenging. First, *fairness* is an *essentially contested construct* that has different theoretical understandings in different contexts [127]. Second, the data itself can be biased due to missing data [161], selective exposure [90], historical prejudices or implicit bias [35, 164]. Third, we need to understand the workflow of the algorithm and be aware that biases can be introduced at any step in the machine learning pipeline. In classification, for instance, there can be six sources of bias: data collection, data preparation, model development, model evaluation, model post-processing, and model deployment [246]. In ranking, however, biased outcomes are often generated by the algorithms due to item popularity [24]. Last but not least, unbiased results often come at the expense of lower accuracy, a trade-off that most algorithms face when mitigating fairness issues [45, 48, 121, 174, 272]. While most of these bias mitigation techniques have been framed as an optimization problem to calibrate algorithms with fairness constraints [66, 111, 148], only a few investigate biased outputs in context of the data [45, 49, 54, 105, 120]. In this direction, it has been shown that biased predictions are mostly caused by biased training data [54]. Therefore, collecting "better quality" of training data, can reduce discrimination in predictive models without sacrificing accuracy (e.g., by collecting more data for the under-represented group) [45].

However, in the context of social networks, where data is *relational*, it is unclear whether balanced samples across groups of people are enough to guarantee good accuracy and unbiased results. Networks are finite sets of nodes and edges that altogether define a structure. This structure encodes properties such as edge density, the directionality of links, fraction of minorities, preferential attachment, and homophily. While pairwise relationships between people have been useful for classification and ranking [6, 43, 106, 165, 193], little is known about the effects of these properties collectively on the performance of machine learning algorithms. This thesis is a step towards that direction.

By proposing and utilizing new network models I generate networks with different structures and measure the correlation between the input and the output. In other words, I look for patterns between network structure and machine learning outcomes (i.e., accuracy and bias). As a result, I create benchmarks for inter-

[1] In other fields also known as the *rich-get-richer* effect [68], the *Matthew* effect [175], *cumulative advantage* [176], and *preferential attachment* [18].

pretable and explainable machine learning on networks to help data scientists understand and estimate what these algorithms will predict given the structure of a social network.

1.2. Problem Statement, Objectives, and Approach

Machine learning algorithms have become an important part in decision making processes. Tasks such as weather and revenue forecasting are becoming more accurate thanks to richer models that take into account lots of historical data. However, the application of algorithms to high-stake situations such as health care, credit scoring, sentencing and policing requires novel evaluation criteria that go beyond predictive performance and take fairness, potential harms, interpretability and explainability into account. The goal is then to understand why an algorithm arrives at certain conclusions or outcomes given new data. This is particularly challenging because unveiling the machine learning black box can be complex not only at explaining how the algorithm works, but also at representing all possible scenarios of the data; in this case: social networks. For instance, what works well for sparse networks, might not work for dense networks, and so on.

This motivated me to study *how edges form* in real-world social networks, and *to what extent machine learning algorithms fail* on social network data. The main goal is to propose: (i) tools to *characterize* edge formation, (ii) benchmarks to evaluate outcomes of classification, ranking and recommendation algorithms applied to social networks, and (iii) interventions to mitigate inequalities or biases found in machine learning outcomes.

To this end, as a starting point, I leverage the fact that social networks possess a structure determined by how nodes connect with each other. In particular, I focus on two well known characteristics found in social networks: preferential attachment [18] and homophily [171]. The former is a mechanism where nodes in a network prefer to connect to other nodes that are popular in terms of the number of connections they already have (e.g., celebrities on Twitter). The latter is a mechanism where nodes tend to connect to other nodes based on their similarities or differences (e.g., connecting only to men on a dating app). This leads to the **first part** of this thesis where I borrow theories of edge formation from the social sciences and physics and propose two techniques to characterize edge formation on real-world social networks. The idea is to generate hypotheses of edge formation using the attributes of nodes (e.g., gender) and the distances between them (e.g., social or physical distance) to then rank them based on their plausibility given the data. The advantage of this approach over traditional model fitting is *interpretability*. Learned model parameters are hard to understand whereas hypotheses are created using the analyst's intuition which helps the understanding of edge formation dynamics. In the **second part** of this thesis I focus on the implications of these mechanisms of edge formation on machine learning algorithms, in particular, classification, ranking and recommendation algorithms. The goal is to find correlations between the structure of the network (e.g., homophily and preferential attachment) and the evaluation metrics of these algorithms. Therefore, I look for such correlations in real-world social networks to show the impact of biased results in the real world. Moreover, I generalize these correlations by generating a wide spectrum of networks with different structures using random network models. In particular, I use a growth model of undirected networks and propose an extended version that is able to generate realistic directed networks with adjustable homophily, preferential attachment, node activity, fraction of minorities, and edge density.

1.3. Research Questions

Aligned to the objectives of this thesis, I define three research questions which allow me to focus on the characterization of edge formation (RQ1) and the role of these mechanisms of edge formation in machine learning, in particular on relational classification (RQ2) and network-based ranking and recommendation algorithms (RQ3).

RQ1 How can we characterize the underlying mechanisms of edge formation of a given network?

RQ2 How do the mechanisms of edge formation influence classification performance and the direction of bias in relational classification?

RQ3 How do the mechanisms of edge formation affect the distribution of opportunities across individuals and minorities in ranking and recommendation algorithms?

1.4. Main Publications

The core chapters of this thesis are based on results from the following articles:

- Article 1 [78]: Lisette Espín-Noboa, Florian Lemmerich, Markus Strohmaier, and Philipp Singer. Janus: A Hypothesis-driven Bayesian Approach for Understanding Edge Formation in Attributed Multigraphs. *Applied Network Science*, 2(1), pp. 1-20. 2017. 10.1007/s41109-017-0036-1.

- Article 2 [79]: Lisette Espín-Noboa, Florian Lemmerich, Simon Walk, Markus Strohmaier, and Mark A. Musen. HopRank: How Semantic Structure Influences Teleportation in PageRank (A Case Study on BioPortal). In *The World Wide Web Conference*, pp. 2708-2714. 2019. 10.1145/3308558.3313487.

- Article 3 [77]. Lisette Espín-Noboa, Fariba Karimi, Bruno Ribeiro, Kristina Lerman, and Claudia Wagner. Explaining Classification Performance and Bias via Network Structure and Sampling Technique. *Applied Network Science*, 6(1), pp. 1-25. 2021. 10.1007/s41109-021-00394-3.

- Article 4 [81]. Lisette Espín-Noboa, Claudia Wagner, Markus Strohmaier, and Fariba Karimi. Inequality and Inequity in Network-based Ranking and Recommendation Algorithms. *Scientific reports*, 12(1), pp. 1-14. 2022. 10.1038/s41598-022-05434-1.

Additionally, the following publications contributed to formulating the basic ideas of this thesis.

- Article 5 [72]: Lisette Espín-Noboa, Florian Lemmerich, Philipp Singer, and Markus Strohmaier. Discovering and Characterizing Mobility Patterns in Urban Spaces: A Study of Manhattan Taxi Data. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 537-542. 2016. 10.1145/2872518.2890468.

- Article 6 [80]: Lisette Espín-Noboa, Claudia Wagner, Fariba Karimi, and Kristina Lerman. Towards Quantifying Sampling Bias in Network Inference. In *Companion Proceedings of The Web Conference*, pp. 1277-1285. 2018. 10.1145/3184558.3191567

- Article 7 [259]: Simon Walk, Philipp Singer, Lisette Espín-Noboa, Tania Tudorache, Mark A. Musen, and Markus Strohmaier. Understanding How Users Edit Ontologies: Comparing Hypotheses About Four Real-World Projects. In *International Semantic Web Conference*, pp. 551-568. 2015. 10.1007/978-3-319-25007-6_32.

- Article 8 [258]: Simon Walk, Lisette Espín-Noboa, Denis Helic, Markus Strohmaier, and Mark A. Musen. How Users Explore Ontologies on the Web: A Study of NCBO's BioPortal Usage Logs. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 775-784. 2017. 10.1145/3038912.3052606.

- Article 9 [153]: Florian Lemmerich, Philipp Singer, Martin Becker, Lisette Espín-Noboa, Dimitar Dimitrov, Denis Helic, Andreas Hotho, and Markus Strohmaier. Comparing Hypotheses about Sequential Data: A Bayesian Approach and its Applications. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 354-357. 2017. 10.1007/978-3-319-71273-4_30.

As a first author, I implemented the methods and experiments of each article and wrote their first drafts. The design of the experiments and improvements of each paper were possible thanks to the collaboration with my co-authors. Furthermore, the paper "HopRank: How Semantic Structure Influences Teleportation in PageRank (A Case Study on BioPortal)" published at the Web Conference 2019 [79], is the result of a two-month fellowship at BMIR, the Stanford Center for Biomedical Informatics Research. Similarly, the papers "Towards Quantifying Sampling Bias in Network Inference", published at the Web Conference 2018 [80], and "Explaining Classification Performance and Bias via Network Structure and Sampling Technique", published at the Applied Network Science journal 2021 [77], are the results of two three-month internships at USC-ISI, the Information Sciences Institute at the University of Southern California. By using the plural form "we" from Chapter 2 to Chapter 5, I acknowledge and honor my co-authors.

1.5. Main Contributions

The main contributions of this thesis are summarized as follows:

1. I propose methods to create and rank hypotheses of edge formation in attributed and non-attributed social networks.

2. I propose methods and benchmarks to evaluate (i) performance and bias in relational classification and (ii) inequality and inequity in network-based ranking and recommendation algorithms.

3. I implement replicable, reproducible and reusable code and data[2].

Besides, this thesis as a whole shows how the interplay between network structure and machine learning affects social phenomena. In principle, the mechanisms of edge formation are *biased connections* that are organically triggered by our preferences or choices. However, these biases are being reinforced by machine learning methods which replicate or amplify what they learn from the data. For instance:

- My systematic study on the effects of network structure on relational classification shows that classification performance is mainly driven by the type of network [77]. This means that the evaluation benchmarks that I propose make relational classification *interpretable* and *explainable*. Interpretable because these benchmarks can tell what the performance and direction of bias will be on average given the sample size, the homophily and the fraction of minorities of a given network. Explainable because thanks to the dynamics between homophily and preferential attachment the results of the classification can be intuitively understood by practitioners. In conclusion, performance and bias in relational classification can be anticipated and therefore mitigated before relying on erroneous results.

[2] For a definition of replicable, reproducible and reusable research see Schoch 2017 [219, 222]

- In the context of ranking and recommendation algorithms, my systematic study of the effects of network structure on PageRank and Who-To-Follow suggest that minorities are not always under-represented in top ranks; they are just not well connected in the network. Certainly, one way to fix a biased ranking is to define and apply diversity constraints or quotas at the top of the rank (e.g., 50% women - 50% men). A more prominent intervention is to ensure that ranked outcomes achieve *proportional representation* across groups at higher ranks [127, 269] (e.g., if the population or pool of applicants is $80\% - 20\%$, the top rank should also be $80\% - 20\%$). By following this criterion, my findings suggest that minorities are not necessarily under-represented for being a smaller group, but actually because they do not receive enough connections from people in the majority group. This means that by connecting strategically in a network, ranking and recommendation algorithms can be fair without calibrating the algorithm or adjusting their outcomes. In some cases, these strategic connections can be easily created without algorithmic intervention (e.g., by citing women more often in scientific papers). However, in cases where this is not possible, we can rely on additional recommender systems to suggest those important missing links.

1.6. Structure of this Thesis

This thesis is split into two parts.

In Part I, I focus on characterizing edge formation on social networks. In **Chapter 2**, I present HopRank to study edge formation in the context of human navigation on BioPortal [263], one of the leading repositories of biomedical ontologies on the Web. In general, HopRank is an algorithm for modeling human navigation on semantic networks. Its key component is a random walker that is biased towards k-hop neighborhoods. This means that it assumes that users besides following links, also follow nodes at certain distances and not at random as suggested by PageRank [193]. I observe such preference towards k-hop neighborhoods on BioPortal. In particular, users navigate within the vicinity of a concept, but they also "jump" to distant concepts less frequently. I compare HopRank with seven other models of human navigation on networks. I run experiments on synthetic and real-world networks, and evaluate BIC scores for model selection. In **Chapter 3**, I show Janus, a hypothesis-driven Bayesian approach that allows to intuitively compare hypotheses about edge formation in multigraphs. For example, in a co-authorship network one hypothesis might be that authors are more likely to collaborate with each other "if they are from the same country", or "if one author is more senior than the other". Once the hypotheses are defined as belief matrices, they are encoded as priors into the Bayesian framework. Then, the final output is a ranking of hypotheses based on their Bayes factors or plausibility given the data. I show experiments on the Kenya contact network [139] and the Higgs Twitter dataset [57], and find that people in the Kenya network tend to interact more often with people from the same household than with people of similar age, while the Twitter reply network in the Higgs dataset is better explained by mentions than by retweets. I compare Janus with QAP [122] and highlight important caveats for further improvements.

In Part II, I focus on explaining how certain mechanisms of edge formation influence machine learning outcomes. In **Chapter 4**, I show how preferential attachment, homophily, fraction of minorities, and edge density influence the performance and bias of relational classification [43, 165], a technique that infers the class label of a node based on the class label of its neighbors. I formulate three research questions to disentangle output bias into input bias (i.e., network structure) and sampling bias (i.e., sampling technique and sample size). Using synthetic networks, I propose benchmarks to evaluate binary classification and corroborate that these benchmarks hold for real-world social networks. Similarly, in **Chapter 5**, I show how the structure of a network influences inequality and inequity in the rank distribution of PageRank [193] and Who-To-Follow [106], two well established network-based ranking and recommendation algorithms. I measure *inequality* as the Gini coefficient of the rank distribution and *inequity* as how fair these algorithms are with respect to the proportional representation of groups in top ranks. I define *fairness in ranking* as a diversity constraint that requires each group to be represented at the top of the rank in proportion to its prevalence in the network. I propose the **D**irected network with **P**referential **A**ttachment and **H**omophily model (DPAH) and compare it with other models to disentangle the individual effects of homophily, preferential attachment, edge density and node activity on the inequality and inequity of top ranks. My main findings suggest that (i) inequality and inequity are positively correlated, (ii) inequity is mainly driven by homophily, and (iii) inequality is driven by the interplay between preferential attachment, homophily, node activity, and edge density.

Table 1.1 provides an overview of the main chapters of this thesis. Each chapter is summarized by showing its relationship with the presented research questions, main publications, data, and methods utilized to achieve the described goals.

This thesis concludes with **Chapter 6** where I summarize the main results and contributions of each chapter. Moreover, I discuss important implications of my findings on real-world applications. Finally, I provide an overview of limitations and future directions where the mechanisms of edge formation can be leveraged to solve specific social phenomena.

Table 1.1.: **Outline.** This table summarizes the main chapters of this thesis. Each chapter is based on an article that answers a particular research question RQ.

Part	Chapter	RQ	Focus	Goal	Data	Methods
I	Chapter 1 based on [79]	RQ1	Human navigation on networks	Characterize navigation patterns on networks using k-hops.	BioPortal ontologies and click streams	HopRank, PageRank [193], Markov Chains [204]
I	Chapter 2 based on [78]	RQ1	Edge formation in social networks	Characterize how edges form in a given social network using information from the nodes.	Social networks with attributed nodes	Janus, QAP [122]
II	Chapter 3 based on [77] [80]	RQ2	Classification on undirected networks	Explain classification performance and bias through network structure and sampling technique.	Scale-free undirected social networks	Relational Classification [43], Collective Inference [227], BA-Homophily [133]
II	Chapter 4 based on [81]	RQ3	Ranking and recommendation on directed social networks	Explain when ranking algorithms reinforce inequality and inequity in directed social networks.	Scale-free directed social networks	PageRank [193], Who-To-Follow [106], DPA, DH, DPAH

Part I.

Characterizing Edge Formation

2. Characterizing Edge Formation on Non-Attributed Networks

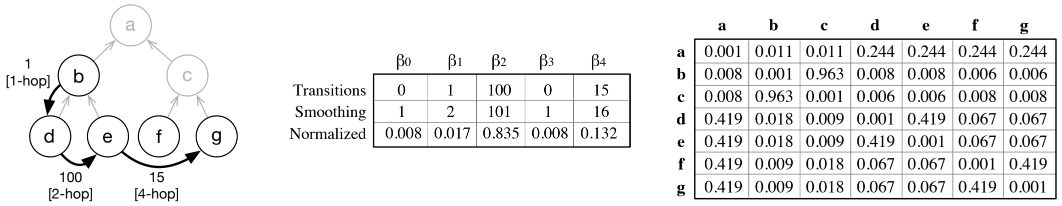
This chapter introduces HopRank, an algorithm for modeling human navigation on semantic networks, and it has been published as a short paper in the Proceedings of The World Wide Web Conference WWW 2019.

HopRank leverages the assumption that users know or can see the whole structure of the network. Therefore, besides following links, they also follow nodes at *certain distances* (i.e., k-hop neighborhoods), and not at random as suggested by PageRank, which assumes only links are known or visible. We observe such preference towards k-hop neighborhoods on BioPortal, one of the leading repositories of biomedical ontologies on the Web. In general, users navigate within the vicinity of a concept. But they also “jump” to distant concepts less frequently. We fit our model on 11 ontologies using the transition matrix of clickstreams, and show that semantic structure can influence teleportation in PageRank. This suggests that users—to some extent—utilize knowledge about the underlying structure of ontologies, and leverage it to reach certain pieces of information. Our results help the development and improvement of user interfaces for ontology exploration.

2.1. Introduction

Ontology Engineering and Ontology Learning are two branches of the Semantic Web whose aim is to accurately build and curate ontologies. The former studies new techniques to improve collaboration among humans while editing ontologies [251, 259], and the latter introduces new methodologies and algorithms to automatically create ontologies by crawling the Web [10, 245]. These efforts represent significant advances in the development of knowledge bases, which represent facts about the real world (e.g., people, diseases). However, there is little knowledge about how users consume such ontologies on the Web. To this end, Walk et al. studied how users browse BioPortal [258]. Their findings suggest that some ontologies influence the way users interact with the website. However, how users navigate through the ontology structure (i.e., from one concept to another) remains unclear.

Problem statement. In this paper, we study the influence of semantic structure on *teleportation* (i.e., jumping to any node chosen at random) in PageRank. For



(a) Transitions

	β_0	β_1	β_2	β_3	β_4
Transitions	0	1	100	0	15
Smoothing	1	2	101	1	16
Normalized	0.008	0.017	0.835	0.008	0.132

(b) HopPortation Vector

	a	b	c	d	e	f	g
a	0.001	0.011	0.011	0.244	0.244	0.244	0.244
b	0.008	0.001	0.963	0.008	0.008	0.006	0.006
c	0.008	0.963	0.001	0.006	0.006	0.008	0.008
d	0.419	0.018	0.009	0.001	0.419	0.067	0.067
e	0.419	0.018	0.009	0.419	0.001	0.067	0.067
f	0.419	0.009	0.018	0.067	0.067	0.001	0.419
g	0.419	0.009	0.018	0.067	0.067	0.419	0.001

(c) Transition Probabilities

Figure 2.1.: **HopRank on semantic networks.** This example illustrates an instance of navigation on an ontology. (a) Shows the underlying network composed by seven concepts ($a, b, \ldots, g$) and six *isASubClassOf* relationships (straight-thin grey arrows). Transitions (curved-thick black arrows) are labeled by the actual number of transitions between concepts, as well as the [k-hop] distance (i.e., shortest path) between them. (b) Illustrates how the HopPortation vector $\vec{\beta}$ is built using transition counts per k-hop. (c) Shows the transition probabilities inferred by HopRank, see Equation (2.1).

example, consider the ontology shown in Figure 2.1(a), where nodes represent *classes* (a.k.a. concepts) and edges *isASubClassOf* relationships. On BioPortal, ontologies are shown vertically as hierarchical trees, and concepts can be explored using the *expand-on-demand* principle. This means that only top level concepts are shown first, and then users are able to expand and collapse as many concepts as they need at any level of the ontology. In other words, users can use and therefore are potentially aware of a *virtually fully connected network* in all stages of navigation. Previous studies [193, 267] have modeled user navigation using PageRank. However, these assume that navigation paths are constrained by links and random teleportation. In our scenario, where the whole structure of an ontology can be visualized at any time, we believe that teleportation is not fully random, but rather biased towards k-hop neighborhoods.

Approach. Motivated by previous studies on information foraging [40, 46, 47, 203], decentralized search [113, 141], and PageRank [1, 103, 112, 193, 267], we propose HopRank, a method for modeling transitions across k-hop neighborhoods on semantic networks. The key idea of this work relies on the HopPortation vector $\vec{\beta}$, which defines the probabilities of transitioning to each k-hop neighborhood. From the PageRank point of view, we can say that teleportation is not fully random, and the probability of following the structure of a page is not based only on one parameter (i.e., probability of following links), but on k parameters, representing all k-hop neighborhoods *reached from the current page.* Technically, we pass the HopPortation vector to a random walker to make biased decisions on which neighborhood to go next. Once this decision is made, the random walker uniformly chooses a concept within that neighborhood.

Contributions. The contributions of this paper are:

1. We empirically show how users leverage the structure of the ontologies on BioPortal by quantifying the proportion of transitions per k-hop neighborhood.

2. We propose HopRank, an algorithm for modeling human navigation on semantic networks.

3. We demonstrate that HopRank outperforms traditional navigation and popularity-based models on BioPortal, especially when users browse ontologies directly without search.

4. We make an implementation of this approach openly available on the Web [74].

2.2. Related Work

BioPortal provides users with a *tree-like explorer* and a *local search* engine to navigate ontologies. In addition, concepts can be *expanded on demand* to see their children nodes. Although these functionalities are exploited differently across ontologies [258], it is unclear how users navigate through the ontology structure. Thus, this section covers previous work on search and navigation on networks.

Search. *Information Foraging* [203] assumes that people, when possible, modify their strategies or the structure of the environment to maximize their gain of valuable information. These patterns are also found in the way humans recall information from memory [114]. Similarly, *berrypicking* [22], a model of online searching, states that queries are not static, but rather evolve, and users commonly gather information in pieces instead of in one large set.

Navigation. *PageRank* [193] is the most popular method to measure the importance of web pages based on their incoming and outgoing links. It relies on an imaginary surfer who is randomly *clicking on links*, and eventually *jumps* to any node in the network. The probability of following links is given by a *damping factor*. Multiple variations have been proposed for improving information retrieval systems, e.g., a biased PageRank [112] to capture the importance of a page more accurately by taken topics into account or a weighted PageRank [267] to assign larger rank values to more popular pages (i.e., preferential attachment) instead of distributing the rank value of a page uniformly to all outgoing pages. Geigl et al. suggest that the behavior of a *random surfer* is almost similar to real users, as long as they do not use search engines [94]. They also find that classical navigation structures, such as navigation hierarchies or breadcrumbs, only exercise limited influence on navigation. Experiments in [233] reveal that memory-less *Markov chains* represent a quite practical model for human navigation on a page level. However, this assumption is violated when the analysis is expanded to a topical level. Helic et al. identify certain configurations of *decentralized search* that are capable of modeling human navigation in information networks [113]. Their findings suggest that navigation on such networks is a two phase process combined with the *exploitation of the known* (i.e., goal-seeking) and the *exploration of the unknown* (i.e., orientation).

User interfaces. Human navigation has also been studied for enhancing interfaces. For instance, [92] explores *fisheye* views to display large information structures such as programs and databases. The intuition behind this paradigm is that users often explore their neighborhood, and distant major landmarks in more detail. Similarly, Van Ham and Perer studied the *search, show context, expand on demand* browsing model in [252], and proposed techniques to design better graph visualization tools.

We propose HopRank—a biased random walker—to model navigation on semantic networks. HopRank builds upon insights from information foraging [114, 203], decentralized search [113, 141] and PageRank [193]. More precisely, we replace the *damping factor* by a HopPortation vector to encode the probabilities of visiting each k-hop neighborhood. The intuition here is that users browse semantically close terms more often than semantically distant ones.

2.3. BioPortal

There exist a large number of ontologies in the biomedical domain. They are highly specialized and therefore expensive to develop. To enable ontology adoption and reuse, effective support for browsing and exploring existing ontologies is required. Towards that goal, the National Center for Biomedical Ontology (NCBO) [185, 186] features BioPortal [26, 192, 263]—one of the leading repositories of biomedical ontologies on the Web—containing currently more than 800 ontologies with more than 13 million ontology classes. On BioPortal, practitioners and experts can access ontologies via Web services and Web browsers. The latter allows users to navigate ontologies by searching specific classes, or by directly browsing their concept hierarchies within a tree-like explorer [258].

Ontologies. We propose to model human navigation on semantic networks using the structure of the underlying ontology. On BioPortal, ontologies are defined as directed networks, where nodes represent *concepts* and edges *isASubClassOf* relationships. Since such edges are usually non-cyclic and have a common root, these ontologies often form trees. Table 2.1 shows 11 of the most visited ontologies in 2015[1]. For instance, LOINC the largest ontology with $175K$ nodes, $153K$ edges, and $74K$ connected components.

Transitions. We analyzed all HTTP requests made in 2015 and extracted $336K$ valid sessions (i.e., after filtering out sessions with less than 2 requests, and requests to ontologies or concepts which do not exist). Each session contains transitions (i.e., a sequence of visited concept pages) triggered by a single

[1] As ontologies can be edited over time, we work with their latest snapshots from 2015.

user (i.e., IP address) without breaks (i.e., pauses of at least 60 minutes). For simplicity, we only consider transitions within the largest connected component (LCC) of each ontology, and discard ontologies with less than 1000 transitions[2]. Overall, we found 11 ontologies and $133K$ transitions between their concepts[3], see Table 2.1 for some key properties.

Navigation types. Based on the HTTP request headers, we inferred 7 navigation types: Details (DE), Direct Click (DC), Direct URL (DU), Expand (EX), External Link (EL), External Search (ES), and Local Search (LS). **DE**: are all clicks made within the *Details* tab of a selected concept. **DC**: are all clicks made on concepts within the tree-like explorer. **DU**: refers to all concept requests without HTTP referrer (e.g., direct URL in the browser). **EX**: considers all clicks on the (+) symbol of a concept, which triggers the expansion of the concept to show all its children nodes. Notice that this request is called only once, even if the symbol is clicked multiple times. The opposite behavior (collapse) is not considered[4]. **EL**: captures all requests coming from external websites that are not search engines. **ES**: are all requests coming from the top 10 most popular external search engines such as Google and Yahoo. **LS**: are all requests made via the local search

[2] Transitions within the LCCs of these ontologies represent 80% of all transitions.

[3] We left out the popular SNOMEDCT ontology due to computational limitations.

[4] Collapse is a client-side functionality, and thus, it is not recorded in the log files.

Table 2.1.: **Datasets.** This table illustrates network properties of 11 of the most popular ontologies on BioPortal in 2015. Ontologies represent networks whose nodes refer to *concepts* and edges *isASubClassOf* relationships. Original number of nodes, edges, and connected components of ontologies are shown under N, E and cc, respectively. Properties of the largest connected component (LCC) of each ontology are shown under N', E', d' and T', where d' refers to the diameter and T' to the number of transitions.

#	**Ontology**	**N**	**E**	**cc**	**N'**	**E'**	**d'**	**T'**
1	CPT	13219	13235	3	13092	13110	15	44651
2	MEDDRA	66506	31863	43493	22889	31738	8	42746
3	NDFRT	35019	34504	522	32074	32080	24	22452
4	LOINC	174513	152683	73518	100871	152558	13	6349
5	ICD9CM	22534	22531	3	22407	22406	12	4434
6	WHO-ART	1852	2997	3	1725	2872	4	2811
7	MESH	165166	24182	145652	16947	21596	31	2623
8	ICD10	12446	11256	1190	11132	11131	10	2288
9	CHMO	2966	3071	3	2964	3071	22	1423
10	HL7	10319	10600	1049	9146	10475	19	1374
11	OMIM	81821	39359	44110	37587	39234	6	1291

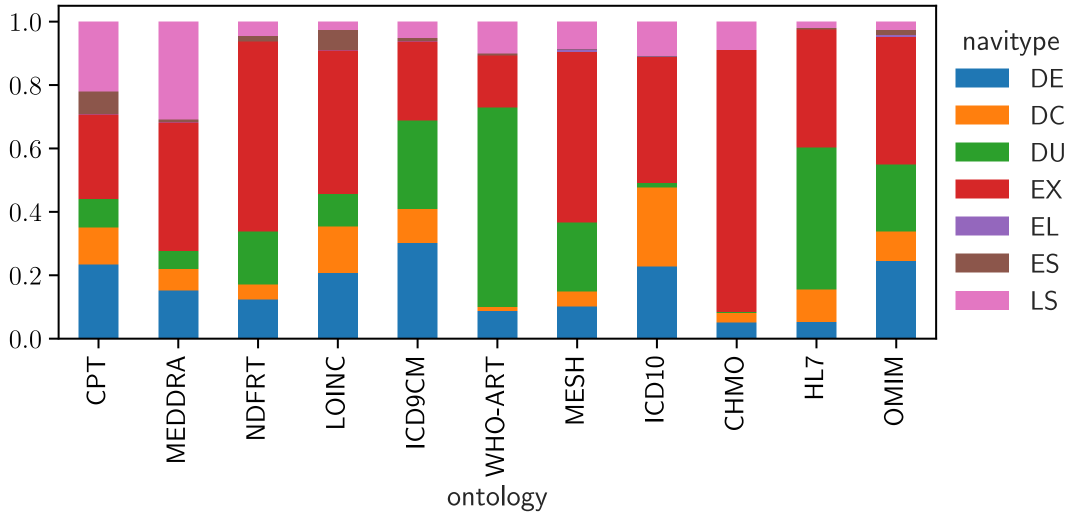


Figure 2.2.: **Navigation Types.** Each bar shows the fraction of transitions within the LCC of each ontology. Stacked bars differentiate types of navigation: details (DE, blue), direct click (DC, orange), direct URL (DU, green), expand (EX, red), external link (EL, purple), external search (ES, brown) and local search (LS, pink). Most ontologies are mainly navigated by *expanding* nodes within the tree-like explorer.

functionality of each ontology. Notice that this search is a 3-step process. First users type a keyword, then the system shows auto-suggestions and finally users click on one of the concepts shown in the auto-suggestion list. We only consider the final step a local search transition. **ALL**: includes all the above-mentioned types. Figure 2.2 shows the distribution of transitions across navigation types for each ontology. In general, most traffic comes from expanding a concept (EX, 44%), followed by local search (LS, 17%), direct URL (DU, 16%) and details (DE, 14%). Surprisingly, direct clicks on concepts (DC) only represent 7% of all transitions. This suggests that users spend substantial time expanding concepts before they find a concept of interest.

2.4. HopRank: A Biased Random Walker

HopRank models human navigation on semantic networks. Imagine a random walker whose decisions on where to go next are biased towards specific k-hop neighborhoods. This bias is what we call *HopPortation*, which encodes the probabilities of transitioning to each k-hop neighborhood. In our model, navigation on networks can be explained as a 2-step process. First, a k-hop neighborhood of the current node i is drawn from a categorical distribution. Second, a node j is randomly chosen within that k-hop neighborhood. Note that this process holds only if the walker is fully or partially aware of the structure of the network (i.e., knows or can see it). Without this prerequisite, and if links are not preferred, then jumps to random pages will be more plausible. In comparison to the classic random walker with teleportation (e.g., PageRank [193]), where its movements are constrained by the damping factor α (i.e., probability of following links), HopRank is constrained by a vector $\vec{\beta}$ containing k different factors, which define the probabilities of going to each k-hop neighborhood from the current location.

Visited k-hop neighborhoods on BioPortal. We aggregate ALL transitions by the shortest distance between two sequentially visited nodes. This distance is referred to as k-hop neighborhood. In Figure 2.3(a) we see that target nodes at large distances are less likely to be visited next. This is expected, since—to some extent—larger distances enclose more branches, therefore more target candidates. Note that ontologies are sorted by diameter in descendant order from MESH to WHO-ART. Interestingly, users tend to hop as far as the ontology's diameter, for $d' \leq 12$. For instance, OMIM's diameter is 6 (see Table 2.1), and 6 is the maximum hop done by users. Otherwise, users (roughly) hop up to two-thirds of the ontology's diameter, for $d' > 12$. For example, MESH's diameter is 31, and the largest hop reached is 19.

Transitions per k-hop neighborhood on BioPortal. Figure 2.3(b) shows the average percentage of transitions across k-hop neighborhoods per navigation type. We see that users on average (ALL, grey) prefer to navigate through 2-hop (41%) and 1-hop (23%) neighbors. In particular, when navigation is triggered by direct clicks (DC, orange) and expand (EX, red). Notice their fast decay when $khop > 8$. Other types of navigation such as external link (EL, purple), and direct URL (DU, green)—which do not leverage the tree-like explorer—tend to reach concepts at larger distances more frequently. Notice their peaks at $khop = \{5, 11\}$, respectively. Interestingly, when users opt for external search (ES, brown), they often click on 2-hop concepts, but also on 12-hop and 15-hop neighbors. Intuitively, the details tab (DE, blue) helps users to click on nearby concepts at $khop \leq 2$, more often than local search (LS, pink), which is more likely to reach concepts at $khop \geq 2$.

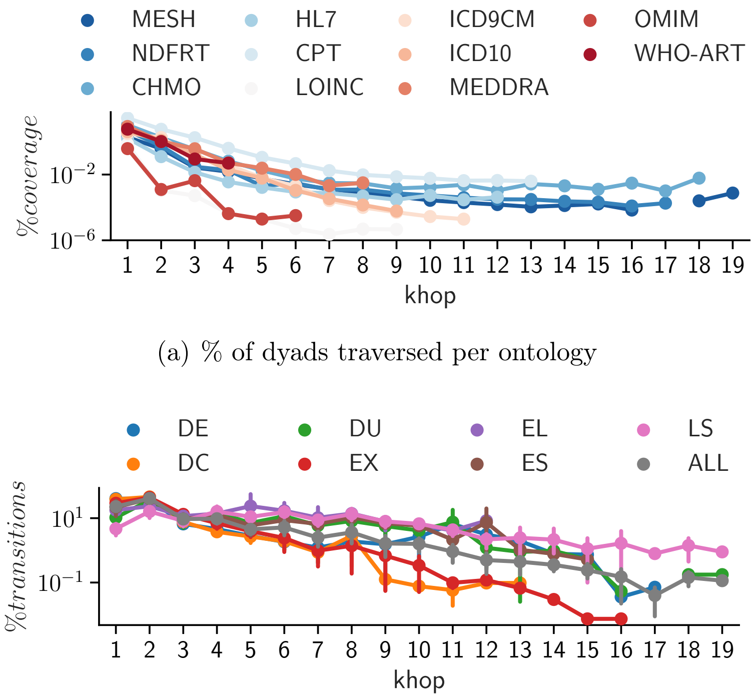
2.5. Models of Human Transitions

In this section, we formally introduce our HopRank model, and recap popular navigation models for comparison. We denote the transition probabilities, and # of parameters according to HopRank and 7 other models that we will use later on for model selection.

We formally represent an ontology[5] as a graph $G = (V, E)$, with $V = (v_1, \ldots v_n)$ being a set of N nodes, and $E = \{(v_i, v_j)\} \in V \times V$ a set of undirected edges[6]. The ontology structure is captured by the adjacency matrix $A_{N\times N} = a_{ij}$, where a_{ij} is 1 if the link exists, 0 otherwise. Transitions are represented by the transition

[5] We focus on its largest connected component (LCC)

[6] Directionality of edges is omitted to calculate shortest paths between all pair of nodes.



(a) % of dyads traversed per ontology

(b) Mean % of transitions per navigation type

Figure 2.3.: **Popularity of k-hops.** (a) Shows the percentage of dyads that are traversed per k-hop neighborhood. Lines represent ontologies and are sorted by their LCC diameter: In descendant order from MESH (darkest blue) to WHO-ART (darkest red). (b) Shows the distribution of transitions across k-hop neighborhoods per navigation type. Percentages are averages across ontologies, and error bars the respective standard deviation. While several k-hop distances are being traversed non-uniformly, most transitions happen across nearby nodes, especially when browsing (DE, DC, EX, ES) 2-hop neighbors. In contrast, non-browsing types (EL, LS, DU) tend to reach more distant nodes more frequently.

matrix $T_{N\times N} = t_{ij}$, where t_{ij} represents the number of transitions between source node i and target node j.

HopRank (HR). Given the HopPortation vector $\vec{\beta}$, the probability of reaching a k-hop neighborhood is denoted by factor $\beta_k \in \vec{\beta}$. The stochastic k-hop matrix M_k encodes all nodes j with a shortest distance k from i. HopRank uniformly distributes β_k across all nodes j at distance k. The limits of k-hop neighborhoods go from 1 (direct edges), to d', the diameter of the ontology G. Noise $\beta_0 = 1 - \sum_{k=1}^{d'} \beta_k$ is added to allow for random jumps and self-loops. Figure 2.1(b) illustrates how the HopPortation vector is computed from the transition counts. *Number of model parameters:* $d' + 1$.

$$P_{HR} = \beta_1 \boldsymbol{M}_1 + \beta_2 \boldsymbol{M}_2 + \cdots + \beta_k \boldsymbol{M}_k + \frac{\beta_0}{N} \tag{2.1}$$

Preferential Attachment (PA). Given the degree matrix $D_{N \times N} = d_{ij} = d_j$, where d_j represents the degree of the target node j. The probability of moving from i to j is proportional to the degree of j. *Number of model parameters:* 0.

$$P_{PA} = \boldsymbol{D} \tag{2.2}$$

Gravitational (Gr). Given the matrix $S_{N \times N} = (sp(i,j) + \epsilon)^2$, where $sp(i,j)$ denotes the shortest path between nodes i and j. The probability of navigating from i to j is proportional to the degree of node j and inversely proportional to the square distance between i and j. We add a smoothing factor ϵ to avoid overflows when dyads are disconnected. In such cases, we set ϵ to the diameter d' of G plus 1, to consider these jumps with a very low probability. Similarly, we set the diagonal (i.e., self-loops) to $\epsilon = d' + 2$. *Number of model parameters:* 0.

$$P_{Gr} = \frac{\boldsymbol{D}}{S} \tag{2.3}$$

Random Walker (RW). Given the damping factor α (i.e., probability of following links), the probability of visiting a node j is proportional to α divided by the degree of the source node i, plus a random choice equally distributed among all nodes. Depending on the α value, a random walker can model four different behaviors: **(i)** $\alpha = 0.0$: random jumps only, **(ii)** $\alpha \approx 1.0$: navigation over links only, **(iii)** $\alpha = 0.85$: PageRank using the commonly used damping factor for navigating the Web [33], and **(iv)** the empirical PageRank which learns the parameter α from the transitions data. *Number of model parameters:* 1 *if empirical,* 0 *otherwise.*

$$P_{RW} = \alpha \boldsymbol{A} + \frac{(1 - \alpha)}{N} \tag{2.4}$$

Markov Chain (MC). We assume that moving to the next node follows a Markov process. Therefore, the probability of moving to a node j only depends on the current node i. These probabilities represent the maximum likelihood, learned from the transition matrix T. Thus, the probability of visiting node j from node i is proportional to the number of transitions t_{ij}. *Number of model parameters:* $N \times (N - 2)$.

$$P_{MC} = \boldsymbol{T} \tag{2.5}$$

Note that $\boldsymbol{M}$, $\boldsymbol{A}$, and all P_* from Equations (2.1) to (2.5) are right stochastic matrices (i.e., each row must sum to 1).

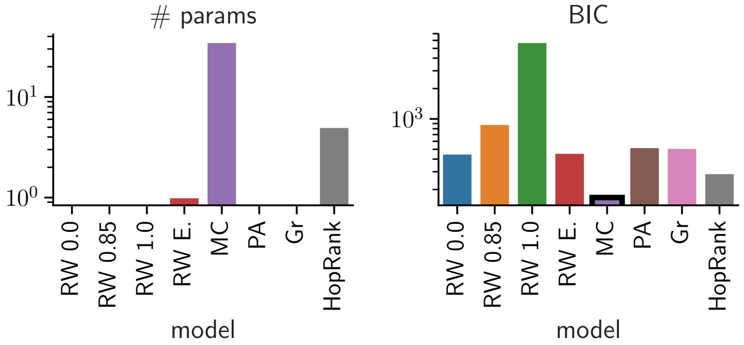


Figure 2.4.: **Results on Synthetic Network from Figure 2.1.** X-axis maps the models at interest. (Left) Number of parameters inferred by each model. (Right) BIC scores: the lower the better at explaining the data. In this example, navigation is best described by Markov chain followed by HopRank.

2.6. Experiments

In this section, we compare the performance of HopRank to the baselines on synthetic and real-world networks.

2.6.1. Model selection

For comparing the models, we employ the Bayesian Information Criterion (BIC) [223] to select the best, i.e., lowest BIC score. BIC evaluates *log-likelihoods LL* (i.e., how likely our transitions are for a given model) and takes into account the *number of model parameters* and *observations* (i.e., # of transitions) to avoid over-fitting.

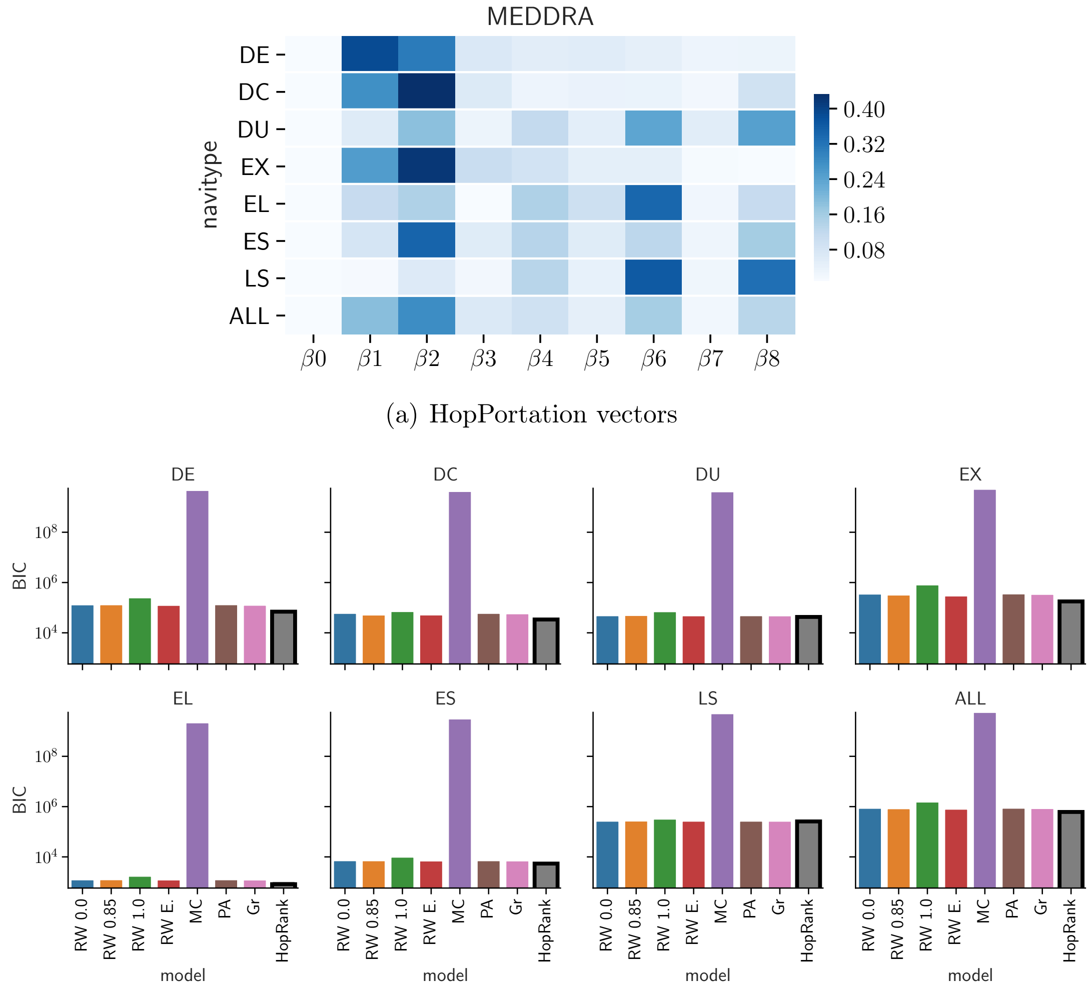
$$BIC = -2 \cdot LL + nparams \cdot log(nobservations), \tag{2.6}$$

$$LL = \sum_{i=1}^{N} \sum_{j=1}^{N} t_{ij} \cdot log(p_{ij}), \tag{2.7}$$

where t_{ij} represents the actual number of transitions from node i to node j, and p_{ij} the probability of transitioning from node i to node j for a given model.

2.6.2. Synthetic network

Setup. The underlying network (structure) is a binary tree composed by $N = 7$ nodes and $|E| = 6$ edges as shown in Figure 2.1(a). Transitions (curved-thick edges) are biased towards 2-hop and 4-hop neighborhoods. These biases are reflected in the HopPortation vector shown in Figure 2.1(b).



(a) HopPortation vectors

(b) Model selection

Figure 2.5.: **Results on MEDDRA.** (a) This heatmap shows the HopPortation vectors learned from the transitions in MEDDRA. Cells represent the probabilities of visiting a certain k-hop neighborhood (column) by a given navigation type (row). In general, 2-hop and 1-hop neighborhoods are more likely to be visited next, regardless of navigation type (ALL). However, distant hops are preferred through *direct URLs* (DU), *external links* (EL), and *local search* (LS). (b) This figure shows the comparison of models across navigation types using BIC scores. We see that HopRank outperforms all baseline models.

Results. Probabilities inferred using Equation (2.1) are depicted in Figure 2.1(c). Figure 2.4 (left) shows the number of parameters inferred by each model. While the Markov chain model (MC) requires 35 parameters, HopRank only needs 5. The empirical PageRank (RW E.) learned a damping factor of $\alpha = 0.01$. This means that users are 1% likely to follow links. In Figure 2.4 (right) we see the comparison of models using BIC scores. In this synthetic network, transitions are best described by the Markov chain model because model parameters (i.e., maximum likelihood) are proportional to the actual transition counts per dyad, and the data structure is very small[7]. In spite of that, HopRank is the second best model and describes navigation better than random (RW 0.0).

2.6.3. Medical Dictionary for Regulatory Activities Terminology (MEDDRA)

Setup. MEDDRA [172] is one of the the largest ontologies in our dataset (see Table 2.1). After pre-processing, its largest connected component (LCC) consists of $23K$ nodes and $43K$ transitions.

Results. Figure 2.5(a) shows the HopPortation vectors learned for each type of navigation in MEDDRA. We see that users mainly navigate through 1, 2, 6, and 8-hop neighbors. For instance, transitions through direct clicks—on a concept (DC), its details (DE) or expand (EX)—mainly follow 1-hop and 2-hop neighbors. However, when transitions are triggered by direct URLs (DU), local search (LS) or external links (EL), users tend to reach distant target nodes (i.e., 6-hop and 8-hop neighbors). Figure 2.5(b) shows the ranking of models according to BIC scores (lower is better). We see that in MEDDRA all types of navigation are best explained by HopRank.

2.6.4. Top11 ontologies in BioPortal

Setup. We fit HopRank and the baseline models to all transitions by ontology and navigation type. These represent $133K$ transitions coming from the 11 ontologies described in Table 2.1.

Results. In Figure 2.6 we highlight the model that explains the number of transitions per ontology and navigation type best (i.e., the model with lowest BIC score). Ontologies are sorted by their number of transitions from CPT (largest) to OMIM (smallest). HopRank outperforms the other models 89% of the time, especially when users browse directly—regardless of the ontology—the tree-like explorer via clicks (DC), details (DE) and expand (EX). When there are not enough observations (i.e., the number of transitions is small), the other models tend to outperform HopRank due to the fact that the other models require fewer parameters and/or it is less likely to find transitions across different k-hop neighborhoods. This is the case for 6 ontologies in certain navigation types. For instance, we found 5 external search (ES) transitions in MESH which are best described by the Gravitational model (Gr). Even though HopRank was a better candidate (i.e., higher log-likelihood), BIC penalized it for having more parameters ($nparams_{HopRank} = 32 > nparams_{Gr} = 0$). Notice that we model

[7] Therefore, number of parameters does not play a very important role in BIC.

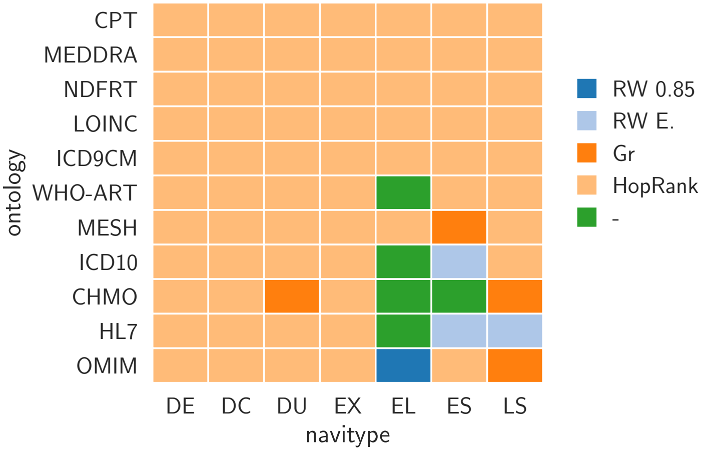


Figure 2.6.: **Model Selection on BioPortal.** This heatmap highlights the model—with lowest BIC score—that best describes the # of transitions per ontology and navigation type. HopRank outperforms the other models 89% of the time, especially when browsing concepts via details (DE), direct click (DC) and expand (EX). When transitions are scarce (i.e., the other 11%), BIC penalizes HopRank since it has more parameters than the other models (except Markov chain).

navigation in ontologies with at least 2 transitions. Ontologies that do not fulfil this condition per navigation type are marked as green cells "-".

2.7. Discussion and Future Work

In this section, we discuss decisions made for data processing, and future directions that can be pursued to improve our results.

Largest connected component (LCC). Surprisingly, ontologies on BioPortal may have multiple connected components. In those cases, only the branch connected to the root *owl:Thing* is shown at first in the tree-like explorer. Disconnected (and hidden) nodes or branches need to be accessed from external pages or local search. For simplicity, we opted to work with the LCC of each ontology with the cost of removing 20% of all transitions. Future work should consider the whole network to study the tradeoffs between number of transitions and random teleportation.

HopRank extensions. More extensions based on network properties or similarity measures between nodes could improve our results. For instance, considering ontologies as directed graphs, and assuming that navigation is not only constrained by distance but also directionality: top-down or bottom-up.

Other types of networks. Even though this paper targets semantic networks, we believe that HopRank can be utilized to model human navigation in other networks, such as the Web or cities. The only assumption required is that users must have background knowledge about the underlying network they are surfing/traveling in.

2.8. Conclusions

In this paper, we introduced the concept of *HopPortation* which states that users—navigating a known or visible network—are biased towards certain k-hop neighborhoods. This is a variation of PageRank, where we assume that teleportation is not fully random but rather distributed non-uniformly across different neighborhoods. We proposed *HopRank*—a biased random walker—to model navigation on semantic networks. Our findings on BioPortal suggest that semantic structure (i.e., shortest path) influences navigation on networks. In particular, users tend to be biased towards certain k-hop neighborhoods depending on the type of navigation. For instance, when manually browsing the tree-like explorer, users tend to hop to nearby concepts, whereas far-away concepts are more likely to be reached by non-browsing types such as external links. These results advance our understanding of how ontologies are actually navigated and consumed, and help to develop and improve user interfaces for ontology exploration.

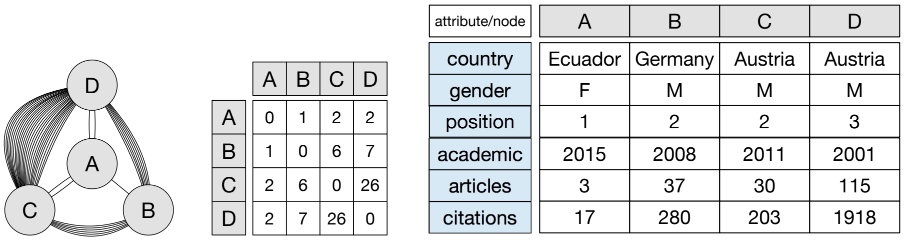
3. Characterizing Edge Formation on Attributed Networks

This chapter introduces Janus, a hypothesis-driven Bayesian approach for understanding edge formation in attributed multigraphs, and it has been published as a full paper in the journal of Applied Network Science, 2017.

Understanding edge formation represents a key question in network analysis. Various approaches have been postulated across disciplines ranging from network growth models to statistical (regression) methods. In this work, we extend this existing arsenal of methods with Janus, a hypothesis-driven Bayesian approach that allows to intuitively compare hypotheses about edge formation in multigraphs. We model the multiplicity of edges using a simple categorical model and propose to express hypotheses as priors encoding our belief about parameters. Using Bayesian model comparison techniques, we compare the relative plausibility of hypotheses which might be motivated by previous theories about edge formation based on popularity or similarity. We demonstrate the utility of our approach on synthetic and empirical data. Janus is relevant for researchers interested in studying mechanisms explaining edge formation in networks from both empirical and methodological perspectives.

3.1. Introduction

Understanding edge formation in networks is a key interest of our research community. For example, social scientists are frequently interested in studying relations between entities within social networks, e.g., how social friendship ties form between actors and explain them based on attributes such as a person's gender, race, political affiliation or age in the network [218]. Similarly, the complex networks community suggests a set of generative network models aiming at explaining the formation of edges focusing on the two core principles of *popularity* and *similarity* [195]. Thus, a series of approaches to study edge formation have emerged including statistical (regression) tools [145, 238] and model-based approaches [134, 195, 239] specifically established in the physics and complex networks communities. Other disciplines such as the computer sciences, biomedical sciences or political sciences use these tools to answer empirical questions;



	A	B	C	D
A	0	1	2	2
B	1	0	6	7
C	2	6	0	26
D	2	7	26	0

attribute/node	A	B	C	D
country	Ecuador	Germany	Austria	Austria
gender	F	M	M	M
position	1	2	2	3
academic	2015	2008	2011	2001
articles	3	37	30	115
citations	17	280	203	1918

(a) Multigraph (b) Adjacency Matrix (c) Node Attributes

Figure 3.1.: **Example.** This example illustrates an unweighted attributed multigraph. (a) Shows a multigraph where nodes represent researchers, and edges scientific articles in which they have collaborated together. (b) Shows the adjacency matrix of the graph, where every cell represents the total number of edges between two nodes. (c) Decodes some attributes per node. For instance, node D shows information about an *Austrian* researcher who started *his* academic career in *2001*. One main objective of Janus is to compare the plausibility of mechanisms derived from attributes for explaining the formation of edges in the graph. For example, here, a hypothesis that researchers have more collaborations if they are from the same country might be more plausible than one that postulates that the multiplicity of edges can be explained based on the relative popularity of authors.

e.g., co-authorship networks[170], wireless networks of biomedical sensors [224], or community structures of political blogs [3].

Problem illustration. Consider for example the network depicted in Figure 3.1. Here, nodes represent authors, and (multiple) edges between them refer to co-authored scientific articles. Node attributes provide additional information on the authors, e.g., their home country and gender. In this setting, an exemplary research question could be: "Can co-authorship be better explained by a mechanism that assumes more collaborations between authors from the *same country* or by a mechanism that assumes more collaborations between authors with the *same gender*?". These and similar questions motivate the main objective of this work, which is to provide a Bayesian approach for understanding how edges emerge in networks based on some characteristics of the nodes or dyads.

While several methods for tackling such questions have been proposed, they come with certain limitations. For example, statistical regression methods based on QAP [122] or mixed-effects models [230] do not scale to large-scale data and results are difficult to interpret. For network growth models [195], it is necessary to find the appropriate model for a given hypothesis about edge formation and thus, it is often not trivial to intuitively compare competing hypotheses. Consequently, we want to extend the methodological toolbox for studying edge formation in networks by proposing a first step towards a hypothesis-driven generative Bayesian framework.

Approach and methods. We focus on understanding edge formation in attributed multigraphs. We are interested in modeling and understanding the multiplicity of edges based on additional network information, i.e., given attributes for the nodes or dyads in the network. Our approach follows a generative storyline. First, we define the model that can characterize the edge formation at interest. We focus on the simple categorical model, from which edges are independently drawn from. Motivated by previous work on sequential data [232], the core idea of our approach is to specify generative hypotheses about how edges emerge in a network. These hypotheses might be motivated by previous theories such as popularity or similarity [195]—e.g., for Figure 3.1 we could hypothesize that authors are more likely to collaborate with each other if they are from the same country. Technically, we elicit these types of hypotheses as beliefs in parameters of the underlying categorical model and encode and integrate them as priors into the Bayesian framework. Using Bayes factors with marginal likelihood estimations allows us to compare the relative plausibility of expressed hypotheses as they are specifically sensitive to the priors. The final output is a ranking of hypotheses based on their plausibility given the data.

Contributions. The main contributions of this work are:

1. We present a first step towards a Bayesian approach for comparing generative hypotheses about edge formation in networks.
2. We provide simple categorical models based on local and global scenarios allowing the comparison of hypotheses for multigraphs.
3. We show that Janus can be easily extended to dyad-attributed multigraphs when multiplex networks are provided.
4. We demonstrate the applicability and plausibility of Janus based on experiments on synthetic and empirical data, as well as by comparing it to the state-of-the-art QAP.
5. We make an implementation of this approach openly available on the Web [73].

Structure. This paper is structured as follows: First, we start with an overview of some existing research on modeling and understanding edge formation in networks in Section 3.2. We present some background knowledge required in this work in Section 3.3 to then explain step-by-step Janus in Section 3.4. Next, we show Janus in action and the interpretation of results by running four different experiments on synthetic and empirical data in Section 3.5. In Section 3.6 we suggest a fair comparison of Janus with the Quadratic Assignment Procedure (QAP) for testing hypotheses on dyadic data. In Section 3.6 we highlight some important caveats for further improvements. Finally, we conclude in Section 3.7 by summarizing the contributions of our work.

3.2. Related Work

We provide a broad overview of research on modeling and understanding edge formation in networks; i.e., *edge formation models* and *hypothesis testing on networks.*

Edge formation models. A variety of models explaining underlying mechanisms of *network formation* have been proposed. Here, we focus on models explaining linkage between dyads beyond structure by incorporating node attribute information. Prominently, the *stochastic blockmodel* [134] aims at producing and explaining communities by accounting for node correlation based on attributes. The *attributed graph* [202] models network structure and node attributes by learning the attribute correlations in the observed network. Furthermore, the *multiplicative attributed graph* [138] takes into account attribute information from nodes to model network structure. This model defines the probability of an edge as the product of individual attribute link formation affinities. *Exponential random graph models* [211] (also called the p^* class of models) represent graph distributions with an exponential linear model that uses feature-structure counts such as reciprocity, k-stars and k-paths. In this line of research, *p1 models* [117] consider expansiveness (sender) and popularity (receiver) as fixed effects associated with unique nodes in the network [100] in contrast to the *p2 models* [211] which account for random effects and assume dyadic independence conditionally to node-level attributes. While many of these works focus on binary relationships, [266] proposes an unsupervised model to estimate continuous-valued relationship strength for links from interaction activity and user similarity in social networks. Recently, the work in [142] has shown that connections in one layer of a multiplex can be accurately predicted by utilizing the hyperbolic distances between nodes from another layer in a hidden geometric space.

Hypothesis testing on networks. Previous works have implemented different techniques to test hypotheses about network structure. For instance, the work in [180] proposes an algorithm to determine whether two observed networks are significantly different. Another branch of research has specifically focused on dyadic relationships utilizing regression methods accounting for interdependencies in network data. Here, we find *Multiple Regression Quadratic Assignment Procedure* (MRQAP) [145] and its predecessor QAP [122] which permute nodes in such a way that the network structure is kept intact; this allows to test for significance of effects. *Mixed-effects models* [230] add random effects to the models allowing for variation to mitigate non-independence between responses (edges) from the same subject (nodes) [264]. Based on the *quasi essential graph* the work in [190] proposes to compare two graphs (i.e., Bayesian networks) by testing and comparing multiple hypotheses on their edges. Recently, *generalized hypergeometric* *ensembles* [41] have been proposed as a framework for model selection and statistical hypothesis testing of finite, directed and weighted networks that allow to encode several topological patterns such as block models where homophily plays an important role in linkage decision. In contrast to our work, neither of these approaches is based on Bayesian hypothesis testing, which avoids some fundamental issues of classic frequentist statistics.

3.3. Background

In this paper, we focus on both *node-attributed* and *dyad-attributed* multigraphs with *unweighted edges without own identity*. That means, each pair of nodes or dyad can be connected by multiple indistinguishable edges, and there are features for the individual nodes or dyads available.

Node-attributed multigraphs. We formally define this as: Let $G = (V, E, F)$ be an unweighted attributed multigraph with $V = (v_1, \dots, v_n)$ being a list of nodes, $E = \{(v_i, v_j)\} \in V \times V$ a multiset of either directed or undirected edges, and a set of feature vectors $F = (f_1, \dots, f_n)$. Each feature vector $f_i = (f_i[1], ..., f_i[c])^T$ maps a node v_i to c (numeric or categorical) attribute values. The graph structure is captured by an adjacency matrix $M_{n\times n} = (m_{ij})$, where m_{ij} is the multiplicity of edge (v_i, v_j) in E (i.e., number of edges between nodes v_i and v_j). By definition, the total number of multiedges is $l = |E| = \sum_{ij} m_{ij}$.

Figure 3.1(a) shows an example unweighted attributed multigraph: nodes represent authors, and undirected edges represent co-authorship in scientific articles. The adjacency matrix of this graph—counting for multiplicity of edges—is shown in Figure 3.1(b). Feature vectors (node attributes) are described in Figure 3.1(c). Thus, for this particular case, we account for $n = 4$ nodes, $l = 44$ multiedges, and $c = 6$ attributes.

Dyad-attributed networks. As an alternative to attributed nodes, we also consider multigraphs, in which each dyad (pair of nodes) is associated with a set of features $\hat{F} = (\hat{f}_{11}, \dots, \hat{f}_{nn})$. Each feature vector $\hat{f}_{ij} = (\hat{f}_{ij}[1], ..., \hat{f}_{ij}[c])^T$ maps the pair of node (v_i, v_j) to c (numeric or categorical) attribute values. The values of each feature can be represented in a separate $n \times n$ matrix. As an important special case of dyad-attributed networks, we study *multiplex networks*. In these networks, all dyad features are integer-valued. Thus, each feature can be interpreted as (or can be derived from) a separate multigraph over the same set of nodes. In our setting, the main idea is then to try and explain the occurrence of a multiset of edges E in one multigraph G with nodes V by using other multigraphs $\hat{G}$ on the same node set.

Bayesian hypothesis testing. Our approach compares hypotheses on edge formation based on techniques from Bayesian hypothesis testing [147, 232]. The elementary Bayes' theorem states for parameters θ, given data D and a hypothesis H that:

$$\overbrace{P(\theta|D,H)}^{\text{posterior}} = \frac{\overbrace{P(D|\theta,H)}^{\text{likelihood}}\overbrace{P(\theta|H)}^{\text{prior}}}{\underbrace{P(D|H)}_{\text{marginal likelihood}}} \tag{3.1}$$

As observed data D, we use the adjacency matrix M, which encodes edge counts. θ refers to the model parameters, which in our scenario correspond to the probabilities of individual edges. H denotes a hypothesis under investigation. The *likelihood* describes, how likely we observe data D given parameters θ and a hypothesis H. The *prior* is the distribution of parameters we believe in before seeing the data; in other words, the prior encodes our hypothesis H. The *posterior* represents an adjusted distribution of parameters after we observe D. Finally, the *marginal likelihood* (also called *evidence*) represents the probability of the data D given a hypothesis H.

In our approach, we exploit the sensitivity of the marginal likelihood on the prior to compare and rank different hypotheses: more plausible hypotheses imply higher evidence for data D. Formally, *Bayes Factors* can be employed for comparing two hypotheses. These are computed as the ratio between the respective marginal likelihood scores. The strength of a Bayes factor can be judged using available interpretation tables [135]. While in many cases determining the marginal likelihood is computationally challenging and requires approximate solutions, we can rely on exact and fast-to-compute solutions in the models employed in this paper.

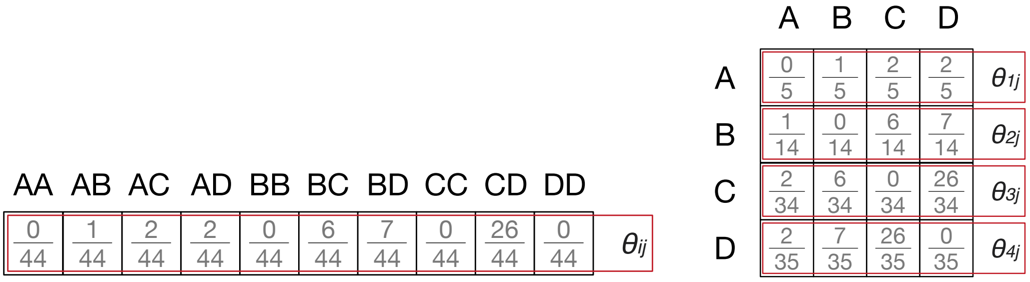
3.4. Approach

In this section, we describe the main steps towards a hypothesis-driven Bayesian approach for understanding edge formation in unweighted attributed multigraphs. To that end, we propose intuitive models for edge formation (Section 3.4.1), a flexible toolbox to formally specify beliefs in the model parameters (Section 3.4.2), a way of computing proper (Dirichlet) priors from these beliefs (Section 3.4.2), computation of the marginal likelihood in this scenario (Section 3.4.3), and guidelines on how to interpret the results (Section 3.4.4). We subsequently discuss these issues one-by-one.

3.4.1. Generative edge formation models

We propose two variations of our approach, which employ two different types of generative edge formation models in multigraphs.

Global model. First, we utilize a simple *global model*, in which a fixed number of graph edges are randomly and independently drawn from the set of all potential edges in the graph G by sampling with replacement. Each edge (v_i, v_j) is sampled from a *categorical distribution* with parameters $\theta_{ij}, 1 \leq i \leq n, 1 \leq j \leq n, \forall ij$:



AA	AB	AC	AD	BB	BC	BD	CC	CD	DD	
$\frac{0}{44}$	$\frac{1}{44}$	$\frac{2}{44}$	$\frac{2}{44}$	$\frac{0}{44}$	$\frac{6}{44}$	$\frac{7}{44}$	$\frac{0}{44}$	$\frac{26}{44}$	$\frac{0}{44}$	θ_{ij}

(a) Global

	A	B	C	D	
A	$\frac{0}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	θ_{1j}
B	$\frac{1}{14}$	$\frac{0}{14}$	$\frac{6}{14}$	$\frac{7}{14}$	θ_{2j}
C	$\frac{2}{34}$	$\frac{6}{34}$	$\frac{0}{34}$	$\frac{26}{34}$	θ_{3j}
D	$\frac{2}{35}$	$\frac{7}{35}$	$\frac{26}{35}$	$\frac{0}{35}$	θ_{4j}

(b) Local

Figure 3.2.: **Multigraph models.** This figure shows two ways of modeling the undirected multigraph shown in Figure 3.1. That is, (a) global or graph-based model to model the whole graph as a single distribution. (b) Local or neighbor-based model to model each node as a separate distribution.

$\sum_{ij} \theta_{ij} = 1$: $(v_i, v_j) \sim Categorical(\theta_{ij})$. This means that each edge is associated with one probability θ_{ij} of being drawn next. Figure 3.2(a) shows the maximum likelihood global model for the network shown in Figure 3.1. Since this is an undirected graph, inverse edges can be ignored resulting in $n(n+1)/2$ potential edges/parameters.

Local models. As an alternative, we can also focus on a *local level*. Here, we model to which other node a specific node v will connect *given that any new edge starting from* v is formed. We implement this by using a set of n separate models for the outgoing edges of the ego-networks (i.e., the 1-hop neighborhood) of each of the n nodes. The ego-network model for node v_i is built by drawing randomly and independently a number of nodes v_j by sampling with replacement and adding an edge from v_i to these nodes. Each node v_j is sampled from a *categorical distribution* with parameters $\theta_{ij}, 1 \leq i \leq n, 1 \leq j \leq n, \forall i : \sum_j \theta_{ij} = 1$: $v_j \sim Categorical(\theta_{ij})$. The parameters θ_{ij} can be written as a matrix; the value in cell (i, j) specifies the probability that a new formed edge with source node v_i will have the destination node v_j. Thus, all values within one row always sum up to one. Local models can be applied for undirected and directed graphs (cf. also discussion in Section 3.6). In the directed case, we model only the outgoing edges of the ego-network. Figure 3.2(b) depicts the maximum likelihood local models for our introductory example.

3.4.2. Hypothesis elicitation

The main idea of our approach is to encode our beliefs in edge formation as Bayesian priors over the model parameters. As a common choice, we employ Dirichlet distributions as the *conjugate priors* of the categorical distribution.

Thus, we assume that the model parameters θ are drawn from a Dirichlet distribution with hyperparameters α: $\theta \sim Dir(\alpha)$. Similar to the model parameters themselves, the Dirichlet prior (or multiple priors for the local models) can be specified in a matrix. We will choose the parameters α in such a way that they reflect a specific belief about edge formation. For that purpose, we first specify matrices that formalize these beliefs, then we compute the Dirichlet parameters α from these beliefs.

Constructing belief matrices

We specify hypotheses about edge formation as *belief matrices* $B = b_{ij}$. These are $n \times n$ matrices, in which each cell $b_{ij} \in \mathbb{R}$ represents a belief of having an edge from node v_i to node v_j. To express a belief that an edge occurs more often (compared to other edges) we set b_{ij} to a higher value.

Node-attributed multigraphs. In general, users have a large freedom to generate belief matrices. However, typical construction principles are to assume that nodes with specific attributes are more *popular* and thus edges connecting these attributes receive higher multiplicity, or to assume that nodes that are *similar* with respect to one or more attributes are more likely to form an edge, cf. [195]. Ideally, the elicitation of belief matrices is based on existing theories.

For example, based on the information shown in Figure 3.1, one could “believe” that two authors collaborate *more frequently* together if: (1) they both are from the same country, (2) they share the same gender, (3) they have high positions, or (4) they are popular in terms of number of articles and citations. We capture each of these beliefs in one matrix. One implementation of the matrices for our example beliefs could be:

- B_1 (same country): $b_{ij} := 0.9$ if $f_i[\text{country}] = f_j[\text{country}]$ else 0.1
- B_2 (same gender): $b_{ij} := 0.9$ if $f_i[\text{gender}] = f_j[\text{gender}]$ else 0.1
- B_3 (hierarchy): $b_{ij} := f_i[\text{position}] \cdot f_j[\text{position}]$
- B_4 (popularity): $b_{ij} := f_i[\text{articles}] + f_j[\text{articles}] + f_i[\text{citations}] + f_j[\text{citations}]$

Figure 3.3(a) shows the matrix representation of belief B_1, and Figure 3.3(b) its respective row-wise normalization for the local model case. While belief matrices are identically structured for local and global models, the ratio between parameters in different rows is crucial for the global model, but irrelevant for local ones.

Dyad-attributed networks. For the particular case of Dyad-Attributed networks, beliefs are described using the underlying mechanisms of secondary multigraphs. For instance, a *co-authorship* network—where every node represents an author with no additional information or attribute—could be explained by a *citation* network under the hypothesis that if two authors frequently cite each other,

	A	B	C	D
A	0.0	0.1	0.1	0.1
B	0.1	0.0	0.1	0.1
C	0.1	0.1	0.0	0.9
D	0.1	0.1	0.9	0.0

(a) Belief matrix B_1

	A	B	C	D
A	0.00	0.33	0.33	0.33
B	0.33	0.00	0.33	0.33
C	0.09	0.09	0.00	0.82
D	0.09	0.09	0.82	0.00

(b) Normalized B_1

	A	B	C	D
A	1.00	2.33	2.33	2.33
B	2.33	1.00	2.33	2.33
C	1.36	1.36	1.00	4.27
D	1.36	1.36	4.27	1.00

(c) Prior $\kappa = 4$

Figure 3.3.: **Prior belief.** This figure illustrates the three main phases of prior elicitation. That is, (a) a matrix representation of belief B_1, where authors are more likely to collaborate with each other if they are from the same country. (b) B_1 normalized row-wise using the local model interpretation. (c) Prior elicitation for $\kappa = 4$; i.e., $\alpha_{ij} = \frac{b_{ij}}{Z} \times \kappa + 1$.

they are more likely to also co-author together. Thus, the adjacency (feature) matrices ($\hat{F}$) of secondary multigraphs can be directly used as belief matrices $B = (b_{ij})$. However, we can express additional beliefs by transforming the matrices. As an example, we can formalize the belief that the presence of a feature tends to inhibit the formation of edges in the data by setting $b_{ij} := -sigm(f_{ij})$, where *sigm* is a sigmoid function such as the logistic function.

Eliciting a Dirichlet prior

In order to obtain the hyperparameters α of a prior Dirichlet distribution, we utilize the pseudo-count interpretation of the parameters α_{ij} of the Dirichlet distribution, i.e., a value of α_{ij} can be interpreted as $\alpha_{ij} - 1$ previous observations of the respective event for $\alpha_{ij} \geq 1$. We distribute pseudo-counts proportionally to a belief matrix. Consequently, the hyperparameters can be expressed as: $\alpha_{ij} = \frac{b_{ij}}{Z} \times \kappa + 1$, where κ is the concentration parameter of the prior. The normalization constant Z is computed as the sum of all entries of the belief matrix in the global model, and as the respective row sum in the local case. We suggest setting $\kappa = n \times k$ for the local model, $\kappa = n^2 \times k$ for the global case, and $k = \{0, 1, ..., 10\}$. A high value of κ expresses a strong belief in the prior parameters. A similar alternative method to obtain Dirichlet priors is the *trial roulette method* [232]. For the global model variation, all α values are parameters for the same Dirichlet distribution, whereas in the local model variation, each row parametrizes a separate Dirichlet distribution.

3.4.3. Computation of the marginal likelihood

For comparing the relative plausibility of hypotheses, we use the marginal likelihood. This is the aggregated likelihood over all possible values of the parameters

θ weighted by the Dirichlet prior. For our set of local models we can calculate them as:

$$P(D|H) = \prod_{i=1}^{n} \frac{\Gamma(\sum_{j=1}^{n} \alpha_{ij})}{\Gamma(\sum_{j=1}^{n} \alpha_{ij} + m_{ij})} \prod_{j=1}^{n} \frac{\Gamma(\alpha_{ij} + m_{ij})}{\Gamma(\alpha_{ij})} \tag{3.2}$$

Recall, α_{ij} encodes our prior belief connecting nodes v_i and v_j in G, and m_{ij} are the actual edge counts. Since we evaluate only a single model in the global case, the product over rows i of the adjacency matrix can be removed, and we obtain:

$$P(D|H) = \frac{\Gamma(\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij})}{\Gamma(\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij} + m_{ij})} \prod_{i=1}^{n} \prod_{j=1}^{n} \frac{\Gamma(\alpha_{ij} + m_{ij})}{\Gamma(\alpha_{ij})} \tag{3.3}$$

Equation (3.3) holds for directed networks. In the undirected case, indices j go from i to n accounting for only half of the matrix including the diagonal to avoid inconsistencies. For a detailed derivation of the marginal likelihood given a Dirichlet-Categorical model see [233, 250]. For both models we focus on the log-marginal likelihoods in practice to avoid underflows.

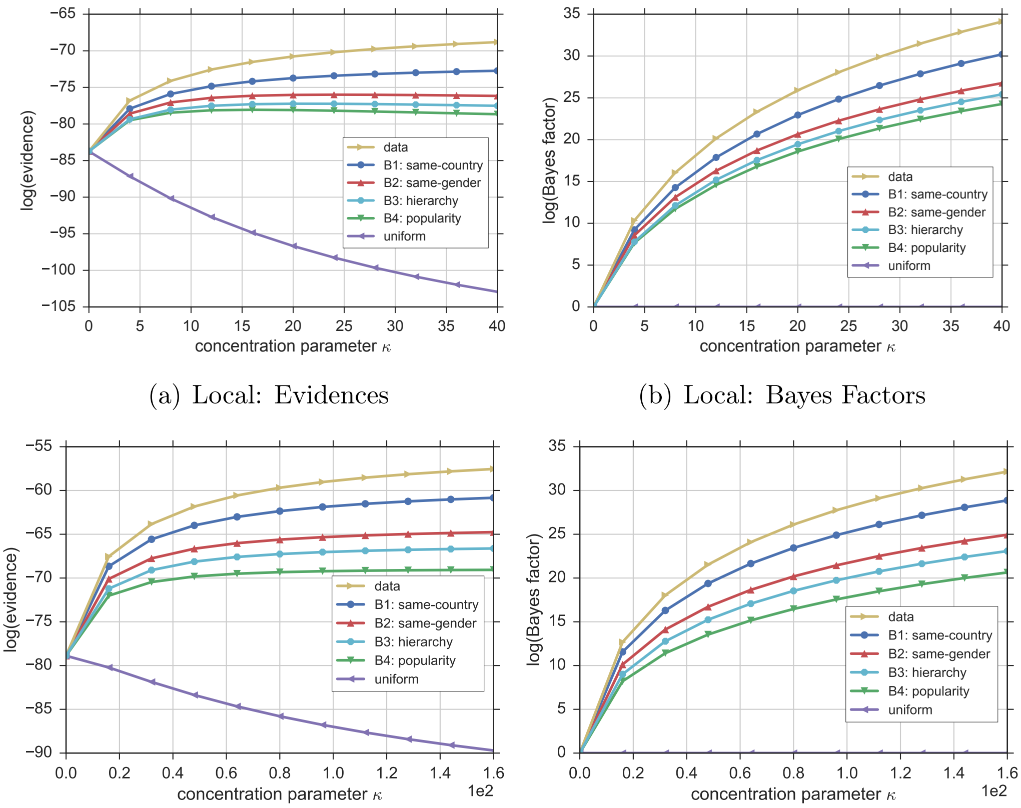
Bayes factor. Formally, we compare the relative plausibility of hypotheses by using so-called *Bayes factors* [135], which simply are the ratios of the marginal likelihoods for two hypotheses H_1 and H_2. If it is positive, the first hypothesis is judged as more plausible. The strength of the Bayes factor can be checked in an interpretation table provided by Kass and Raftery [135].

3.4.4. Application of the method and interpretation of results

We now showcase an example application of our approach featuring the network shown in Figure 3.1, and demonstrate how results can be interpreted.

Hypotheses. We compare four hypotheses (represented as belief matrices) B_1, B_2, B_3, and B_4 elaborated in Section 3.4.2. Additionally, we use the *uniform* hypothesis as a *baseline* which assumes that all nodes are equally likely to connect to each other, i.e., $b_{ij} = 1$ for all i, j. Hypotheses that are not more plausible than the uniform cannot be assumed to capture relevant underlying mechanisms of edge formation. We also use the *data* hypothesis as an upper bound for comparison, which employs the observed adjacency matrix as belief: $b_{ij} = m_{ij}$.

Calculation and visualization. For each hypothesis H and every κ, we can elicit the Dirichlet priors (cf. Section 3.4.2), determine the aggregated marginal likelihood (cf. Section 3.4.3), and compare the plausibility of hypotheses in contrast to the uniform hypothesis at the same κ by calculating the logarithm of the Bayes factor as $log(P(D|H)) - log(P(D|H_{uniform}))$. We suggest two ways of visualizing the results, i.e., plotting the marginal likelihood values, and showing the Bayes factors on the y-axis as shown in Figures 3.4(a) and 3.4(b) respectively for



(a) Local: Evidences

(b) Local: Bayes Factors

(c) Global: Evidences

(d) Global: Bayes Factors

Figure 3.4.: **Ranking of hypotheses for the introductory example.** (a,b) Represent results using the local model and (c,d) results of the global model. Rankings can be visualized using (a,c) the marginal likelihood or evidence (y-axis), or (b,d) using Bayes factors (y-axis) by setting the uniform hypothesis as a baseline to compare with; higher values refer to higher plausibility. The x-axis depicts the concentration parameter κ. For this example, from an individual perspective (local model) authors from the multigraph shown in Figure 3.1 appear to prefer to collaborate more often with researchers of the same country rather than due to popularity (i.e., number of articles and citations). In this particular case, the same holds for the global model. Note that all hypotheses outperform the uniform, meaning that they all are reasonable explanations of edge formation for the given graph.

the local model. In both cases, the x-axis refers to the concentration parameter κ. While the visualization showing directly the marginal likelihoods carries more information, visualizing Bayes factors makes it easier to spot smaller differences between the hypotheses.

Interpretation. Every line in Figures 3.4(a) to 3.4(d) represents a hypothesis using the local (top) and global models (bottom). In Figures 3.4(a) and 3.4(c), higher evidence values mean higher plausibility. Similarly, in Figures 3.4(b) and 3.4(d) positive Bayes factors mean that for a given κ, the hypothesis is judged to be more plausible than the uniform baseline hypothesis; here, the relative Bayes factors also provide a ranking. If evidences or Bayes factors are increasing with κ, we can interpret this as further evidence for the plausibility of expressed hypothesis as this means that the more we believe in it, the higher the Bayesian approach judges its plausibility. As a result for our example, we see that the hypothesis believing that two authors are more likely to collaborate if they are from the same country is the most plausible one (after the data hypothesis). In this example, all hypotheses appear to be more plausible than the baseline in both local and global models, but this is not necessarily the case in all applications.

3.5. Experiments

We demonstrate the utility of our approach on both synthetic and empirical networks.

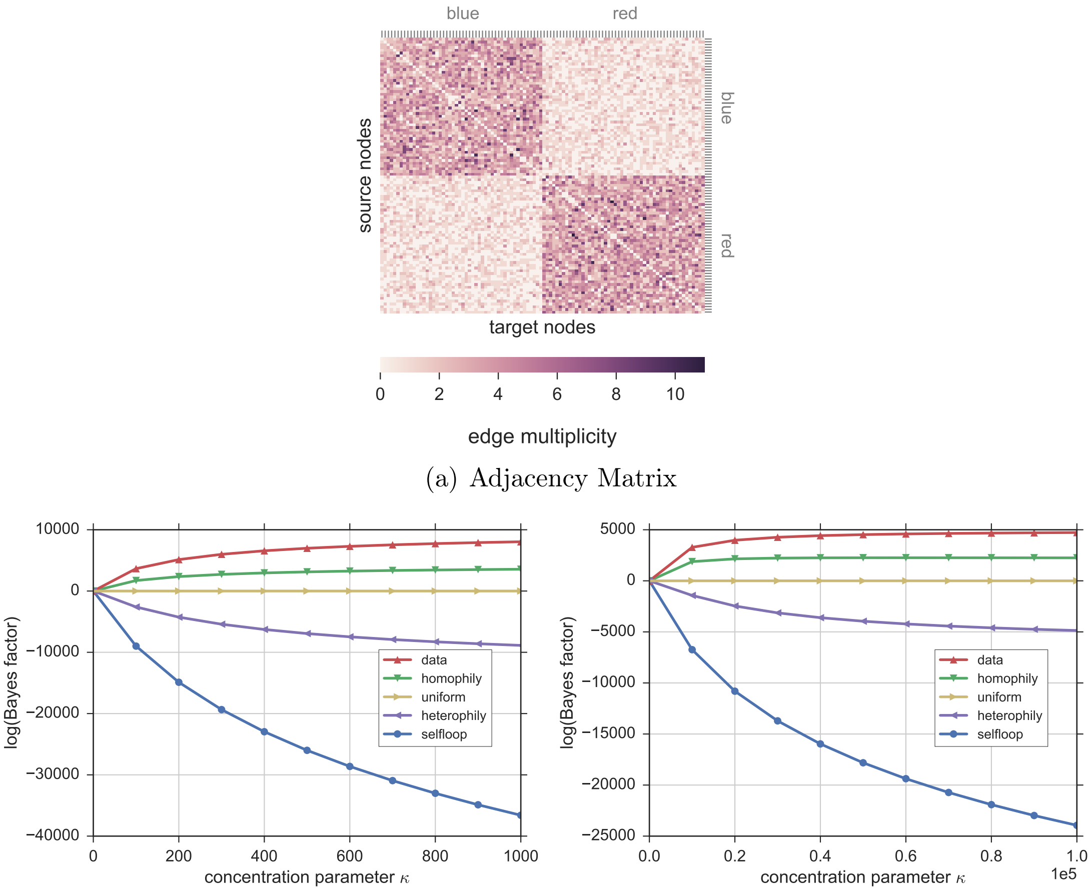
3.5.1. Synthetic attributed multigraph

We start with experiments on a synthetic attributed multigraph. Here, we control the underlying mechanisms of how edges in the network emerge and thus, expect these also to be good hypotheses for our approach.

Network. The network contains 100 nodes where each node is assigned one of two colors with uniform probability. For each node, we then randomly drew 200 undirected edges where each edge connects randomly with probability $p = 0.8$ to a different node of the same color, and with $p = 0.2$ to a node of the opposite color. The adjacency matrix of this graph is visualized in Figure 3.5(a).

Hypotheses. In addition to the uniform baseline hypothesis, we construct two intuitive hypotheses based on the node color that express belief in possible edge formation mechanics. First, the *homophily* hypothesis assumes that nodes of the same color are more likely to have more edges between them. Therefore, we arbitrary set belief values b_{ij} to 80 when nodes v_i and v_j are of the same color, and 20 otherwise. Second, the *heterophily* hypothesis expresses the opposite behavior; i.e., $b_{ij} = 80$ if the color of nodes v_i and v_j are different, and 20 otherwise. An additional *self-loop* hypothesis only believes in self-connections (i.e., diagonal of adjacency matrix).

Results. Figures 3.5(b) and 3.5(c) show the ranking of hypotheses based on their Bayes factors compared to the uniform hypothesis for the local and global models respectively. Clearly, in both models the homophily hypothesis is judged as the most plausible. This is expected and corroborates the fact that network



(a) Adjacency Matrix

(b) Local: Bayes Factors

(c) Global: Bayes Factors

Figure 3.5.: **Ranking of hypotheses for synthetic attributed multigraph.** In (a), we show the adjacency matrix of a 100-node 2-color random multigraph with a node correlation of 80% for nodes of the same color and 20% otherwise. One can see the presence of homophily based on more connections between nodes of the same color; the diagonal is zero as there are no self-connections. In (b,c), we show the ranking of hypotheses based on Bayes factors when compared to the uniform hypothesis for the local and global models respectively. As expected, in general the homophily hypothesis explains the edge formation best (positive Bayes factor and close to the data curve), and while the heterophily and self-loop hypotheses provide no good explanations for edge formation in both local and global cases—they show negative Bayes factors.

connections are biased towards nodes of the same color. The heterophily and self-loop hypotheses show negative Bayes factors; thus, they are not good hypotheses about edge formation in this network. Due to the fact that the multigraph lacks of self-loops, the self-loop hypothesis decreases very quickly with increasing strength of belief κ.

3.5.2. Synthetic multiplex network

In this experiment, we control the underlying mechanisms of how edges in a dyad-attributed multigraph emerge using multiple multigraphs that share the same nodes with different link structure (i.e., multiplex) and thus, expect these also to be good hypotheses for Janus.

Network. The network is a *configuration model* graph [189] with parameters $n = 100$ (i.e., number of nodes) and degree sequence $\overrightarrow{k} = k_i$ drawn from a power law distribution of length n, where k_i is the degree of node v_i. The adjacency matrix of this graph is visualized in Figure 3.6(a).

Hypotheses. Besides the uniform hypothesis, we include ten more hypotheses derived from the original adjacency matrix of the configuration model graph where only a percentage of edges ϵ gets shuffled. The bigger the ϵ the less plausible the hypothesis since more shuffles can modify drastically the original network.

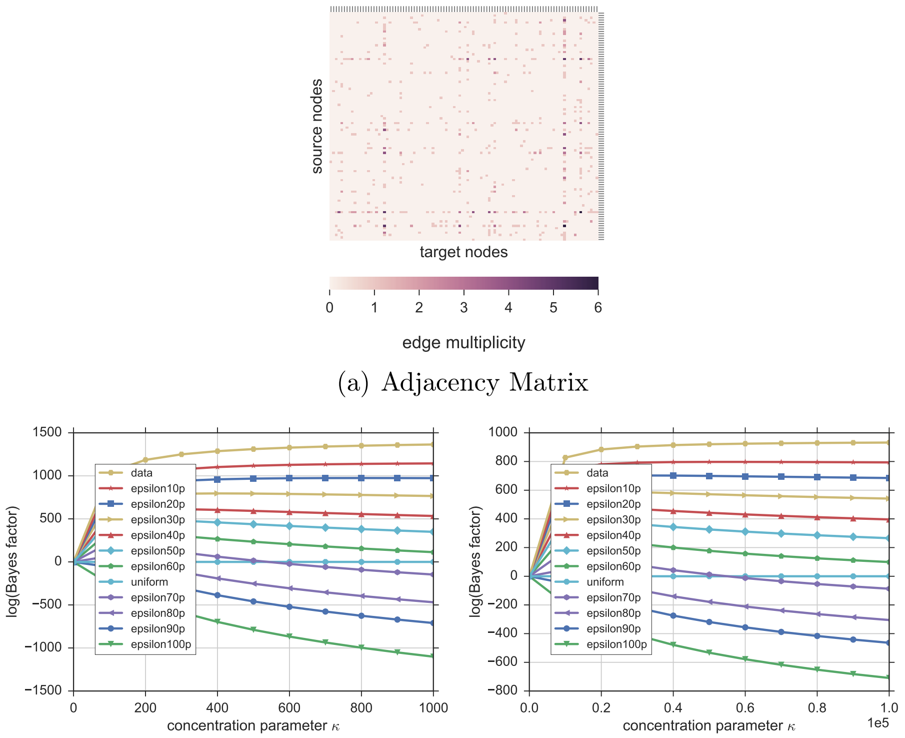
Results. Figures 3.6(b) and 3.6(c) show the ranking of hypotheses based on their Bayes factors compared to the uniform hypothesis for the local and global model respectively. In general, hypotheses are ranked as expected, from small to big values of ϵ. For instance, the *epsilon10p* hypothesis explains best the *configuration model* graph—represented in Figure 3.6(a)—since it only shuffles 10% of all edges (i.e., 10 edges). On the other hand the *epsilon100p* hypothesis shows the worst performance (i.e., Bayes factor is negative and far from the data curve) since it shuffles all edges, therefore it is more likely to be different than the original network.

3.5.3. Empirical attributed multigraph

Here, we focus on a real-world contact network based on wearable sensors.

Network. We study a network capturing interactions of 5 households in rural Kenya between April 24 and May 12, 2012 [139, 240]. The undirected unweighted multigraph contains 75 nodes (persons) and 32 643 multiedges (contacts) which we aim to explain. For each node, we know information such as gender and age (encoded into 5 age intervals). Interactions exist within and across households. Figure 3.7(a) shows the adjacency matrix (i.e., number of contacts between two people) of the network. Household membership of nodes (rows/columns) is shown accordingly.

Hypotheses. We investigate edge formation by comparing—next to the uniform baseline hypothesis—four hypotheses based on node attributes as prior beliefs. (i) The *similar age* hypothesis expresses the belief that people of similar age are more likely to interact with each other. Entries b_{ij} of the belief matrix B are set to the inverse age distance between members: $\frac{1}{1+abs(f_i[age]-f_j[age])}$. (ii) The *same*



(a) Adjacency Matrix

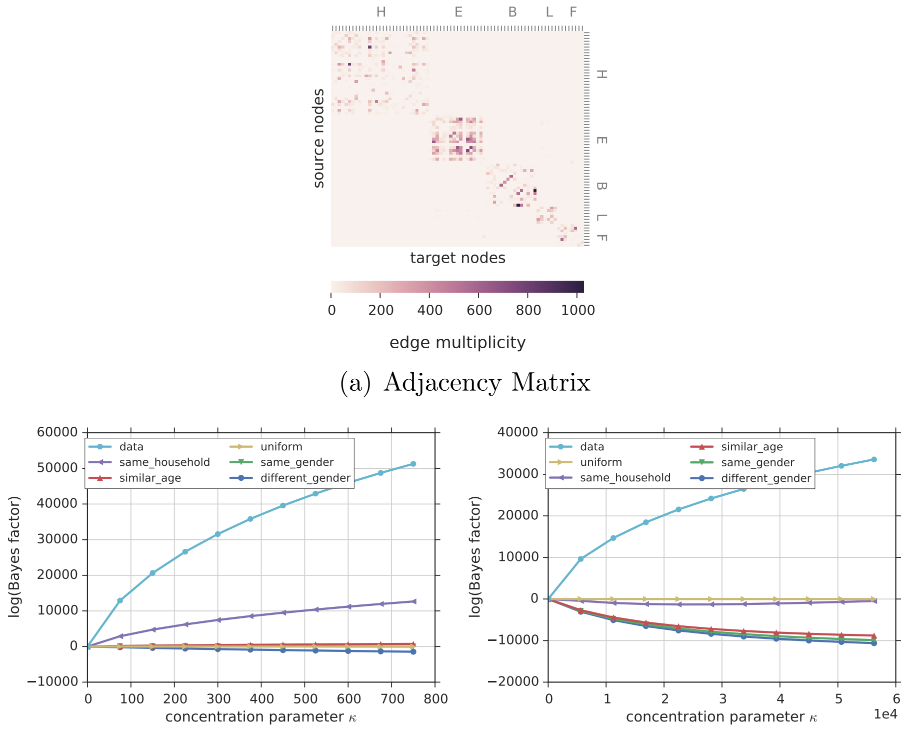
(b) Local: Bayes Factors

(c) Global: Bayes Factors

Figure 3.6.: **Ranking of hypotheses for synthetic multiplex network.** In (a), we show the adjacency matrix of a configuration model graph of 100 nodes and power-law distributed degree sequence. In (b,c), the ranking of hypotheses is shown for the local and global model, respectively. As expected, hypotheses are ranked from small to big values of ϵ since small values represent only a few changes in the original adjacency matrix of the graph. Both models show that when the original graph changes at least 70% of its edges the new graph cannot be explained better than random (i.e., uniform).

household hypothesis believes that people are more likely to interact with people from the same household. We arbitrarily set b_{ij} to 80 if person v_i and person v_j belong to the same household, and 20 otherwise. (iii) With the *same gender* hypothesis we hypothesize that the number of same-gender interactions is higher than the different-gender interactions. Therefore, every entry b_{ij} of B is set to 80 if persons v_i and v_j are of the same gender, and 20 otherwise. Finally, (iv) the *different gender* hypothesis believes that it is more likely to find different-gender than same-gender interactions; b_{ij} is set to 80 if person v_i has the opposite gender of person v_j, and 20 otherwise.

Results. Results shown in Figures 3.7(b) and 3.7(c) show the ranking of hypotheses based on Bayes factors using the uniform hypothesis as baseline for the local and global model, respectively. The local model Figure 3.7(b) indicates that the



(a) Adjacency Matrix

(b) Local: Bayes Factors

(c) Global: Bayes Factors

Figure 3.7.: **Ranking of hypotheses for Kenya contact network.** (a) Shows the adjacency matrix of the network with node ordering according to household membership. Darker cells indicate more contacts. (b,c) Display the ranking of hypotheses based on Bayes factors, using the uniform hypothesis as baseline for the local and global model respectively. Using the local model (b) the *same household* hypothesis is ranked highest followed by the *similar age* hypothesis which also provides positive Bayes Factors. On the other hand, the *same* and *different gender* hypotheses are less plausible than the baseline (uniform hypothesis) in both the local and global case. In the global case (c) all hypotheses are bad representations of edge formation in the Kenya contact network. This is due to the fact that interactions are very sparse, even within households. Results are consistent for all κ.

same household hypothesis explains the data the best, since it has been ranked first and it is more plausible than the uniform. The *similar age* hypothesis also indicates plausibility due to positive Bayes factors. Both the *same* and *different gender* hypotheses show negative Bayes factors when compared to the uniform hypothesis suggesting that they are not good explanations of edge formation in this network. This gives us a better understanding of potential mechanisms producing underlying edges. People prefer to contact people from the same household and similar age, but not based on gender preferences. Additional experiments could further refine these hypotheses (e.g., combining them). In the general case of the global model in Figure 3.7(c), all hypotheses are bad explanations of the Kenya network. However, the *same-household* hypothesis tends to go upfront the uniform for higher values of κ, but still far form the data curve. This happens due to the fact that the interaction network is very sparse (even within same households), thus, any hypothesis with a dense belief matrix will likely fall below or very close to the uniform.

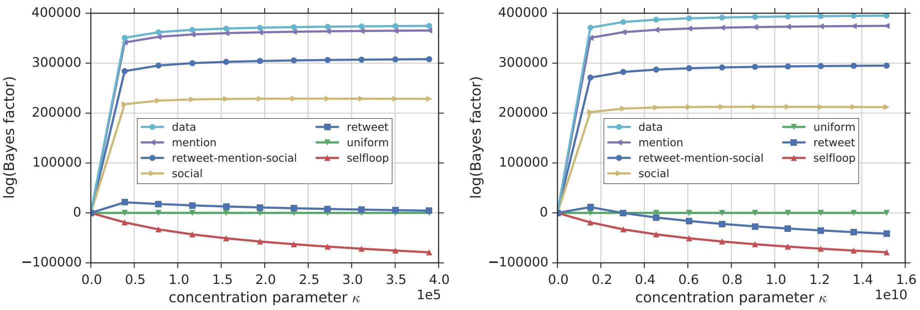
3.5.4. Empirical multiplex network

This empirical dataset consists of four real-world social networks, each of them extracted from Twitter interactions of a particular set of users.

Network. We obtained the Higgs Twitter dataset from SNAP [237]. This dataset was built upon the interactions of users regarding the discovery of a new particle with the features of the elusive Higgs boson on the 4^{th} of July 2012 [57]. Specifically, we are interested on characterizing edge formation in the *reply network*, a directed unweighted multigraph which encodes the replies that a person v_i sent to a person v_j during the event. This graph contains 38 918 nodes and 36 902 multiedges (if all edges from the same dyad are merged it accounts for 32 523 weighted edges).

Hypotheses. We aim to characterize the reply network by incorporating other networks—sharing the same nodes but different network structure—as prior beliefs. In this way we can learn whether the interactions present in the reply network can be better explained by a retweet or mentioning or following (social) network. The *retweet* hypothesis expresses our belief that the number of replies is proportional to the number of retweets. Hence, beliefs b_{ij} are set to the number of times user v_i retweeted a post from user v_j. Similar as before, the *mention* hypothesis states that the number of replies is proportional to the number of mentions. Therefore, every entry b_{ij} is set to the number of times user v_i mentioned user v_j during the event. The *social* hypothesis captures our belief that users are more likely to reply to their friends (in the Twitter jargon: followees or people they follow) than to the rest of users. Thus, we set b_{ij} to 1 if user v_i follows user v_j and 0 otherwise. Finally, we combine all the above networks to construct the *retweet-mention-social* hypothesis which captures all previous hypotheses at once. In other words, it reflects our belief that users are more likely to reply to their friends and (at the same time) the number of replies is proportional to the number of retweets and mentions. Therefore the adjacency matrix for this hypothesis is simply the sum of the three networks described above.

Results. The results shown in Figure 3.8 suggest that the *mention* hypothesis explains the reply network very well, since it has been ranked first and it is very close to the data curve, in both local 3.8(a) and global 3.8(b) models. The *retweet-mention-social* hypothesis also indicates plausibility since it outperforms the uniform (i.e., positive Bayes factors). However, if we look at each hypothesis



(a) Local: Bayes Factors

(b) Global: Bayes Factors

Figure 3.8.: **Ranking of hypotheses for the reply Higgs network.** (a,b) Ranking of hypotheses based on Bayes factors when compared to the uniform hypothesis using multiplexes for the local and global models, respectively. In both cases, the *mention* hypothesis explains best the reply network, since it is ranked first and very close to the data curve. This might be due to the fact that replies inherit a user mention from whom a tweet was originally posted. We can see that the combined *retweet-mention-social* hypothesis is the second best explanation of the reply network. This is mainly due to the mention hypothesis which performs extremely better than the other two (social and retweet). The *social* hypothesis can also be considered a good explanation since it outperforms the uniform. The *retweet* hypothesis tends to perform worse than the uniform in both cases for increasing number of κ. Similarly, the *self-loop* hypothesis drops down below the uniform since there are only very few self-loops in the *reply* network data.

individually, we can see that the combined hypothesis is dominated mainly by the *mention* hypothesis. The *social* hypothesis is also a good explanation of the number of replies since it outperforms the uniform hypothesis. *Retweets* and *self-loops* on the other hand show negative Bayes factors, suggesting that they are not good explanations of edge formation in the reply network. Note that the retweet curve in the local model has a very strong tendency to go below the uniform for higher numbers of κ. These results suggest us that the number of replies is proportional to the number of mentions and that usually people prefer to reply other users within their social network (i.e., followees).

3.6. Discussion

Next, we discuss some aspects and open questions related to the proposed approach.

Comparison to existing method. While we have already demonstrated the plausibility of Janus based on synthetic and empirical datasets, we want to discuss how our results compare to existing state-of-the-art methods. A simple alternative approach to evaluate the plausibility of beliefs as expressed by the belief matrices is to compute a Pearson correlation coefficient between the entries in the belief matrix and the respective entries in the adjacency matrix of the network. To circumvent the difficulties of correlating matrices, they can be flattened to vectors that are then passed to the correlation calculation. Then, hypotheses can be ranked according to their resulting correlation against the data. However, by flattening the matrices, we disregard the direct relationship between nodes in the matrix and introduce inherent dependencies to the individual data points of the vectors used for Pearson calculation. To tackle this issue, one can utilize the Quadratic Assignment Procedure (QAP) as mentioned in Section 3.2. QAP is a widely used technique for testing hypotheses on dyadic data (e.g., social networks). It extends the simple Pearson correlation calculation step by a significance test accounting for the underlying link structure in the given network using shuffling techniques. For a comparison with our approach, we executed QAP for all datasets and hypotheses presented in Section 3.5 using the `qaptest` function included in the `statnet` [108, 109] package in `R` [207].

Table 3.1.: **QAP on synthetic dyad-attributed network (multiplex).** List of correlation coefficients for each hypothesis tested. Statistically highly significant p-values ($p < 0.001$) are marked by (**). Last two columns show ranking of hypotheses according to Janus for the local and global models. By omitting the uniform hypothesis in Janus (rank 7) we can see that the ranking of hypotheses by correlation aligns with the rankings given by Janus for the multiplex given in Section 3.5.2.

Hypothesis	**Correlation Coefficient**	**P-Value**	**Janus Ranking Local**	**Janus Ranking Global**
Epsilon10p	0.939	0.0**	1	1
Epsilon20p	0.863	0.0**	2	2
Epsilon30p	0.787	0.0**	3	3
Epsilon40p	0.704	0.0**	4	4
Epsilon50p	0.636	0.0**	5	5
Epsilon60p	0.461	0.0**	6	6
Epsilon70p	0.352	0.0**	8	8
Epsilon80p	0.242	0.0**	9	9
Epsilon90p	0.142	0.0**	10	10
Epsilon100p	0.010	0.238	11	11

Overall, we find in all experiments strong similarities between the ranking provided by the correlation coefficients of QAP and our rankings according to Janus. Exemplary, Table 3.1 shows the correlation coefficients and p-values obtained with QAP for each hypothesis tested on the synthetic multiplex described in Section 3.5.2 as well as the ranking of hypotheses obtained from Janus for the local and global models (leaving the uniform hypothesis out). However, in other datasets minor differences in the ordering of the hypotheses could be observed between the two approaches.

Compared to QAP, Janus yields several advantages, but also some disadvantages. First, by utilizing our belief matrix as priors over parameter configurations instead of fixed parameter configurations themselves, we allow for tolerance in the parameter specification. Exploring different values of tolerance expressed by our parameter κ allows for more fine-grained and advanced insights into the relative plausibility of hypotheses. Contrary, simple correlation takes the hypothesis as it is and calculates a single correlation coefficient that does not allow for tolerances.

Second, by building upon Bayesian statistics, the significance (or decisiveness) of results in our approach is determined by Bayes factors, a Bayesian alternative to traditional p-value testing. Instead of just measuring evidence *against* one null hypothesis, Bayes Factors allow to directly gather evidence *in favor* of a hypothesis compared to another hypothesis, which is arguably more suitable for ranking.

Third, QAP and MRQAP, and subsequently correlation and regression, are subject to multiple assumptions which our generative Bayesian approach circumvents. Currently, we employ QAP with simplistic linear Pearson correlation coefficients. However, one could argue that count data (multiplicity of edges) warrants advanced generalized linear models such as Poisson regression or Negative Binomial regression models.

Furthermore, our approach intuitively allows to model not only the overall network, but also the ego-networks of the individual nodes using the local models presented above. Finally, correlation coefficients cannot be applied for all hypotheses. Specifically, it is not possible to compute it for the uniform hypothesis since in this case all values in the flatten vector are identical. However, our method currently does not sufficiently account for dependencies within the network as it is done by specialized QAP significance tests. Exploring this issue and extending our Bayesian approach into this direction will be a key subject of future work.

Runtime performance. A typical concern often associated with Bayesian procedures are the excessive runtime requirements, especially if calculating marginal likelihoods is necessary. However, the network models employed for this paper allow to calculate the marginal likelihoods—and consequently also the Bayes factors—efficiently in closed form. This results in runtimes, which are not only competitive with alternative methods such as QAP and MRQAP, but could be calculated up to 400 times faster than MRQAP in our experiments as MRQAP requires many data reshuffles and regression fits. Furthermore, the calculation (of Bayesian evidence) could easily be distributed onto several computational units, cf. [23].

Local vs global model. In this paper, we presented two variations of our approach, i.e., a local and a global model. Although both model substantially different generation processes (an entire network vs. a set of ego-networks), our experiments have shown that hypotheses in the global scenario are ranked mostly the same as the ones using the local model. This is also to be expected to some degree since the constructed hypotheses did not explicitly expressed a belief that outgoing links are more likely for some nodes.

Inconsistency of local model. For directed networks, the local ego-network models can assemble a full graph model by defining a probability distribution for the degrees of the source nodes of edges. For undirected networks, this is not directly possible as e.g., the ego-network model for v_A generated an edge from v_A to v_B, but the ego-network model for node v_B did not generate any edge to v_A. Note that this does not affect our comparison of hypotheses as we characterize the network.

Single edges. As mentioned in Section 3.3, Janus focuses on multigraphs, meaning that edges might appear more than once. This is because we assume that a given node v_i, with some probability p_{ij}, will be connected *multiple* times to any other node v_j in the local models. The same applies to the global model where we assume that a given edge (v_i, v_j) will appear *multiple* times within the graph with some probability p_{ij}. For the specific case of single edges (i.e., unweighted graphs), where $m_{ij} \in 0, 1$, one might consider other probabilistic models to represent such graphs (e.g., Bernoulli distribution).

Sparse data-connections. Most real networks exhibit small world properties such as high clustering coefficient and fat-tailed degree distributions meaning that the adjacency matrices are sparse. While comparison still relatively judges the plausibility, all hypotheses perform weak compared to the data curve as shown in Figure 3.7(b). As an alternative, one might want to limit our beliefs to only those edges that exist in the network, i.e., we would then only build hypotheses on how edge multiplicity varies between edges.

Other limitations and future work. The main intent of this work is the introduction of a hypothesis-driven Bayesian approach for understanding edge formation in networks. To that end, we showcased this approach on simple categorical models that warrant extensions, e.g., by incorporating appropriate models for other types of networks such as weighted or temporal networks. We can further investigate how to build good hypotheses by leveraging all node attributes, and infer subnetworks that fit best each of the given hypotheses. In the future, we also plan an extensive comparison to other methods such as mixed-effects models and p^* models. Ultimately, our models also warrant extensions to adhere to the degree sequence in the network, e.g., in the direction of multivariate hypergeometric distributions as recently proposed in [41].

3.7. Conclusions

In this paper, we have presented a Bayesian framework that facilitates the understanding of edge formation in node-attributed and dyad-attributed multigraphs. The main idea is based on expressing hypotheses as beliefs in parameters (i.e., multiplicity of edges), incorporate them as priors, and utilize Bayes factors for comparing their plausibility. We proposed simple local and global Dirichlet-categorical models and showcased their utility on synthetic and empirical data. For illustration purposes our examples are based on small networks. We tested our approach with larger networks obtaining identical results. We briefly compare Janus with existing methods and discuss some advantages and disadvantages over the state-of-the-art QAP. In future, our concepts can be extended to further models such as models adhering to fixed degree sequences. We hope that our work contributes new ideas to the research line of understanding edge formation in complex networks.

Part II.

Influence of Edge Formation in Machine Learning

4. The Influence of Edge Formation in Classification

This chapter proposes evaluation benchmarks to better understand the influence of network structure on the performance and bias of relational classification, a machine learning technique that infers the class label of a node based on the class label of its neighbors. This work has been published as a full paper in the journal of Applied Network Science, 2021.

Social networks are very important carriers of information. For instance, the political leaning of our friends can serve as a proxy to identify our own political preferences. This explanatory power is leveraged in many scenarios such as business decision-making and scientific research to infer missing attributes using machine learning. However, factors affecting the performance and the direction of bias of these algorithms are not well understood. To this end, we systematically study how structural properties of the *network* and the *training sample* influence the results of collective classification. Our main findings show that (i) mean classification performance can empirically and analytically be predicted by structural properties such as homophily, class balance, edge density, and sample size, (ii) small training samples are enough for heterophilic networks to achieve high and unbiased classification performance, even with imperfect model estimates, (iii) homophilic networks are more prone to bias issues and low performance when group size differences increase, (iv) when sampling budgets are small, partial crawls achieve the most accurate model estimates, and degree sampling achieves the highest overall performance. Our findings help practitioners to better understand and evaluate their results when sampling budgets are small or when no ground-truth is available.

4.1. Introduction

People connect with others through online social network platforms [123], scientific collaboration networks [188], and other peer-to-peer platforms [151]. All these connections are leveraged by certain systems to recommend new content and new connections. In turn, these recommendations are often based on algorithms that rely on individuals' information such as gender, political leaning or credit score. In practice, however, often only partial information about individuals is available due to API quotas (e.g., very large networks). In this scenario, collective classification[1] [96, 165, 187] can be used to infer individual's attributes using information from their neighbors and a few *seeds* (i.e., individuals with known attributes). The advantage of collective classification over traditional machine learning techniques[2] is that the former does not require the data to be independent and identically distributed, which is important when dealing with networked data, as the class label of a node may depend on the class label of its neighbors in the network.

However, little is known about the impact of network structure on the performance and the direction of bias of collective classification. For instance, many social networks demonstrate a property known as *homophily*, which is the tendency of individuals to associate with others who are similar to them, e.g., with respect to gender or ethnicity [171]. Furthermore, the *class balance* or distribution of individual attributes over the network is often uneven, with coexisting groups of different sizes, e.g., one ethnic group may dominate the other in size. A challenge for inference is then to be accurate and unbiased with each individual and group in the network, regardless of its structure. However, the variety of network types—as well as many choices for the sampling method, modeling, and inference—make it difficult to choose the best combination of methods for a particular problem. A further complication is that ground truth data is not always available to evaluate results.

Therefore, it becomes crucial to understand how these algorithms work and under which conditions they discriminate against certain groups of people (e.g., minorities). To that end, our work aims at providing decision makers with: (i) evaluation guidelines to assess the impact of different network types and sampling techniques on collective classification, and (ii) a reproducible and reusable tool to identify performance and bias issues on new networks, sampling techniques, classifiers and inference algorithms. Our findings also shed light on the design of better algorithms to mitigate biases coming from networked data.

Research questions. In this work we systematically study different factors that may influence the performance and bias of collective classification. These factors relate to structural properties of the *network* and the *training sample* (i.e., random sampled subgraph with labeled nodes, so-called seed nodes) involved in the beginning of the inference process (see Figure 4.1).

- **RQ1:** How does *network structure* (i.e., *homophily*, *class balance*, and *edge density*) affect the overall performance of collective classification?
- **RQ2:** How does the choice of the *sampling technique* affect the overall performance of collective classification and its parameter estimation?
- **RQ3:** How does network structure and the choice of sampling technique influence the direction of bias in collective classification?

[1] A technique that combines relational classification and collective inference.

[2] which rely only on node attributes and ignore relationships with other nodes.

Approach and methods. We utilize a network model that allows us to generate scale-free networks with tunable homophily and class balance [133]. One advantage of this model is that it generates node-attributed networks with power-law degree distributions which have been observed in many large-scale social networks [17]. More importantly, it only requires two main input parameters (homophily and class balance), and thus the behavior of the model is analytically tractable [133]. The homophily parameter h ranges from 0 to 1, and it allows us to generate networks with a broad range of group mixing ranging from heterophilic networks ($0 \leq h < 0.5$) to neutral networks ($h = 0.5$), and homophilic networks ($0.5 < h \leq 1$).

Furthermore, we follow definitions and pseudo-codes from [165] to implement the *network-only Bayes* classifier (nBC) and the *relaxation labeling* inference algorithm (RL). We measure classification performance in terms of ROCAUC, assess the quality of the model parameters using squared estimation errors (SE), and extend the balanced accuracy [34] to compare the *true positive rates* (TPR) of each class—also known as sensitivity and specificity in binary classification—to assess the direction of bias.

Contributions. Our contributions are two-fold: First, we propose a methodology to assess classification performance and bias on real-world networks when no ground-truth is available. Second, we demonstrate analytically and empirically that collective classification mean performance, estimation error, and bias are predictable and mainly depend on homophily, class balance, edge density, and sample size. In particular, we show that: (i) Larger training samples are required in undirected networks with homophilic connections to achieve a similar high classification performance compared to heterophilic networks. (ii) Samples obtained by partial crawls allow to learn the most accurate model estimates, and the most accurate results are obtained when using degree sampling. (iii) Classification results are often less biased in heterophilic networks than in homophilic networks regardless of class imbalance. Last but not least, we make our code and data openly available [75].

4.2. Related Work

We cover previous work on analyzing the performance of collective classification from both *algorithmic* and *network* bias perspectives. Additionally, we review literature on *network sampling* to identify characteristics of *seed nodes* (i.e., labeled nodes) that could potentially improve the performance of collective classification.

Collective classification performance. From the literature, we know that each component in the collective classification pipeline can be implemented in different ways, and it has been shown that different implementations (or combinations) may perform better or worse depending on certain properties of the network and sample size [165, 227, 274]. For instance, Macskassy and Provost [165] evaluated the influence of relational classifiers (**RC**) together with collective inference algorithms (**CI**) and sample size using random node sampling, and conclude that the network-only Bayes classifier (**nBC**) is almost always significantly and often substantially worse than other RCs. When samples are small, relaxation labeling (**RL**) is best among all CIs, whereas weighted-voting (**wvRN**) and class-distribution (**cdRN**) are best among all RCs. When samples are large, all CIs perform similarly well, and network-only link-based (**nLB**) is best among all RCs. More recent work by [274] concluded that as the sample size increases it is better to learn a model using nBC than with wvRN. Note that this work utilizes synthetic networks and assumes edge weights to be $w_{ij} = 1$, opposite to the work by Macskassy and Provost [165] where they run their experiments on real-world networks and thus utilize real weight values $w_{ij} \in \mathbb{R}$. While all these contributions are very important and relevant, they have mostly focused on the performance of RCs and CIs. Besides, their findings are not comparable since they use different datasets, different configurations of RC and CI, and different evaluation metrics. In our work we focus on the performance of nBC and RL, and systematically vary some properties of the network and the training sample. We choose this combination because it has been shown that RL outperforms other CIs [165], and the model parameters of nBC are easy to interpret, i.e., they reflect the network properties of interest. Besides, nBC outperforms wvRN [274]. Moreover, we use ROCAUC as a standard measure of classification performance and compare the true positive rates of each class to assess the direction of bias.

Network bias. The influence of homophily and edge density on both RC and CI was studied in [227] on synthetic networks. They found that, as homophily (or density) increases, the accuracy of classification improves drastically over non-relational classifiers. It has also been shown that certain RC such as wvRN perform poorly on heterophilic networks [62]. Besides, when networks are neutral (i.e., when nodes are connected at random), no classifier—even with the largest training dataset—can beat a random classifier [80]. We build upon these findings and broaden the spectrum of network types by varying not only homophily and density, but also class balance (i.e., fraction of minorities). Note that a new metric called *monophily* shows that similarity among friends-of-friends (a.k.a., 2-hop neighbors) can improve relational classification results, especially in the neutral case [6]. However, we focus on 1-hop neighbors to better understand the role of homophily and class balance in collective classification.

Sampling bias on networks. Previous research has studied the robustness of network samples from different angles. For instance, a range of network properties such as degree and betweenness centrality have been found to be sensitive to the choice of sampling methods [29, 52, 93, 124, 144, 152, 155, 156, 257, 260]. These efforts have shown that network estimates become more inaccurate with lower sample coverage, but there is a wide variability of these effects across different measures, network structures and sampling errors. In terms of benchmarking network sampling strategies, Coscia and Rossi [51] show that it is not enough to ask which method returns the most accurate sample (in terms of statistical properties); one also needs to consider API constraints and sampling budgets. In the context of collective classification, Yang et al. [268] demonstrated that certain sampling techniques such as snowball sampling and random walks could lead to biased parameter estimates, and then corrected such bias by exploiting a general crawling method [12] that produces unbiased model estimates. We leverage the nature of these estimates to verify whether perfect estimates always lead to perfect classification performance and unbiased results while varying the sampling budget.

Fairness in classification. In recent years, there has been an increase of research focusing on mitigating bias [61, 146, 208] and guaranteeing individual and group fairness while preserving accuracy in classification algorithms [25, 67, 132, 271]. Many definitions of fairness have been proposed [173, 253] and the most used are equalized odds [111], equal opportunity [111], counterfactual fairness [148], demographic parity, and fairness through awareness [66]. While all this body of research focuses on fairness influenced by the attributes of the individuals, recent research proposes a new notion of fairness that is able to capture the relational structure of individuals [84, 276]. The main difference between this line of research and our methodology is that instead of ensuring a fair algorithm, we focus on *explaining discrimination* [173] via input and sampling bias. By doing so, we gain a better understanding of the direction of bias (i.e., why and when collective classification discriminates against certain groups of people). Consequently, we simplify the classification task to work with only one (protected) binary attribute (e.g., gender) which in turn plays the role of a target and a membership class.

To our best knowledge, there is no systematic study that explores the interplay between sampling, network structure, performance, and bias in collective classification, and we aim to fill this gap.

4.3. Methods: Classification on Networks

We focus on *classification on networks* as a *semi*-supervised machine learning technique, where categorical class labels of records are predicted by exploiting both the labeled and the unlabeled part of the data [168]. In particular, we study *relational classification* together with *collective inference* (a.k.a., collective classification [227]), two techniques used to infer missing attributes of nodes using information from their neighbors. Figure 4.1 shows the four requirements of collective classification: (i) Data: a network with unlabeled nodes. (ii) Training sample: a subgraph with known labels sampled from the network. (iii) Models: local and relational models learned from the training sample to encode class priors and conditional probabilities, respectively. (iv) Collective Inference: a systematic process where models are fitted to the ego networks of each unlabeled node to infer their posterior class probabilities.

Next, we describe (i) networks of interest, (ii) network sampling, and (iii) the modeling and inference processes utilized in this work.

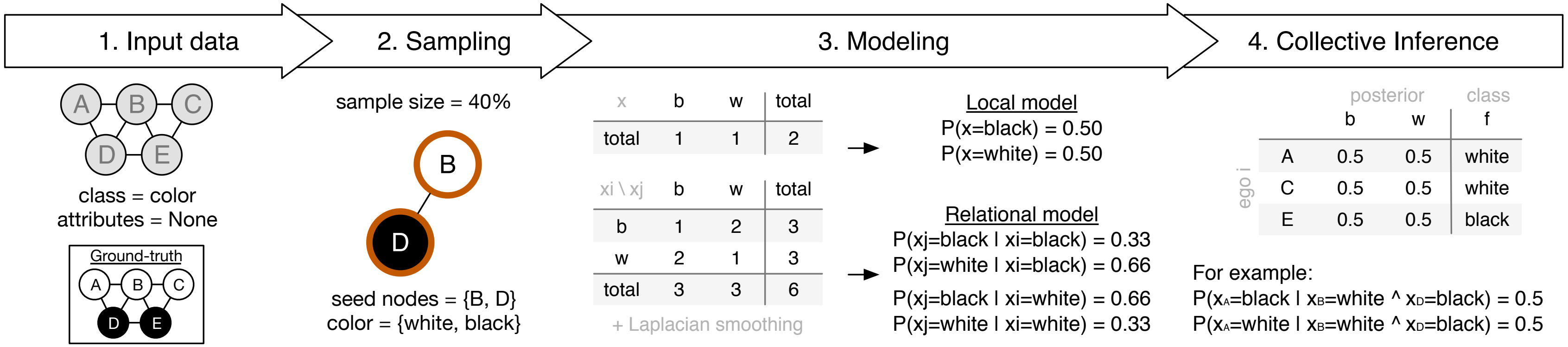


Figure 4.1.: **Classification on networks.** This example describes the collective classification workflow. (1) A network $G = (V, E, C)$ is given, and it is defined by a set of nodes V, edges E, and class labels C. Nodes are known to belong to a binary class *color* $\in \{white, black\}$, and no additional attributes are given. Therefore, the goal of the collective classification is to infer the correct class label of nodes. To achieve this (2) a set of seed nodes is sampled, together with their class labels, to create a subgraph $G_{\text{seeds}} \subset G$. Then, (3) the local model learns the probability of each class label, e.g., $P(x = black) = 0.5$, while the relational model learns the probability of neighbor v_j having a particular class label conditioned on the class label of a node v_i, e.g., $P(x_j = white|x_i = black) = 0.66$. Finally, in (4) the collective inference process assigns posterior probabilities to each unlabeled node using their 1-hop neighbors. For example, the probability that node A is black (or white) is conditioned on the color of its neighbors B and D. Notice that the inference is performed collectively, i.e., at the same time for all nodes.

4.3.1. Input data: An attributed network

We define the input network as: Let $G = (V, E, C)$ be an attributed unweighted graph with $V = \{v_1, ..., v_n\}$ being a set of N nodes, $E \subseteq V \times V$ a set of undirected edges, and $C = \{c_1, ..., c_n\}$ a list of binary class labels where each element c_i represents the class membership of node v_i.

The homophily parameter H is the probability of nodes with the same class label to be connected. Homophily values range from 0 to 1. Networks with homophily $H = 0.5$ are referred to as *neutral*, otherwise they are *heterophilic* if $H < 0.5$, or *homophilic* when $H > 0.5$. Class balance B captures the fraction of minority nodes—with respect to C—in the network. A network is *balanced* when all class labels have the same number of nodes ($B = 0.5$), otherwise it is *unbalanced* ($B < 0.5$). Edge density $d = \frac{2|E|}{N(N-1)}$ represents the fraction of existing edges out of all possible edges in G.

To generate such networks, we refer to the preferential attachment-based model with adjustable homophily proposed in [133]. In this model, each node is assigned one binary class label, e.g., $color \in \{white, black\}$[3]. The probability of node v_i to connect to node v_j is given by:

$$\Pi_{ij} = \frac{h_{ij} k_j}{\sum_l h_{il} k_l} \tag{4.1}$$

where k_i is the degree of node v_i and h_{ij} is the homophily between nodes v_i and v_j. For simplicity, in our synthetic networks, we assume that homophily is symmetric and complementary: $h_{aa} = h_{bb} = H$ and $h_{ab} = h_{ba} = 1 - H$.

Example. Figure 4.1 shows an attributed network (see the ground-truth in "1. Input data"), where nodes are assigned one *color*, either *white* or *black*. Since only 3 out of 7 edges are same-color connections, this network is heterophilic ($H \approx 0.43$). This network is also unbalanced ($B = \frac{2}{5} = 0.4$) because the number of black nodes ($N_b = 2$) is different from the number of white nodes ($N_w = 3$).

Note that in practice, often the list of class labels C is unknown or incomplete. Therefore, values for B and H are either not available or inaccurate. However, in our experiments we assume that the ground-truth is given (see Section 4.7 for a real use case).

4.3.2. Sampling: The observed network

The goal of sampling is to split the network into *training* and *testing* samples. First, a subgraph $G_{\text{seeds}} = (\hat{V}, \hat{E}, \hat{C})$ is extracted from G in order to learn the model parameters (see Section 4.3.3). Nodes $\hat{V} \subset V$ that belong to the training sample G_{seeds} are called *seed nodes*, and they are a percentage *pseeds* of nodes selected by the sampling method. Similarly, edges $\hat{E} \subseteq \hat{V} \times \hat{V}$ are all links ($\hat{E} \subset E$) in the induced subgraph between seed nodes $\hat{V}$. Class labels $\hat{C} \subset C$ are automatically known by the classification algorithm after sampling. The testing

[3] Minority group always refers to black nodes.

sample includes all nodes and edges of G, but only nodes $v_i \in V - \hat{V}$ are target for classification. In this paper we explore four widely used sampling methods: random nodes, random edges, degree ranking, and partial crawls.

Random nodes. This is the most used and basic sampling method where a percentage *pseeds* of random nodes is selected. The training sample then contains the selected nodes and all edges among them. Note that in the case of unbalanced networks, this sample will be biased towards nodes in the majority class.

Random edges. This technique randomly selects edges (and their nodes) until it reaches a specific percentage *pseeds* of nodes. The training sample then contains this percentage of nodes and the selected edges. Note that when sampling by edges, the resulting sample will be biased towards hubs since they get higher chances to be picked multiple times through their multiple connections.

Degree rank. We rank all nodes by their degree in descending order and select the top *pseeds*% of nodes. While computing the degree of all nodes is expensive, the idea is to verify whether high degree nodes improve the inference [159]. The main difference compared to edge sampling is that degree sampling selects hubs without their neighbors (which may have low degree).

Partial crawls. Yang et al. [268] proposed a crawl-aware parameter estimator of peer-effect based on random walk tours. The procedure is detailed in [12], but roughly described as follows. First, a fraction p_{sn} of nodes in V is sampled by some arbitrary sampling procedure (the method is insensitive to sampling biases in this phase). These sampled nodes are called a *super node* S. Second, t different random walks are performed starting at nodes in S, ending when the random walker encounters any node also in S. The starting node is chosen from S proportional to the number of edges from nodes in S to nodes outside S. The random walk progresses by moving to a random neighbor of the current visited node. Note that all nodes in the super node together with the crawled nodes in every tour belong to G_{seeds}[4]. The random walker stops once G_{seeds} contains a percentage *pseeds* of the total number of nodes in G. It has been shown that partial crawls can provide unbiased estimates of network statistics [12]. Thus, we verify whether perfect model estimates always lead to perfect classification.

In RQ2 we compare the influence that each of these sampling techniques has on classification performance and parameter estimation. In RQ3 we examine their bias with respect to minority and majority groups. In RQ1 we explore the impact of the network structure on classification performance. Thus, to ensure that the sampling method does not influence the performance we only use random node sampling.

Example. Following the example in Figure 4.1 (see "2. Sampling"), the subgraph extracted via random nodes consists of: 40% randomly selected nodes

[4] This is an adaptation of the original algorithm to work with semi-supervised learning. That is, instead of crawling an unknown network, we extract a subgraph by sampling class labels of nodes from a known network.

$\hat{V} = \{B, D\}$, their class labels (color) $\hat{c}_B = \textit{white}$ and $\hat{c}_D = \textit{black}$, and all edges between them $\hat{E} = \{(B, D)\}$.

4.3.3. Modeling and collective inference: Estimates

Collective classification in networked data [96, 129, 165] learns correlations between attributes of linked nodes from observed data, and transfers this knowledge simultaneously to the unseen nodes. This process consists of three components: local model, relational model, and collective inference. To isolate the effects of network and training sample, we fix the classification algorithm as follows. We (i) learn the local model LC from the nodes in the training sample, (ii) learn the relational model RC from the nodes and edges in the training sample using *Bayesian* statistics, and (iii) infer class values using *relaxation labeling* as the collective inference process CI. Therefore, the probability of a node $v_i \in V - \hat{V}$ with neighbors $\mathscr{N}_i$ taking on class $x_i = c$ is given by:

$$\overbrace{P(x_i = c|\mathscr{N}_i)}^{\text{posterior}} = \frac{\overbrace{P(x = c)}^{\text{prior}} \cdot \overbrace{P(\mathscr{N}_i|x_i = c)}^{\text{likelihood}}}{\underbrace{P(\mathscr{N}_i)}_{\text{marginal likelihood}}} \tag{4.2}$$

where $P(\mathscr{N}_i|x_i = c) = \prod_{v_j \in \mathscr{N}_i} P(x_j = \widetilde{x}_j|x_i = c)$ and $\widetilde{x}_j$ is the actual class observed at node v_j. Parameters in the local and relational model include the prior probability $P(x = c)$ of any node being of class c, and conditional probabilities $P(x_j = \widetilde{x}_j|x_i = c)$ that a neighboring node v_j has class $\widetilde{x}_j$ given that node v_i has class c.

Parameter estimation. The model parameters are inferred from the nodes and edges in the training sample. The estimates learned using the partial crawls algorithm are defined in [268]. The estimates learned using random nodes, random edges, and degree ranking are calculated as follows:

$$P(x = c) = \frac{1}{|\hat{V}|} \sum_{\hat{v}_i \in \hat{V}} \mathbb{1}\{\hat{c}_i = c\} \tag{4.3}$$

$$P(x_j = a|x_i = c) = \frac{\sum_{(\hat{v}_i, \hat{v}_j) \in \hat{E}} \mathbb{1}\{\hat{c}_i = c\} \cdot \mathbb{1}\{\hat{c}_j = a\}}{\sum_{(\hat{v}_i, \hat{v}_j) \in \hat{E}} \mathbb{1}\{\hat{c}_i = c\}} \tag{4.4}$$

Inference (relaxation labeling). Once the model parameters are learned, relaxation labeling initializes each unlabeled node with the prior probabilities. Then, rather than estimating one node at a time and updating the graph right away, the current estimations are frozen so that at step $t + 1$ all vertices will be updated based on the estimations of step t. The updating step takes into consideration a decay constant to regulate the influence of neighboring nodes in every iteration [165].

Example. During the modeling phase in Figure 4.1 (see "3. Modeling") we learn the prior probabilities (e.g., $P(x = black) = 0.5$) and the conditional probabilities (e.g., $P(x_j = white|x_i = black) = 0.66$). Continuing to the inference phase in Figure 4.1 (see "4. Collective Inference"), the *relaxation labeling* first initializes the posterior probabilities of all unlabeled nodes using the class priors, and then iterates through all unlabeled nodes simultaneously to infer their posterior probabilities using the Bayes theorem, see Equation (4.2).

4.4. Experimental Setup

To explore the interplay between network structure, sampling techniques, and the performance and bias of classification, we systematically vary structural properties of the *network* and the *training sample* by fixing the classification algorithm as explained below.

Synthetic networks. We generate 330 undirected networks G using the model by Karimi et al. [133], referred to as BA-Homophily[5], and adjust four parameters: number of nodes $N = 2\,000$, class balance or fraction of minorities $B \in \{0.1, 0.3, 0.5\}$, homophily $H \in \{0.0, 0.1, \ldots, 1.0\}$, and edge density $d \in \{0.004, 0.02\}$[6]. Networks are generated 5 times in each configuration to control for random fluctuations. We omitted results using smaller and bigger networks, i.e., $N \in \{500, 10000\}$, since their mean ROCAUC scores are very similar to the ones obtained with $N = 2000$ (see Supplementary Figure 4.9). The main difference is that the variance of ROCAUC reduces with larger networks. Due to this similarity we decided to show only results with $N = 2\,000$.

Training samples. Subgraphs G_{seeds} contain a percentage *pseeds* of nodes from G that are selected by one of the following sampling methods: random nodes (*nodes*), random edges (*edges*), degree rank (*degree*) and partial crawls (*partial_crawls*). We assume that G_{seeds} is completely observed, which means that we know the class labels of nodes and all or some edges among them. We vary $pseeds \in \{5\%, 10\%, 20\%, ..., 90\%\}$ to measure the impact of sample size on classification. In the particular case of the partial crawls, we set the size of the super node to $|S| = p_{sn} \times N$, where $p_{sn} = 0.01$, and the number of tours t to as many as necessary until reaching $pseeds \times N$ of nodes in G_{seeds}. For each *pseeds*, we run the classification algorithm 10 times.

Classification algorithm. We focus on uni-variate network classification, which means that the linkage structure in the network is modeled with the class label of

[5] The acronym for Barabási-Albert Homophilic network.

[6] Density in the BA-Homophily model originally was adjusted via minimum degree $m \in \{4, 20\}$. Since degrees are power-law distributed and nodes have minimum degree of m, the larger the value of m the higher the density d.

the nodes and no information from additional node attributes. In particular, we choose the *network-only Bayes classifier* (nBC) as the relational model (RC), and apply *relaxation labeling* (RL) as the collective inference algorithm (CI). We use this combination for two reasons. First, it has been shown that RL outperforms other CIs when training samples are small, and when training samples are large any CI performs equally well [165]. Second, the nBC model parameters are easy to interpret since they are based on network structure (i.e., class priors relate to class balance, and conditional probabilities to homophily). Additionally, we show that the overall trend of classification performance vs. network structure does not vary with a second RC, namely *LINK classifier* [278].

Evaluation. We quantify the performance of the classification using three different metrics: (i) **ROCAUC score**[7] to estimate the overall performance of the collective classification, (ii) **squared estimation errors (SE)** between global and sample parameters to assess the quality of the parameter estimation, and (iii) a comparison between the **true positive rates** of each class to measure the direction of bias. Note that when working with unbalanced data, a classifier may achieve high overall performance even if it often misclassifies instances of the minority class. By comparing the positive rates—sensitivity and specificity in binary classification—we disentangle the direction of bias and assess how well the algorithm classified both, minority and majority classes.

4.5. Results

Using *naive Bayes* and *relaxation labeling*, we classify nodes as either *white* or *black* using different sample sizes and different evaluation metrics (see Section 4.4). In the following we will discuss our results and answer the three research questions which we raised before.

4.5.1. RQ1: How does *network structure* affect the overall performance of collective classification?

We analyze to what extent the structure of the network (i.e., *homophily*, *class balance* and *edge density*) impacts classification performance. We measure performance of collective classification using ROCAUC scores, where each value can be interpreted as the probability of distinguishing between classes.

Overall performance vs. network structure

Figure 4.2 shows the classification performance on synthetic networks with number of nodes $N = 2000$ and edge density $d \in \{0.004, 0.02\}$ (rows)[8]. Class balance is defined by the parameter B (columns). Homophily H ranges from 0 to 1

[7] Area under the receiver operating characteristic curve.

[8] A different visualization can be found in Supplementary Figure 4.8.

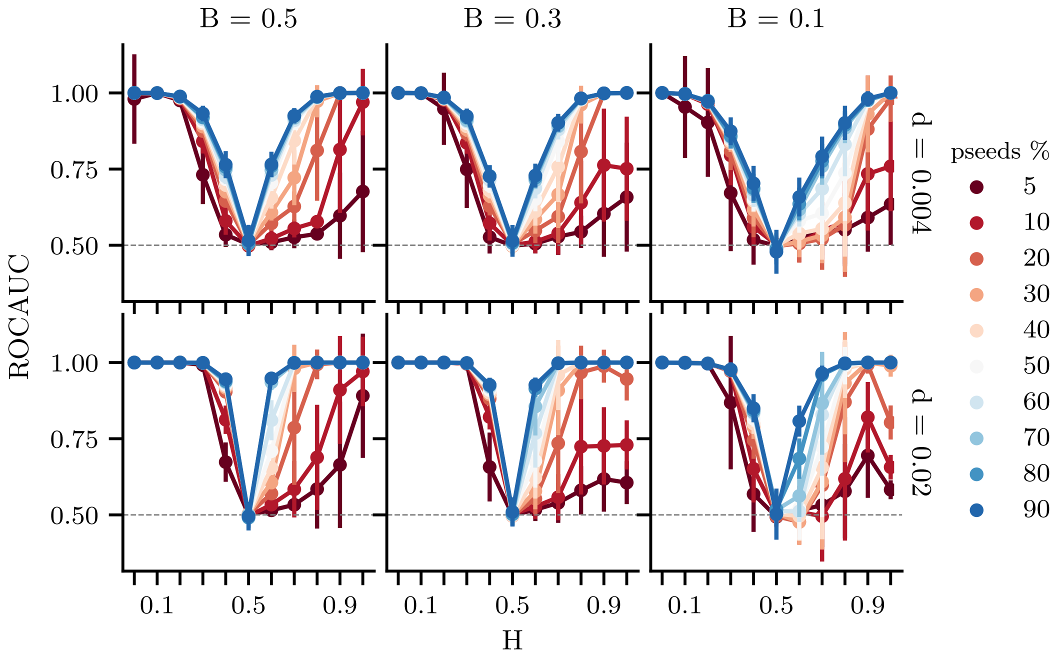


Figure 4.2.: **RQ1: Network structure, sample size and overall performance using random node sampling and network-only Bayes classifier (nBC).** Classification performance, measured by ROCAUC scores, is shown on the y-axis for networks with $N = 2000$ nodes and different levels of homophily (x-axis), edge density (rows), class balance (columns), and sample size (colors). Dots represent mean ROCAUC scores over 50 runs, and error bars their respective standard deviation. In general we see that: (i) neutral networks $H = 0.5$ cannot be classified better than a random classifier; (ii) heterophilic networks $H < 0.5$ require smaller samples to achieve high and stable classification performance compared to homophilic networks $H > 0.5$; (iii) Dense networks $d = 0.02$ achieve higher ROCAUC compared to sparse networks $d = 0.004$.

(x-axis). Sample size, using random node sampling is shown as the percentage *pseeds* of nodes (colors), and the overall performance as ROCAUC scores (y-axis). At first glance, from Figure 4.2 we notice four main patterns. **(i)** As expected, classification performance on neutral networks ($H = 0.5$) is always similar to a random classifier. **(ii)** Surprisingly, heterophilic networks ($H < 0.5$) require smaller samples to achieve high and stable classification performance compared to homophilic networks ($H > 0.5$). **(iii)** ROCAUC scores are neither stable nor consistent (i.e., high variance) in the homophilic regime when samples are very small. In other words, classification performance varies widely. **(iv)** Dense networks (d=0.02) achieve higher classification performance compared to sparse networks (d=0.004) around $H = 0.5 \pm 0.3$, i.e., $\overline{\text{ROCAUC}}_{d=0.02,H=0.5\pm0.3} = 0.82 > \overline{\text{ROCAUC}}_{d=0.004,H=0.5\pm0.3} = 0.74$.

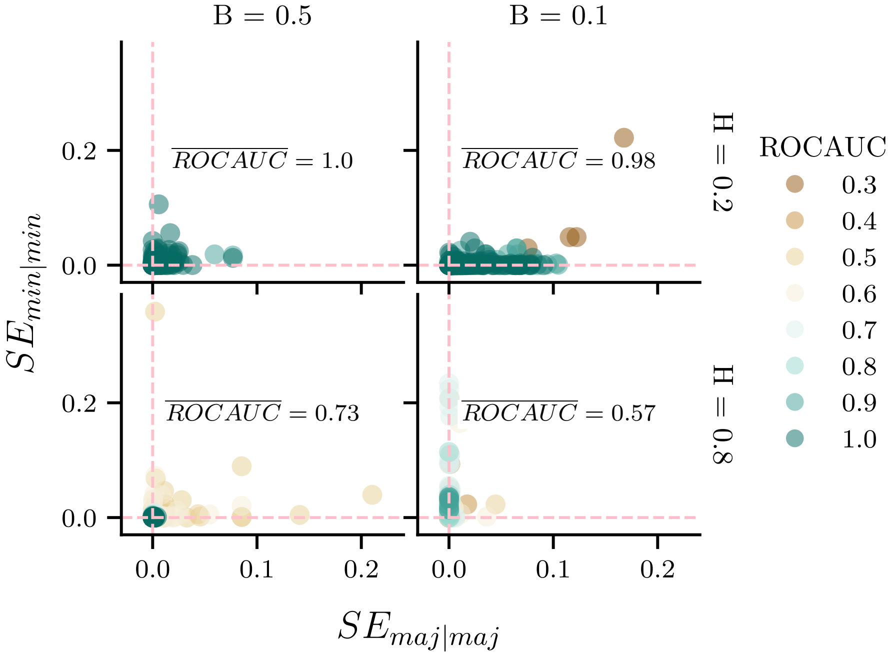


Figure 4.3.: **RQ1: Estimation errors on small samples using random node sampling.** These are the squared estimation errors (SE) of the conditional probabilities $P_{maj|maj}$ (x-axis) and $P_{min|min}$ (y-axis) learned from small samples ($pseeds \leq 30\%$) using random nodes on sparse networks ($N = 2000, d = 0.004$) with different levels of homophily (rows) and class balance (columns). Mean ROCAUC scores within each type of network is shown as $\overline{ROCAUC}$. In general we see that homophilic networks require lower estimation errors to achieve high performance (green). Also, in unbalanced networks, low estimation error within minorities correlates with high performance in heterophilic networks, while in homophilic networks the correlation is between low estimation error within majorities and high performance.

Why is heterophily easier to predict?

In Figure 4.2 we see an asymmetry between homophilic ($H > 0.5$) and heterophilic ($H < 0.5$) regimes for small samples (red lines) and all class balance levels B (columns). To explain this discrepancy, we turn to the properties of the sampling error[9] and the network structure: Undirected networks only contain three types of edges, e.g., black-white, white-white, and black-black. In the heterophilic regime,

[9] The error caused by observing a sample instead of the whole population.

only one type of edge is prevalent (black-white), while in the homophilic regime two types are equally prevalent (white-white, black-black). In general, for small training samples (e.g., *pseeds* $\leq 30\%$), the probability of correctly observing each type of edge is very low. Consequently, the parameter estimation is prone to be wrong. However, its impact depends on the class balance and homophily of the network.

Balanced networks, B=0.5. First, note that the probability of observing a black-black edge in the synthetic network can be calculated analytically given the homophily (H), the class balance (B), and the degree exponents of the groups (β) as follows:

$$P_{bb} = \frac{B^2 H(1-\beta_w)}{Z} \tag{4.5}$$

where, Z is a normalization constant, and β_b and β_w are the exponents of the degree distribution for the *black* and *white* nodes, respectively. For the detailed analytical derivations and values of β see [133]. Similarly, the probability of observing a black-white edge is given by:

$$P_{bw} = \frac{B(1-B)(1-H)[(1-\beta_b)+(1-\beta_w)]}{Z} \tag{4.6}$$

In the heterophilic case ($H = 0.2$), the probability of observing a black-white edge in the whole graph is 0.8. Thus, the sampling error in a small sample follows $(0.8|\hat{E}|)^{-\frac{1}{2}}$, where $|\hat{E}|$ is the total number of edges in the sample. In the homophilic case ($H = 0.8$), the probability of observing a black-black edge is 0.4 and a white-white edge is also 0.4. The sampling error for *each* homophilic class is then $(0.4|\hat{E}|)^{-\frac{1}{2}}$ which individually are smaller than the error in the heterophilic case but adding them together they are larger. These sampling errors are reflected in the *estimation error* calculated here as the squared distance between the model parameter inferred from the training sample ($P\{.\}$) and the global network ($\theta\{.\}$):

$$SE\{.\} = (P\{.\} - \theta\{.\})^2 \tag{4.7}$$

We see these errors in the left-most column of Figure 4.3, where the x-axis refers to $SE_{maj|maj}$, and the y-axis to $SE_{min|min}$. Note that large errors in homophilic networks ($H = 0.8$) lead to low overall performance (brown). However, there are some cases where performance is also low even though such errors are small. This means that homophilic networks are more sensitive to the precision of the parameter estimation because it requires: $P_{maj|maj} = P_{min|min}$.

Unbalanced networks, B<0.5. In addition to the sampling error explained above, the group size differences and the inherent structure of the network add additional complexity to the learning process. This happens because of the interplay between homophily and preferential attachment which enables the formation of all different types of connections. For instance, in *homophilic networks* ($H = 0.8$), minority nodes will be mainly attracted by other minority nodes. However, due to the preferential attachment, minority nodes will also be partly attracted to majority nodes. On the other hand, majority nodes will be mostly connected to other majority nodes due to both mechanisms. Therefore, the estimation error of the conditional probability $P_{maj|maj}$ is on average lower than the estimation error for $P_{min|min}$, as shown at the bottom-right plot in Figure 4.3. The same principle applies to *heterophilic networks* ($H = 0.2$). In this case, even though most edges are heterophilic, networks will also contain edges between nodes of the same type but in significantly different proportions. Since there is only a very limited number of minority nodes, there can only be a very limited number of edges between them. That is not the case for majorities because they can connect to many more majorities. Therefore, though locally they connect to a few other majorities, globally there are many edges within this group. This gives an advantage to small samples because the randomly selected majority nodes are likely to be either disconnected[10] or connected to other minority nodes that are in the training sample. Thus, the classifier learns that the network is heterophilic. This explains why heterophilic networks can achieve high overall performance even when estimation errors are high for $P_{maj|maj}$ as shown in the top-right plot in Figure 4.3. This holds as long as $\frac{P_{maj|maj}}{P_{min|maj}} \times \frac{P_{min|min}}{P_{maj|min}} < 1$, otherwise the classifier believes that the network is extremely homophilic.

Finally, besides these conditional probabilities, class priors are also important in the collective inference. Thus, in the balanced case ($B = 0.5$), we expect the class priors to be the same: $P_{min} = P_{maj} = 0.5$; if this condition is not fulfilled, the classifier initially believes that one group is more prevalent than the other[11]. In the unbalanced case ($B = 0.1$), however, it is enough to identify the minority group correctly, regardless of its actual group size.

To what extent do these results depend on the algorithm?

For interpretability reasons we chose the network-only Bayes classifier (nBC) as relational model, since its model parameters correlate with the homophily and class balance of the network. However, it is unclear whether the results shown in Figure 4.2 are to some extent an artifact of nBC. Therefore, we run the classification algorithm on the same networks by changing the relational model. We choose the LINK classifier [6, 278], which learns a regularized logistic regression. The features of a node are the entire row of the adjacency matrix and the outcome variable is the node's class. In this case, the model parameters are not based on the classes of the nodes (as in nBC), but purely on all nodes in the network. Results using this new setup are shown in Supplementary Figure 4.7. We see that the main patterns—compared to the results using nBC—persist. Classification performance achieves its best scores with high levels of homophily and heterophily, and it drops when networks are neutral. Also, classification on heterophilic networks is just slightly better than classification on homophilic networks. However,

[10] When their neighbors belong to the majority group but are not in the sample.

[11] In fact, larger fluctuations are more likely in small samples.

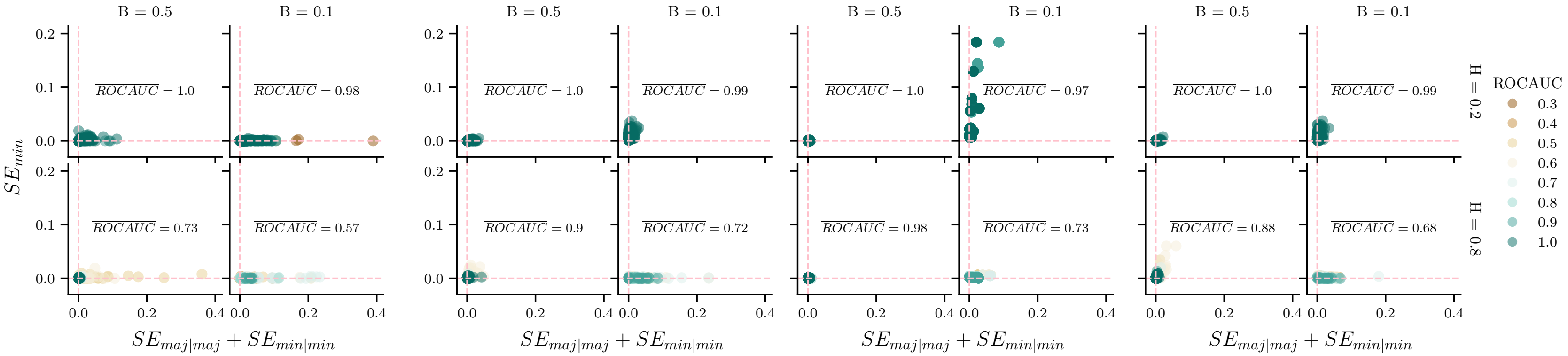
the most notorious difference between LINK and nBC is the performance across sample sizes. First, we notice that when using nBC, performance drops drastically when using small training samples on homophilic networks. Second, in this regime performance is not stable (i.e., high variance), see Figure 4.2. These two issues do not appear in the results when using LINK, see Supplementary Figure 4.7. Therefore, we can conclude that *performance, in terms of ROCAUC scores, is mainly driven by the type of network* (i.e., the interplay between homophily, class balance, edge density and preferential attachment). When it comes to sample size, nBC gets penalized by small samples since their fluctuations introduce noise in the model parameters, while the parameters of LINK never change.

4.5.2. RQ2: How does the choice of the *sampling technique* affect the overall performance of collective classification and its parameter estimation?

In Section 4.5.1 we learned that certain properties of the network structure help in the parameter estimation even when training samples are very small. Now, we compare random node sampling with three other sampling methods, two of them are *biased* towards high degree nodes (*random edge sampling* and *degree ranking*), and one is *unbiased* (*partial crawls*); more details in Section 4.3.2.

Since the focus is on the sampling techniques, we fix the number of nodes and edge density of networks to $N = 2000$ and $d = 0.004$, respectively. We also omit results on neutral networks, and large sample sizes since their performance is either consistent or often very high. Results are shown in Figure 4.4. The x-axis represents the sum of the squared estimation errors of conditional probabilities $P_{maj|maj}$ and $P_{min|min}$, the y-axis shows the squared estimation error of the class prior P_{min}, and colors represent the overall performance.

Random nodes vs. other sampling techniques. First, if we look at the estimation errors from the class prior and the conditional probabilities separately (as shown in Figure 4.4) we notice that random edges, degree sampling, and partial crawls are better at estimating conditional probabilities than random nodes. This is because conditional probabilities are based on connections between nodes and all three sampling methods exploit these connections during the sampling. Second, not surprisingly, random node sampling is on average better at estimating class priors since it observes a random sample of nodes, and the class prior only depends on the prevalence of node attributes. Third, on average degree sampling achieves the highest performance ($\overline{\text{ROCAUC}} \approx 0.91$) followed by random edges, partial crawls, and random nodes ($\overline{\text{ROCAUC}} \approx 0.81$). On the other hand, partial crawls sampling provides the most accurate estimates followed by random edges, random nodes, and degree ranking. However, depending on the structure of the network these sampling techniques may improve or worsen their overall performance and parameter estimation as described below.



(a) Random Nodes

(b) Random Edges

(c) Degree Sampling

(d) Partial Crawls

Figure 4.4.: **RQ2: Parameter estimation and overall performance on small samples.** These are the squared estimation errors of class priors (y-axis) and conditional probabilities (x-axis) with their respective overall performance (colors) on small samples ($pseeds \leq 30\%$) using different sampling techniques: (a) random nodes, (b) random edges, (c) degree sampling, and (d) partial crawls. Class balance and homophily values are shown as columns B and rows H, respectively. Mean ROCAUC scores per network type (B,H) are shown as $\overline{ROCAUC}$. Conditional probabilities obtained by partial crawls ($SE_{maj|maj} + SE_{min|min} < 0.18$) are the most accurate, followed by random edges, degree sampling and random nodes. However, the most accurate class priors are obtained by random nodes ($SE_{min} < 0.02$), followed by random edges, partial crawls, and degree rank. In terms of performance, on average degree sampling achieves the highest ROCAUC followed by random edges, partial crawls, and random nodes. Surprisingly, perfect estimates ($\sum SE = SE_{min} + SE_{maj|maj} + SE_{min|min} \approx 0$) do not guarantee perfect performance (ROCAUC ≈ 1.0), and vice versa.

Trade-off between homophily and class balance. In terms of overall performance, all sampling techniques perform equally well in heterophilic networks in both the balanced and unbalanced regimes ($\overline{\text{ROCAUC}}_{H=0.2} \approx 0.97$). Similarly, all sampling techniques perform equally well in homophilic networks ($\overline{\text{ROCAUC}}_{H=0.8} \approx 0.76$). However, this performance is proportional to the class balance: low for unbalanced networks ($\overline{\text{ROCAUC}}_{H=0.8,B=0.1} \approx 0.67$), and high for balanced networks ($\overline{\text{ROCAUC}}_{H=0.8,B=0.5} \approx 0.85$). Last but not least, we also see in Figure 4.4 that the most accurate estimates across sampling techniques are obtained in balanced networks, especially when they are also heterophilic (more details in Supplementary Figure 4.10).

Which sampling technique should we use?

If the goal is to achieve high overall performance ($\overline{\text{ROCAUC}} \approx 1.0$) with a small sample, random edge sampling or partial crawls should be used in *heterophilic* networks[12], and degree sampling in *homophilic* networks, as long as the degree of nodes is available, otherwise random edges should be considered. However, if the goal is to achieve good quality of estimates ($\sum SE = SE_{min} + SE_{maj|maj} + SE_{min|min} \approx 0$) with a small sample, then the most accurate estimates are obtained by degree ranking (followed by partial crawls) when networks are *balanced*, and partial crawls when networks are *unbalanced* (see Supplementary Figure 4.10).

4.5.3. RQ3: How does network structure and the choice of sampling technique influence the direction of bias in collective classification?

Now, we explore how classification mistakes are distributed across both classes. If mistakes are concentrated in one class, the classifier is biased against that class. For example, when data is unbalanced, a majority class classifier will be highly accurate, but misclassify—or be biased against—the minority class. To disentangle *how well the algorithm classifies both minority and majority* classes, we extend the balanced accuracy [34] to assess the direction of bias. We then compare the true positive rates (TPR) of each class as follows:

$$bias = \frac{TPR_{min}}{TPR_{min} + TPR_{maj}} \tag{4.8}$$

Since our classification task is on a binary attribute, TPR_{min} refers to *sensitivity* and TPR_{maj} refers to TNR_{min}[13] or *specificity.* This bias score ranges from 0 to 1. Depending on its value, classification can be interpreted as: (a) bias < 0.5:

[12] In this regime these sampling techniques are on average slightly better than the other methods, though all of them perform equally well.

[13] True negative rate of the minority class.

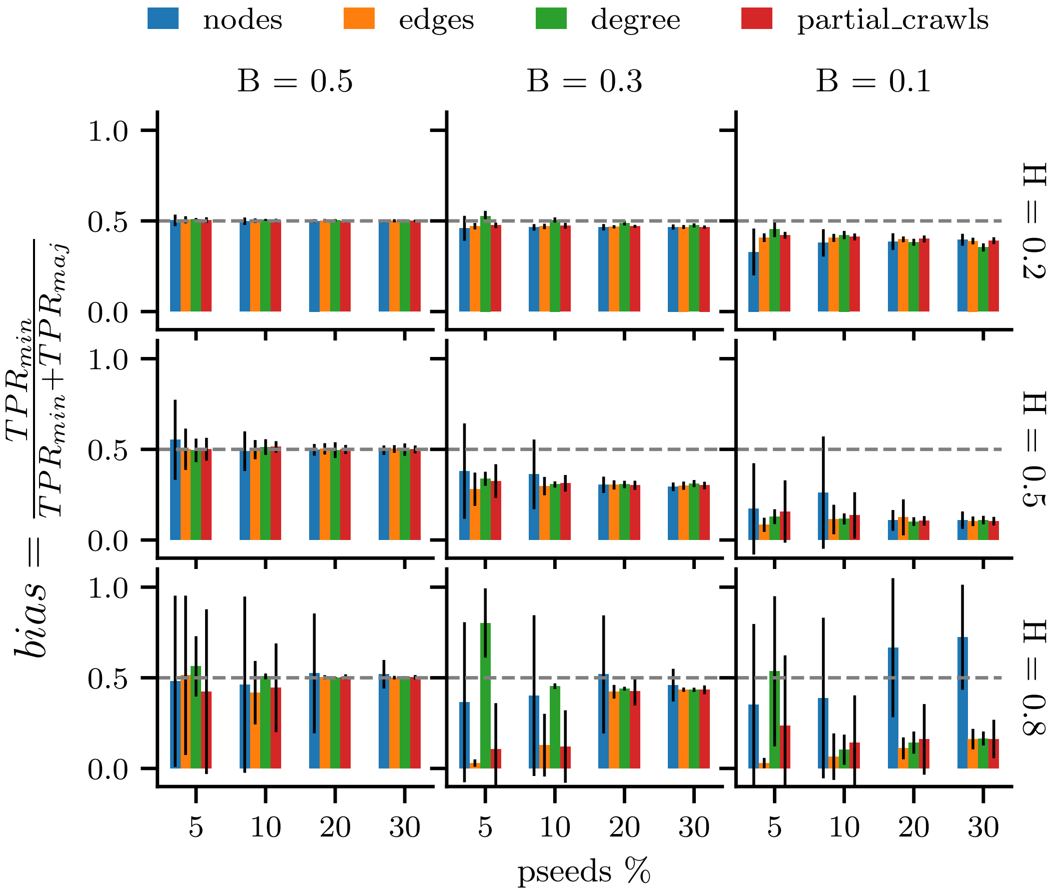


Figure 4.5.: **RQ3: Direction of bias.** We measure the direction of bias by comparing the true positive rates of each class. The unbiased case is $bias = 0.5$, when both classes have the same true positive rates. Otherwise, results are biased towards majority nodes $bias < 0.5$, or towards minority nodes $bias > 0.5$. These are observations on small samples ($pseeds \leq 30\%$) where fluctuations are high, after this point bias scores converge or get better (i.e., towards $bias = 0.5$). We see that classification on heterophilic networks is less biased than in neutral and homophilic networks.

biased towards majorities (or against minorities), (b) bias > 0.5: biased towards minorities (or against majorities), and (c) bias $= 0.5$: unbiased.

Results on networks with fixed number of nodes ($N = 2000$) and fixed density ($d = 0.004$), using the four sampling techniques, are shown in Figure 4.5. Large samples ($pseeds > 30\%$) are not shown since bias scores converge at that point for almost all cases[14].

[14] For larger samples in networks with $B = 0.1$ and $H = 0.8$, the bias score slightly increases (but it is still biased against the minority nodes).

On average, classification results are unbiased in balanced networks ($B = 0.5$). Additionally, when class balance decreases ($B < 0.5$), classification results are often biased towards majority nodes. However, depending on the level of homophily of the network, the bias score decreases considerably in neutral and homophilic networks ($H \in \{0.5, 0.8\}$), or just slightly in heterophilic networks ($H = 0.2$). Notice as well that in the homophilic regime the standard deviation is high. This means that the variation with respect to which group is classified correctly is high. These results are consistent across all sampling methods, and indicates that unbiased results are more robust to changes in group-size and sampling choice in heterophilic networks than in neutral and homophilic networks.

Surprisingly, there are a few cases where classification is biased against majority nodes (bias > 0.5). Specifically in homophilic networks when nodes are sampled randomly (blue) or by degree (green). Thus, their classification performance is low $ROCAUC < 0.6$ (see Supplementary Figures 4.11(a) and 4.11(c)). On the other hand, when classification results are unbiased (bias $= 0.5$) or biased against minority nodes (bias < 0.5), their classification performance is high $ROCAUC > 0.8$ and inversely proportional to the sum of estimation errors (see Supplementary Figure 4.11).

Future work should investigate new sampling methods or classifiers that focus on overcoming the bias issue of collective classification especially in homophilic and neutral networks. For instance, one promising direction that has been proven to improve performance, especially for neutral networks, is to look at friends-of-friends similarities in the parameter estimation [6].

4.6. Empirical Networks

Finally, we focus on five real-world networks, described in Table 4.1, and show that the utilized network model [133], allows for computing a baseline for the performance that a collective classifier can achieve on empirical social networks.

Real-world networks. *Sexual contact network:* The Escorts dataset represents a network of sexual contacts from Brazil [212]. Nodes are of two types: client or escort. *Friendship networks:* Swarthmore42 and Caltech36 are University networks which include friendship links between user's Facebook pages [249]. Every node in each network represents a member of the school. For the purpose of our experiments we choose the attribute $gender \in \{1(\text{female}), 2(\text{male})\}$ as class label. *Hyperlink network:* This is a hyperlink network of American politicians in Wikipedia [256]. We consider reciprocal edges in order to treat the network as undirected. We use the politician's gender as class label. *Following network:* The GitHub dataset is a large social network of GitHub developers [216]. Nodes are developers who have starred at least 10 repositories and edges are mutual follower relationships between them. Nodes possess a binary attribute that describes whether the user is a web or a machine learning developer.

Table 4.1.: **Empirical networks.** Structural properties of five real-world networks: Escorts, Swarthmore42, Caltech36, Wikipedia, and GitHub. In addition to the properties of interest, we report β, the power-law exponent of the degree distribution computed as described in [133]. N_{fit} and m_{fit} represent the number of nodes and minimum degree utilized to generate synthetic networks, respectively.

dataset	Escorts	Swarthmore42	Caltech36	Wikipedia	GitHub
N	16730	1519	701	2132	37700
m	1	1	1	1	1
class	role	gender	gender	gender	dev
minority	escort	2 (m)	1 (f)	female	1 (ML)
B	0.40	0.49	0.33	0.15	0.26
E	39044	53726	15464	3143	289003
d	0.0003	0.05	0.06	0.001	0.0004
β	2.87	5.50	4.90	2.87	2.54
H	0.00	0.52	0.54	0.64	0.84
N_{fit}	14338	208	179	2893	9830
m_{fit}	2	2	2	2	2

We remove nodes without class label, and nodes with no edges. Note that all these networks are scale-free (i.e., power-law degree distributed). Table 4.1 shows the value of the power-law exponent β of each dataset, and Supplementary Figure 4.12 shows their degree distributions.

BA-Homophily networks. For each dataset we generate 5 synthetic networks using the BA-Homophily model with symmetric homophily (e.g., $h_{aa} = h_{bb}$), and adapt the algorithm to fulfill the edge density condition [75]. We pass to the algorithm five parameters shown in Table 4.1: number of nodes N_{fit}, class balance B, homophily H, edge density d, and minimum degree m_{fit}[15].

[15] Although m is required for the BA-Homophily model, it does not affect the classification results because the behavior of the network is independent of m [133]

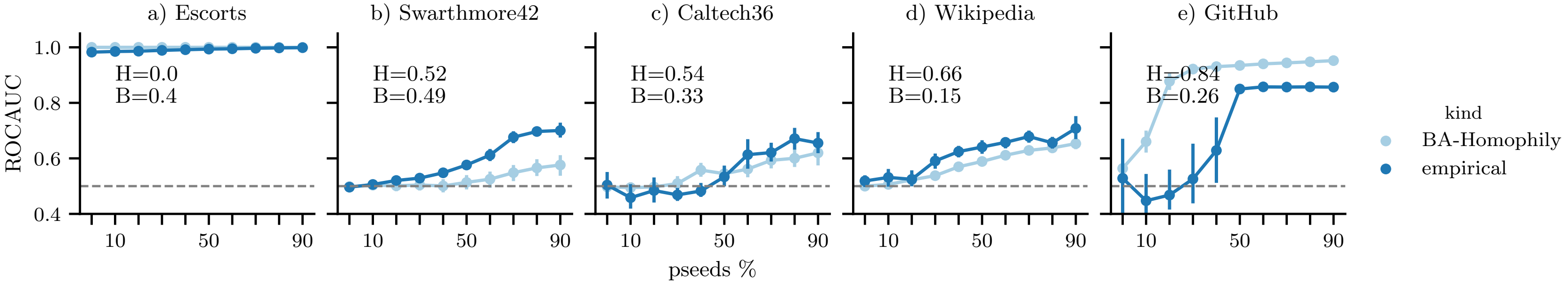


Figure 4.6.: **Classification performance on empirical networks.** We run the collective classification algorithm using random node sampling on five networks: (a) sexual contact network, (b,c) university networks, (d) reciprocal hypher-link network and (e) developer-follower network. Properties of the networks are shown as H for homophily and B for class balance. ROCAUC scores using different sample sizes are shown on the y- and x-axis, respectively. Results from real networks are shown as “empirical” (dark blue), and their synthetic counterparts as “BA-Homophily” (light blue). Overall, results from synthetic networks follow a similar pattern as the empirical counterpart, except GitHub when samples are small.

Results on empirical networks

As in Section 4.4, we run the collective classification algorithm 10 times for each sample size, and report its performance using mean ROCAUC scores. In Figure 4.6 we see the classification performance (y-axis) on five real-world networks (columns) for different sample sizes (x-axis). The (expected) ROCAUC using each synthetic network is shown as “BA-Homophily” (light blue), and the (observed) ROCAUC using the real networks is shown as “empirical” (dark blue). In general, we see that the synthetic networks are able to mimic the collective classification performance of real-world networks. As expected, the Escorts network has the best fit between data and model, because it is extremely heterophilic (see Section 4.5.1). The Caltech and Wikipedia networks also show very good fit. Note that Swarthmore42 has a perfect fit only for small samples ($pseeds < 50\%$). Conversely, when training samples are small, GitHub shows high variance. These discrepancies might be due to other network properties that the model does not capture. For instance, rich mixing patterns might get ignored by summarizing homophily with a global statistic [199, 200]. In other words, while the global behavior is captured in the sample, some nodes can exhibit local differences. Similarly, real-world networks might exhibit asymmetric homophily ($h_{min,min} \neq h_{maj,maj}$) and high clustering [118].

Additionally, note that results in Figure 4.6 have been sorted by H (homophily) in ascending order. We see a similar pattern in Figure 4.2, where ROCAUC scores besides getting higher with larger sample sizes, they also get higher upper-bounds with values of homophily far from $H = 0.5$.

4.7. Discussion and Future Work

Many popular applications rely on peer information and other types of relational data. For instance, peer-to-peer lending systems [13] allow people to borrow/lend money to connected friends. Such systems utilize machine learning algorithms to infer credit scores of individuals using the credit score of their friends (i.e., high/low risk) [107, 157, 160]. In such cases, it is extremely important to understand and explain the overall performance of these algorithms, as well as their impact on different groups, especially minorities. Our results highlight that especially in homophilic and neutral networks, minorities may be at a disadvantage when the classifier is trained on a small sample.

Additionally, we show that homophily, class balance, edge density, and sample size impact the performance and the direction of bias of collective classification. Interestingly, we find that larger training samples are required in undirected networks with homophilic connections to achieve a similar high classification performance compared to heterophilic networks. This fundamental difference between heterophilic and homophilic networks can be explained by the sampling error and the network structure; in particular undirected edges, class balance, homophily, and preferential attachment (see Section 4.5.1). This suggests that the structure of a social network can help to infer the needed sample size to achieve high classification performance, and to be aware of potential bias issues.

Our comparison of sampling techniques suggests that although partial crawls sampling provides the most accurate estimates (i.e., the prior and conditional probabilities learned from their training samples are closest to what we observe in the full network), the inherent bias of degree and edge sampling (towards high degree nodes), help the classifier to achieve very high performance. This suggests that accurate estimates do not always lead to perfect classification performance, and vice versa. Concretely, we observed that when sampling by degree on heterophilic and unbalanced networks, perfect classification performance can also be achieved based on imperfect estimates, in particular the class prior. This can be explained by the fact that in heterophilic networks, minority nodes have high degree, and heterophilic edges are predominant. Thus, a very small sample (sampled by degree) will mostly contain minority nodes and a few majority nodes that are linked to the minorities (see Section 4.5.1). Surprisingly, we also find that perfect estimates do not guarantee perfect classification performance, especially in homophilic and unbalanced networks across all sampling techniques. One explanation could be that only a few unlabeled nodes are connected to the seed nodes as 1-hop neighbors. Thus, long cascades of unlabeled nodes (in k-hops; $k > 1$) might propagate erroneous information. Further studies are needed to explain this behavior.

Use case. Due to the design of our experiments, we always had access to the class balance and homophily of the network. In a real scenario, however, these properties might be unknown. Then, how can we evaluate results on new datasets if the structure of the network is unknown? It has been shown in [12] that partial crawls can obtain unbiased and reliable node and edge statistics. We corroborated that in Section 4.5.2. Therefore, in cases when these properties are unknown, a small sample taken by partial crawls can already capture accurately the structure statistics of the whole network, e.g., class priors and conditional probabilities (see Section 4.3.3).

Limitations. Our results are limited to undirected networks to simplify our exposition. However, results can and should be extended to directed networks including larger sets of attributes per node, and different levels of asymmetry between intra-group connections (i.e., asymmetric homophily) in both directed and undirected networks. Additionally, while in this work we focused on social networks with preferential attachment and homophily, further research may investigate the impact of other mechanisms of edge formation on relational classification. For instance, by disentangling the effects of homophily and triadic closure [8, 118]. Finally, we focus on one specific collective inference method (relaxation labeling) and mostly on one relational model (network-only Bayes classifier, nBC). The comparison of multiple methods is a potential avenue to disentangle algorithmic bias. For instance, when we compared results using nBC and LINK (see Section 4.5.1), we learned that their main differences rely on the training sample size. In particular, we found that the model parameters of nBC get distorted when training samples are small. These fluctuations affect performance since the nBC parameters must learn the correct class balance and homophily of the network in order to increase the chances of achieving high performance. Even when these parameters are correct, if the training sample is too small, unlabeled nodes might not reach sufficient seed nodes and this might lead to erroneous collective inference. This does not occur in LINK since its parameters only depend on the presence or absence of a node as a neighbor.

4.8. Conclusions

Collective classification is often used to infer missing attributes of nodes in networks. However, which factors impact its performance? And under which conditions is inference biased towards minority or majority groups? This paper provides answers to these questions by systematically analyzing the impact of network structure and sampling technique on the performance of collective classification. Our findings suggest that (i) mean classification performance can empirically and analytically be predicted by homophily, class balance, edge density, and sample size, (ii) networks with homophilic connections require larger training samples than heterophilic networks to achieve comparable performance, (iii) when sampling budgets are small, on average, partial crawls and edge sampling achieve the most accurate model estimates, and (iv) classification results are often less biased in heterophilic networks than in homophilic networks regardless of class imbalance.

4.9. Supplementary Material

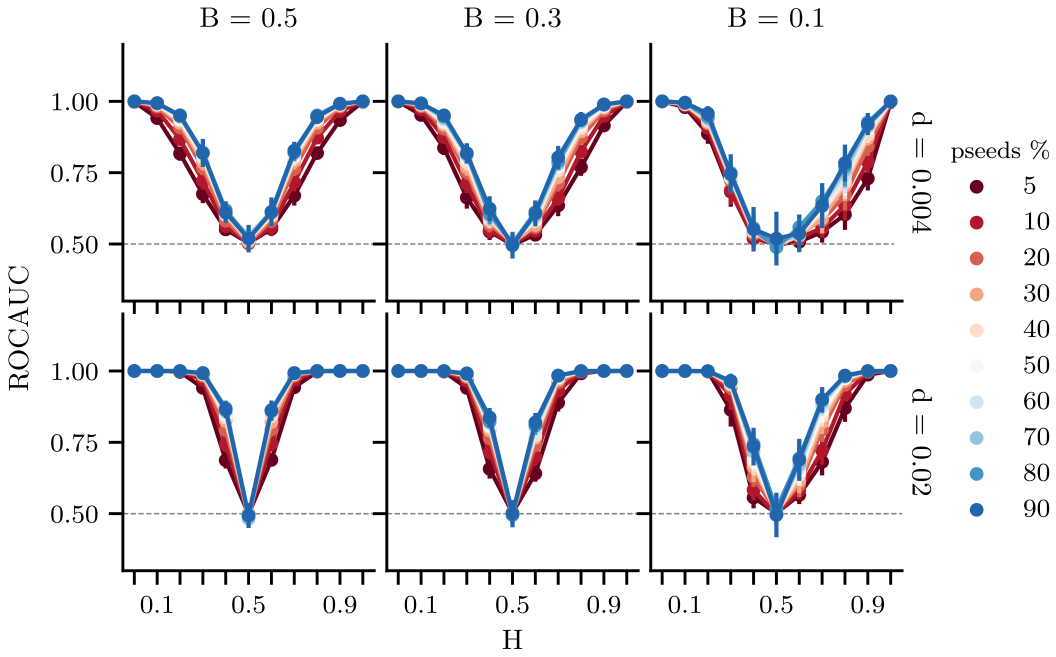


Figure 4.7.: **RQ1: Network structure, sample size and overall performance using random node sampling and LINK classifier.** These results reflect the interplay between network structure, sample size and classification performance. In Figure 4.2 results were obtained using the network-only Bayes classifier (nBC), while in this figure results were obtained using the LINK classifier. Overall, the main patterns persist: Classification on neutral networks is as good as a random classifier, and classification performance is best in the extreme heterophilic and homophilic regimes. However, results using LINK are more consistent across samples sizes (i.e., performance differences between small and large samples are small), and even within small training samples (i.e., variance is low). Also, ROCAUC scores using nBC are slightly higher than the ones obtained with LINK.

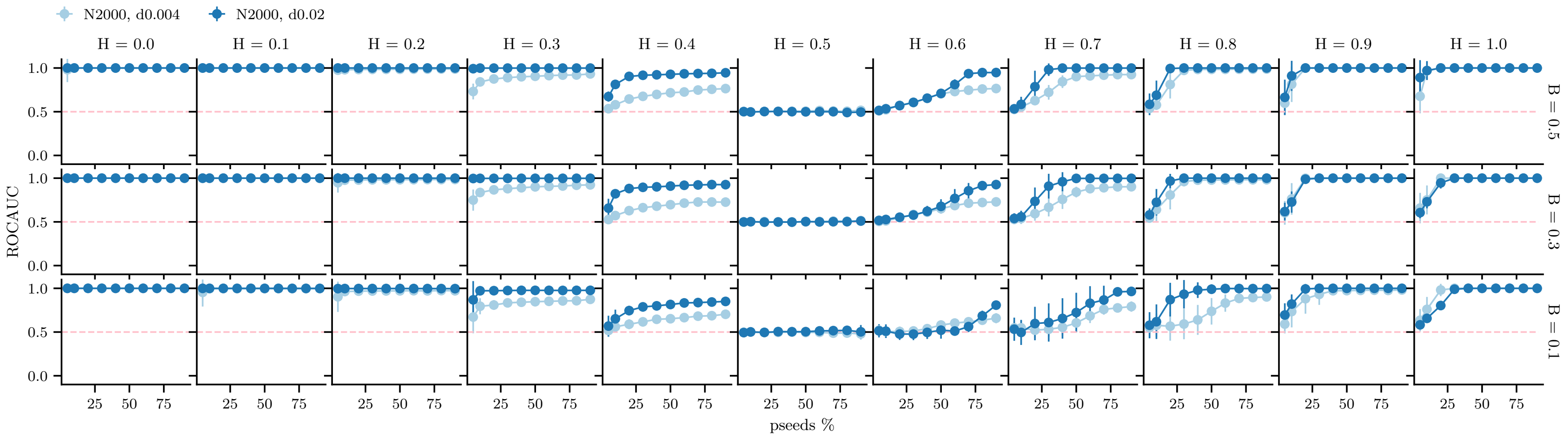


Figure 4.8.: **RQ1: Overall performance vs. network structure.** Columns represent homophily values and rows capture class balance (or fraction of minorities). The y-axis shows ROCAUC scores (performance) and the x-axis the percentage of nodes in the training sample. Colors refer to the type of network: dark blue represents *dense* networks with 2000 nodes and edge density of 0.02, and light blue lines refer to *sparse* networks with density of 0.004 and the same number of nodes. The main difference between sparse and dense networks is visible around neutral networks ($H = 0.5 \pm 0.2$). In these cases, higher density improves performance. We also see that in all cases, the larger the training sample, the higher the performance, specially in homophilic networks ($H > 0.5$). Heterophilic networks ($H < 0.5$), on the other hand, always achieve perfect performance ($ROCAUC \approx 1.0$) regardless of sample size. Performance is worst on neutral networks ($H = 0.5$).

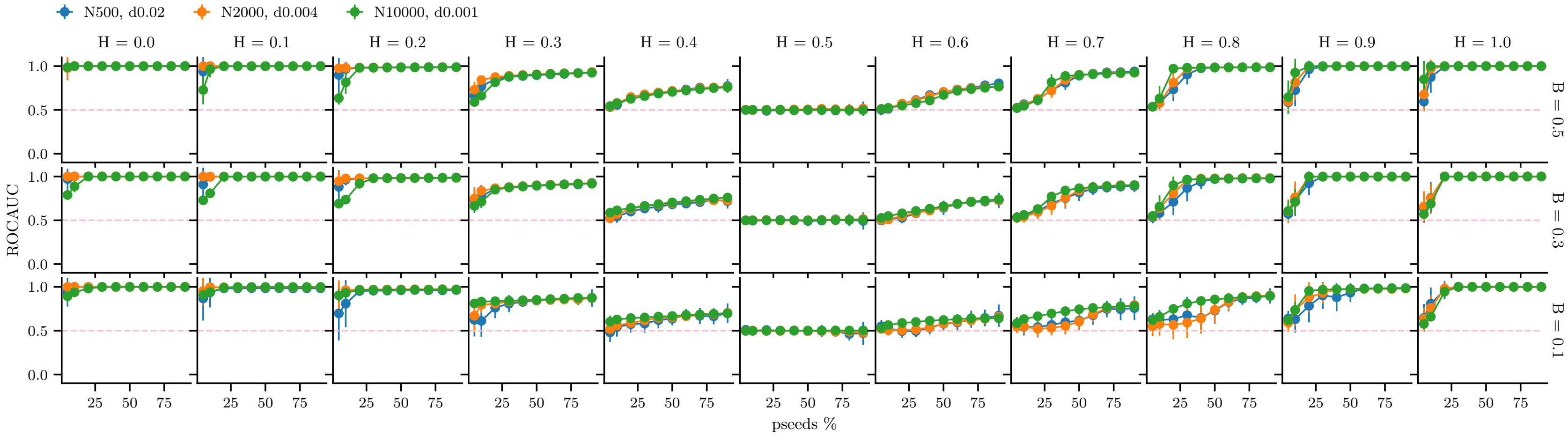
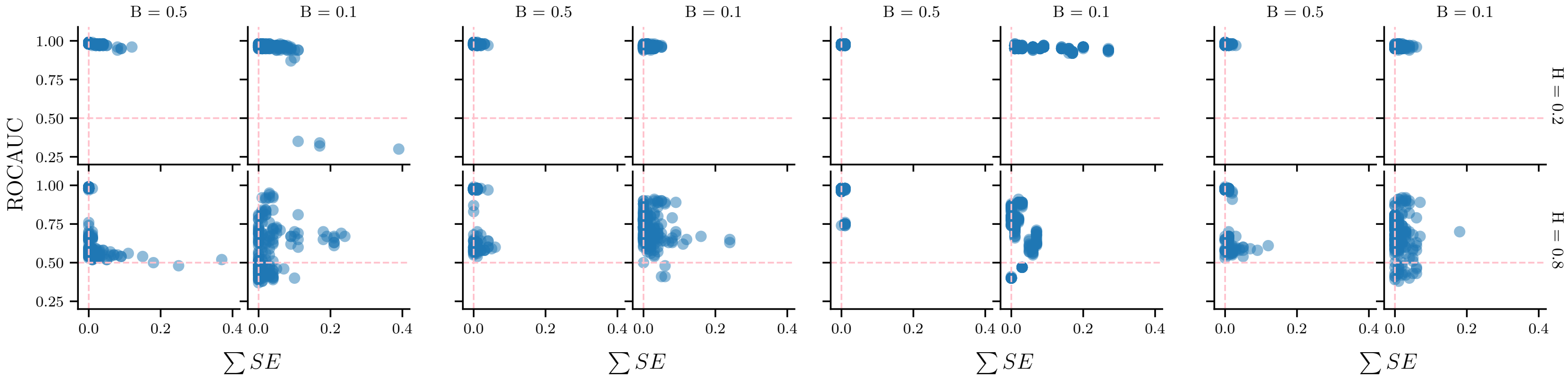


Figure 4.9.: **RQ1: Overall performance vs. network size.** Columns represent homophily values and rows capture class balance (or fraction of minorities). The y-axis shows ROCAUC scores (performance) and the x-axis the percentage of nodes in the training sample. Colors refer to the size of the network: blue, orange and green represent networks with 500 (small), 2000 (medium) and 10000 (large) nodes, respectively. All these networks possess a minimum degree of $m = 4$. In other words, they are sparse with density $\{0.02, 0.004, 0.001\}$ for small, medium and large networks, respectively. Overall, the mean performance is very similar across all network sizes. The main difference is that the larger the network, the lower the variance in performance (i.e., more stable). Moreover, we see that larger networks help the classification to be slightly more accurate, especially for small samples. However, when networks are large, heterophilic, and group size differences are small ($N = 10000$, $H \leq 0.3$, $B \leq 0.3$), mean performance drops when samples are small ($pseeds < 20\%$).



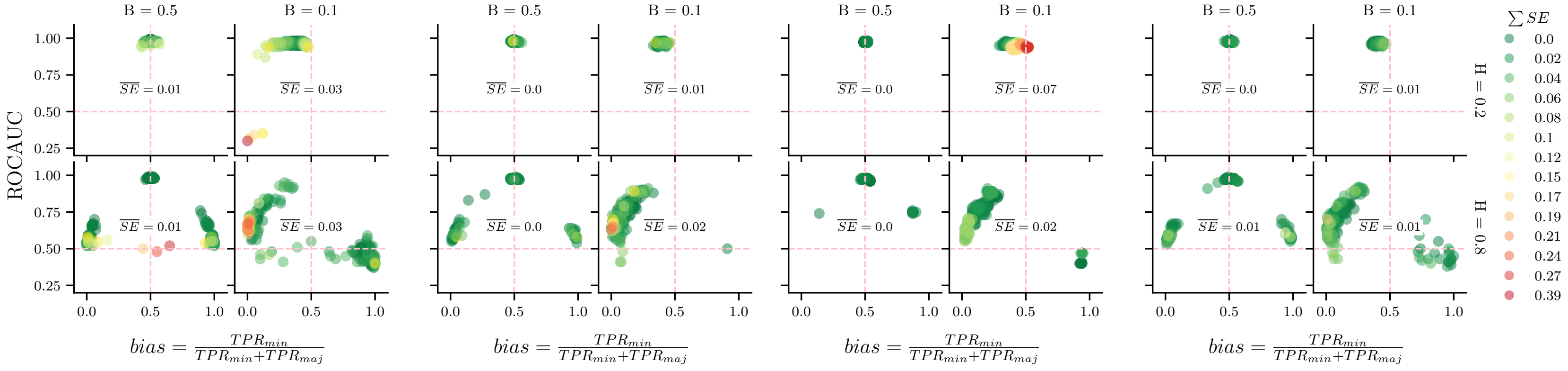
(a) Random Nodes

(b) Random Edges

(c) Degree Sampling

(d) Partial Crawls

Figure 4.10.: **RQ2: Overall performance vs. overall parameter estimation on small samples.** These are the sum of squared estimation errors of class priors and conditional probabilities (x-axis) with their respective ROCAUC scores (y-axis). These are observations on small samples ($pseeds \leq 30\%$) using different sampling techniques: (a) random nodes, (b) random edges, (c) degree sampling, and (d) partial crawls. We see that across sampling techniques, heterophilic networks are consistent w.r.t performance, except a few cases in random nodes. In contrast, performance in homophilic networks tends to be more noise specially in terms of performance (i.e., high variance). Estimations errors are higher when sampling by random nodes and random edges.



(a) Random Nodes (b) Random Edges (c) Degree Sampling (d) Partial Crawls

Figure 4.11.: **RQ1, RQ2, RQ3: Overall performance, bias, and parameter estimation on small samples.** This plot captures the relationship between network structure (class balance B as columns, homophily H as rows), sample size (only small samples *pseeds* $\leq 30\%$), parameter estimation (sum of estimation errors SE as colors), bias (x-axis) and classification performance (y-axis) using random node sampling. $\overline{SE}$ values refer to the mean sum of estimation errors in the corresponding type of network. We see the following patterns: (i) Heterophilic networks achieve the highest ROCAUC scores and their parameter estimation is almost perfect (i.e., sum of errors is close to zero; green), except a few cases in degree sampling where ROCAUC is high and errors are high (as shown in Section 4.5.2 and Section 4.7, this is due to erroneous class prior estimation). (ii) Unbiased results on heterophilic networks correlates with high performance and low estimation errors on unbalanced networks. Results on heterephilic and unbalanced networks can be only slightly biased towards majority nodes without affecting much performance and estimation of model parameters. (iii) Unbiased classification in homophilic networks correlates with high performance and perfect parameter estimation. This only occurs when networks are balanced. Unbiased results often correlate with low performance. When homophilic networks are unbalanced, this bias can affect minorities but also majority nodes.

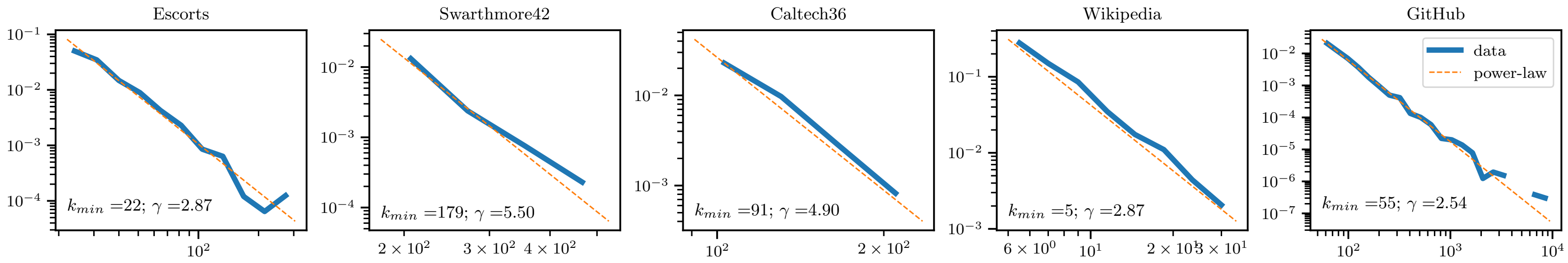


Figure 4.12.: **Degree distribution of empirical networks.** Our results in Section 4.5 are based on synthetic networks. These findings generalize to other existing real-world networks under the assumption that edges form following the preferential attachment and homophily mechanisms. In this plot we show that the empirical networks used in our experiments (see Section 4.6) are scale-free networks.

5. The Influence of Edge Formation in Ranking

This chapter proposes evaluation benchmarks to better understand the influence of network structure on the outcome of ranking and recommendation algorithms. In particular, we investigate and assess inequality and inequity in the distribution of opportunities across individuals and minorities in top ranks. This work has been published as a full paper in the journal of Scientific Reports, 2022.

Though algorithms promise many benefits including efficiency, objectivity and accuracy, they may also introduce or amplify biases. Here we study two well-known algorithms, namely PageRank and Who-to-Follow (WTF), and show to what extent their ranks produce *inequality* and *inequity* when applied to directed social networks. To this end, we propose a **d**irected network model with **p**referential **a**ttachment and **h**omophily (DPAH) and demonstrate the influence of network structure on the rank distributions of these algorithms. Our main findings suggest that (i) inequality is positively correlated with inequity, (ii) inequality is driven by the interplay between preferential attachment, homophily, node activity and edge density, and (iii) inequity is driven by the interplay between homophily and minority size. In particular, these two algorithms *reduce*, *replicate* and *amplify* the representation of minorities in top ranks when majorities are homophilic, neutral and heterophilic, respectively. Moreover, when this representation is reduced, minorities may improve their visibility in the rank by connecting strategically in the network. For instance, by increasing their out-degree or homophily when majorities are also homophilic. These findings shed light on the social and algorithmic mechanisms that hinder equality and equity in network-based ranking and recommendation algorithms.

5.1. Introduction

Online social networks and information networks have become integral parts of our everyday life. However, the opportunities offered by such networks are often constrained not only by our previous interactions [32, 37, 38, 50, 182], but also by algorithms. For instance, algorithms could make some people or content more visible than others via classification [77], ranking or recommendations [2]. In this regard, search engines and recommender systems are increasingly used for various applications such as whom to follow, whom to cite, or whom to hire. Typically, these applications use algorithms to order items (e.g., people and academic papers) based on “importance” or “relevance”, and may therefore produce social inequalities by discriminating certain individuals or groups of people in top ranks. In fact, it has been shown that recommender systems such as *Who-to-Follow* (WTF) [106] tend to increase the popularity of users who are already popular [2, 24, 244]. A similar effect has been found in *PageRank* [193], where nodes in high ranks stabilize their position and give little opportunity to other nodes to occupy higher positions [97]. This tendency towards the “popular” arises because these algorithms harness structural information, in particular, the in- and out-degree of nodes. For this reason, modeling the directionality of links—which is often left out for simplicity—is crucial to really understand how these algorithms work on different types of networks.

However, social networks are complex systems, and many other structural properties may also alter the distribution of nodes and groups in the ranking. For example, previous studies have shown that *homophily*—the tendency to connect to similar others—affects the visibility of minorities in degree rankings [133] and people recommender systems [83]. Consequently, it can reinforce societal issues such as the glass ceiling effect [11, 53, 243] and the invisibility syndrome [89]. Despite these findings, little is known about the extent to which the combination of multiple structural properties can alter the visibility of minorities in top ranks from ranking and recommendation algorithms. A further complication is that debiasing ranking outcomes and making them fair is very challenging since they can be mitigated in different ways [273]: by intervening on the score distribution of candidates [140], on the ranking algorithm [9], or on the ranked outcome [269]. While most of these studies tackle fairness in ranking, they do not explore the effects of networked data in ranking. This paper is a step towards this goal. Since such algorithms are so deeply involved in social, economic, and political processes, we need to first understand how our connections affect them to then apply appropriate interventions towards fair results.

To this end, we propose DPAH, a network model that generates directed scale-free networks with binary-attributed nodes. It encodes two main mechanisms of edge formation found in social networks: *homophily* and *preferential attachment* [18, 30, 171] (see Section 5.5 for more details). Moreover, it allows to control for the *fraction of minorities*, *edge density*, and the *skewness of the out-degree distribution*. By using this model, we systematically study how these structural properties of social networks impact the ranking of nodes in PageRank and WTF. In particular, we investigate two ranking issues, inequality and inequity, and show how they get affected by the ranking algorithm together with the type of network. We measure *inequality* by quantifying the skewness of the rank distribution of nodes that PageRank and WTF produce, and *inequity* as how well-represented the minorities are in the top of the rank compared to the proportion of minorities in the network. In this work we study both ranking issues and measure their correlation. Furthermore, we quantify them globally using the whole rank distribution, and locally within each top-k% rank. The goal is to identify both the overall inequality and inequity trend that these algorithms

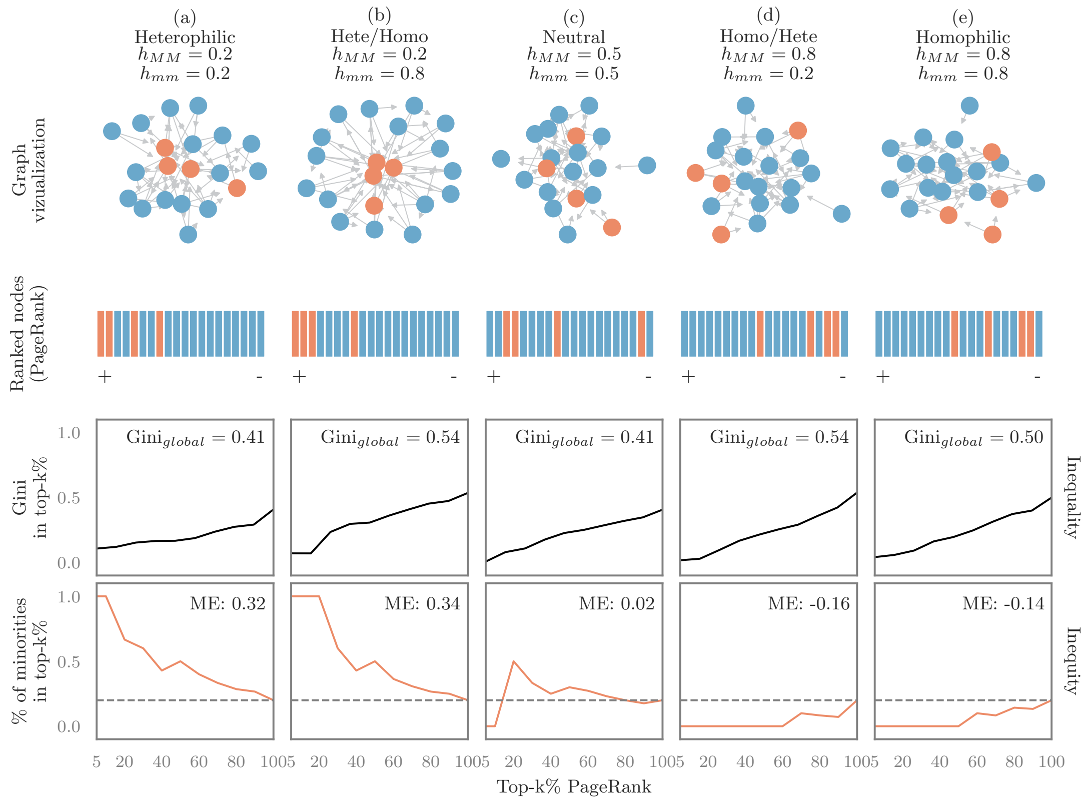


Figure 5.1.: **Inequality and inequity.** Every column represents a network with certain level of homophily. All networks contain 20 nodes: 20% belong to the minority group (orange), and 80% to the majority group (blue). Edges follow a preferential attachment with homophily mechanism. The top row shows the graph and the level of homophily within groups (MM: majorities and mm: minorities). The second row shows all nodes in descending order (from + to -) based on their PageRank scores. The third row represents the rank *inequality*: Gini coefficients of the rank distribution for every top$-k$% (black line). Gini_{global} refers to the Gini coefficient of the entire rank distribution (i.e., at top-100%). We see that the lower the k, the lower the Gini of the rank distribution. The bottom row represents the rank *inequity*: Percentage of minorities found in each top-k% of the rank distribution (orange line). ME is the mean error of these percentages compared to a fair baseline or diversity constraint (i.e., how much the orange line deviates from the dotted line across all top-k's). Here we see three main patterns: (a,b) When the majority group is heterophilic, minorities are on average over-represented, $ME > 0.0$. (d,e) When majorities are homophilic, minorities are on average under-represented, $ME < 0.0$. (c) When both groups are neutral, the observed fraction of minorities is almost as expected, $ME \approx 0$.

produce, and the tipping points where minorities start gaining visibility in the top of the rank.

As an example, consider the *directed networks* shown in Figure 5.1. Every column represents a network with two types of nodes, minority (orange) and majority (blue), and different levels of homophily within groups. Homophily h, is a parameter ranging from 0 to 1 and determines the tendency of two nodes of the same color to be connected. h_{MM} and h_{mm} represent homophily within majorities and minorities, respectively. When nodes are ranked using PageRank (second row), the position of the minorities in the rank varies *systematically.* For instance, when majorities are heterophilic ($h_{MM} = 0.2$, columns a and b), minorities often appear at the top (+). In contrast, when majorities are homophilic ($h_{MM} = 0.8$, columns d and e), minorities tend to appear at the tail of the rank (-). Next, we explain this systematic ranking behavior in top ranks by further varying the structure of the network.

5.2. Results

5.2.1. Inequality and inequity in ranking

Inequality refers to the dispersion or distribution of *importance* among *individuals.* This importance is the ranking score assigned to every node by the algorithm. We compute the *Gini* coefficient of the rank distribution to measure how far the ranking scores of individuals deviate from a totally equal distribution (see Methods for more details). As shown in Figure 5.2, a very low Gini score (Gini < 0.3) means that individuals are very similar with respect to their ranking scores. If the Gini score is extremely high (Gini ≥ 0.6), it means that only a few individuals capture most of the rank. In other words, the rank distribution is very skewed. Values in between ($0.3 \leq$ Gini < 0.6) represent moderate skewed distributions. Note that we measure inequality globally by using the whole rank distribution, and locally for each top-k%. From our example in Figure 5.1, we see that PageRank on average generates moderate skewed ranking distributions for all the depicted networks ($Gini_{global} \approx 0.5$). However, for very small top-k%'s, the Gini is very low. This means that the top individuals possess very similar ranking scores.

Inequity refers to *group* fairness. In particular, it measures the error distance between the fraction of minorities in the top-k% and a given fair baseline (e.g., a diversity constraint or quota). This baseline may be adjusted depending on the context of the application [64, 234, 273]. Here, a ranking is fair when its top-k% preserves the proportional representation of groups in the network (i.e., equivalent to demographic parity [66, 273]). Therefore, the error represents the local inequity at each top-k%, and ME the mean of these errors across all top-k% ranks or global inequity. As shown on the last row of Figure 5.1, we measure the *local inequity* in two steps. First, we compute the fraction of minorities that

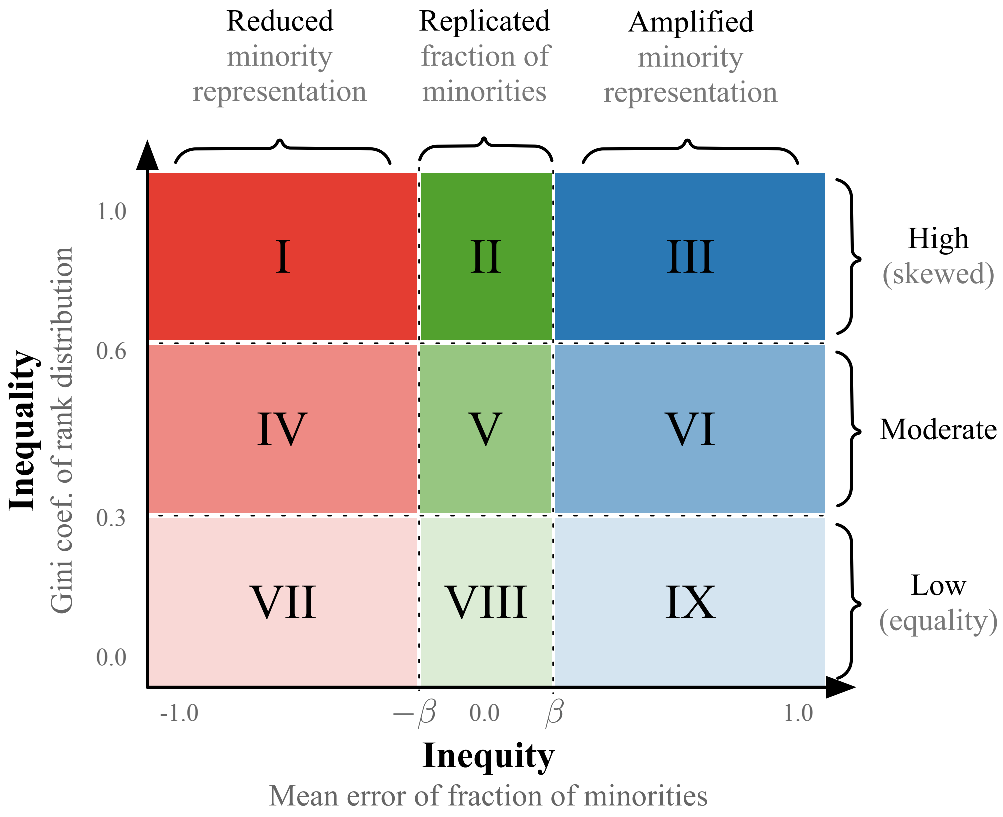


Figure 5.2.: **Regions of disparity.** We measure *inequality* (y-axis) as the skewness of the rank distribution, and *inequity* (x-axis) as the mean differences between the proportional representation of groups in top-k% ranks and the network. Highly skewed distributions lie in regions I to III (darker colors), and fair rankings, where minorities are well represented in top ranks, lie in regions II, V, VIII (green). We set $\beta = 0.05$ which is arbitrary and allows for a flexible region of *group fairness*.

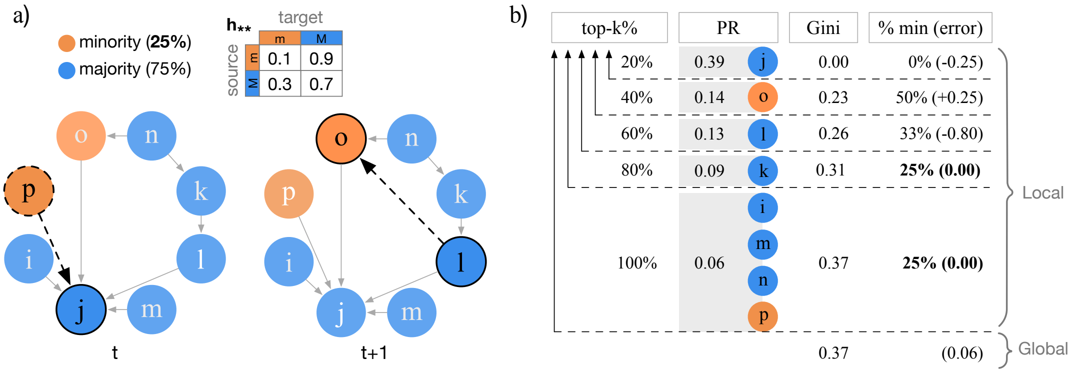
appear in each top-k% rank (orange line). Second, we compute the error between the observed fraction of minorities in each top-k% rank and a fair baseline (e.g., the actual fraction of minorities in the network, in this example 20%). Then, we average these error scores across all top-k% ranks to determine the *global inequity* score (ME values). Ideally, a fair ranking should reach $ME = 0$. However, in order to allow for small fluctuations we introduce the smoothing factor β. Thus, a fair ranking is such that $-\beta \leq ME \leq \beta$. The value of β is arbitrary, and allows for a smooth definition of "low mean error" or fairness. We set $\beta = 0.05$. As shown in Figure 5.2, when $ME > \beta$, then minorities are over-represented in the top-k% (blue region). When $ME < -\beta$, then minorities are under-represented (red region), otherwise the ranking is representing very well the minorities in the top of the rank (green region). Alternatively, we can say that the top rank (i) *replicates* the proportional representation of groups when ME is zero or very low, (ii) *amplifies* the representation of minority nodes when $ME > \beta$, and (iii) *reduces* the representation of minority nodes—and benefits the majority group—when $ME < -\beta$. Note that $ME \approx 0$ may be an artifact of a numerical cancellation as in (c), the neutral case in Figure 5.1. In such cases, we could argue that the ranking is still fair since overall it was biased towards both groups across all top-k's.

Finally, we refer to the relationship between inequality and inequity as *disparity*. For example, if a ranking distribution achieves $Gini = 0.65$ and $ME = 0.5$, we say that the disparity lies in the region III (dark blue), i.e., high inequality and high inequity, see Figure 5.2.

5.2.2. Growth network model with homophily and directed links

In order to examine the effect of homophily on the ranking of minorities in social networks, first we need to develop realistic network models that capture not only a variety of group mixing, but also the directionality of links. Many online social networks are directed networks in their nature, including the follower-followee structure on Twitter, citation networks [143], and the hyperlink structure of the Web. Directed links are the key components of many algorithms such as Google Scholar [215], PageRank and Who-to-Follow.

To this end, we propose DPAH, a **d**irected **p**referential **a**ttachment with **h**omophily network growth model. We generate these networks by adjusting the number of nodes $n = 2000$, the edge density $d = 0.0015$, the fraction of minorities $f_m \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, the in-class homophily $h_{MM}, h_{mm} \in \{0.0, 0.1, ..., 1.0\}$, and the power-law exponents of the activity distributions $\gamma_M = \gamma_m = 3.0$. We refer to the minority group as m, and to the majority group as M. Note that the between-class homophily is the complement of the in-class homophily. That is, $h_{Mm} = 1 - h_{MM}$ and $h_{mM} = 1 - h_{mm}$. Furthermore, an activity score is assigned to every node. This score is drawn from a power-law distribution and determines



top-k%	PR		Gini	% min (error)
20%	0.39	j	0.00	0% (-0.25)
40%	0.14	o	0.23	50% (+0.25)
60%	0.13	l	0.26	33% (-0.80)
80%	0.09	k	0.31	**25% (0.00)**
100%	0.06	i m n p	0.37	**25% (0.00)**
			0.37	(0.06)

Figure 5.3.: **DPAH model and ranking of nodes.** a) Illustration of the directed network model with preferential attachment and homophily (DPAH). First, $n = 8$ nodes are created and randomly labeled according to the fraction of minorities $f_m = 0.25$. Then, the following algorithm repeats until a desired edge density is fulfilled. At time t, a source node p is drawn from a power-law (activity) distribution, and a target node j is drawn with a probability proportional to the product of its in-degree $k_j^{in} = 4$ and the pair-wise homophily $h_{pj} = h_{mM} = 0.9$. At time $t+1$, a new edge is added between nodes $l \rightarrow o$ based on the same mechanism. b) The PageRank score of each node is shown under PR. Nodes in each top-k% of the rank are grouped based on the unique PageRank scores. In this example, the top-60% of nodes concentrate most of the PageRank and their scores are somewhat similar (i.e., low Gini). Also, the ranking is fair from top-80% onwards, since they capture the same fraction of minorities as in the population, 25%. Local values are measured per top-k%, and global values are measured using the whole distribution for inequality (Gini), and the average across all top-k% ranks for inequity (mean error).

with what probability the existing node becomes active to create additional links to other nodes. This means that more active nodes possess higher out-degree (see Section 5.5 for more details). Each combination of network structure is generated 10 times, nodes are ranked using PageRank and WTF separately, and inequality and inequity scores are computed and averaged across network types (and top-k's for local disparity) for each algorithm.

Figure 5.3 (a) illustrates the generation of a network using the DPAH model. First, n labeled nodes are created. In this example $n = 8$. Then, at time t, node p is selected as source node with a probability proportional to its activity. Then, p connects to an existing node j with a probability related to their pair-wise homophily h_{pj} and preferential attachment that is based on k_j^{in}, the in-degree of node j. By this process, we ensure that the out- and in-degree distributions of nodes follow seemingly power-law distributions that have been observed in many large social networks [255]. The algorithm stops once the network reaches an expected density. Note that source nodes can be either new nodes joining the network for the first time (e.g., node p at time t) or existing nodes (e.g., node l at time $t+1$). Since the network size is given, "a node joining the network for the first time" is a 0-degree node that has been selected to create its first edge. Once a source node connects to a target node successfully, the source node becomes available in the next rounds to become a target candidate. This means that in the beginning the model faces a cold start problem since there are no existing (target) nodes to connect to. Thus, the first 1% of new edges are between a source node (drawn from the activity distribution) and any other node with probability as in Equation (5.3). For the sake of completeness we show the computation of local and global disparities of this network in Figure 5.3 (b).

5.2.3. How do homophily and directional links influence the ranking of minorities globally and locally?

Global disparity. As expected, we found that the Gini coefficient of the rank distributions is large. $Gini_{\text{global}} \geq 0.6$ (regions I, II and III; dark colors) for both PageRank (see Figure 5.4) and WTF (see Supplementary Figure 5.8). As we will see later, this is mainly due to the preferential attachment mechanism [87, 194]. Moreover, we find that on average: (i) Balanced networks ($f_m = 0.5$) can get a fair ranking (green) when both groups possess the same homophily scores ($h_{MM} = h_{mm}$). The same applies for neutral networks ($h_{MM} = h_{mm} = 0.5$) regardless of their fraction of minorities ($f_m \leq 0.5$). (ii) When the fraction of minorities decreases ($f_m < 0.5$), groups can be fairly represented in the rank in two regimes: First, when both groups are homophilic, homophily within minorities must be higher than homophily within majorities ($h_{mm} > h_{MM} > 0.5$). Second, when both groups are heterophilic, homophily within majorities must be higher than homophily within minorities ($h_{mm} < h_{MM} < 0.5$) to balance the importance of groups.

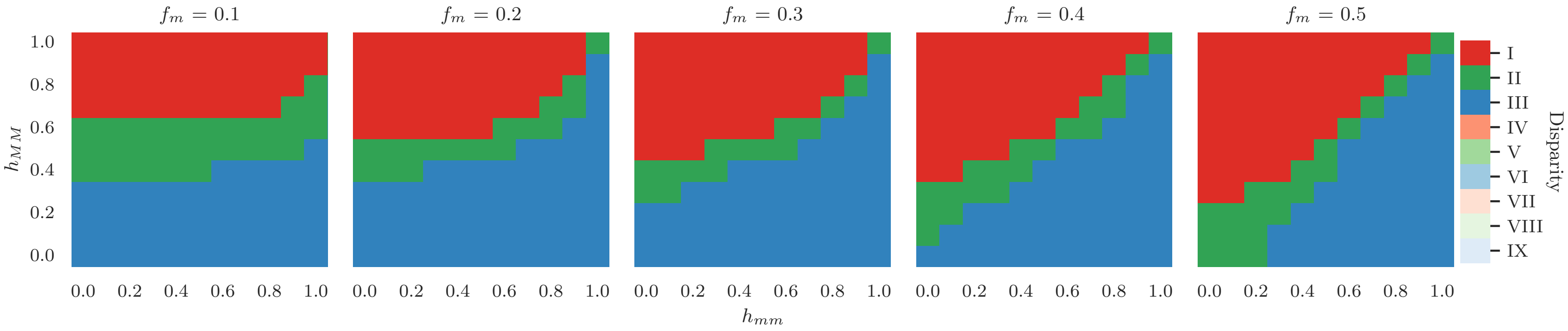


Figure 5.4.: **The effects of homophily and fraction of minorities in the global disparity of PageRank.** Columns represent the fraction of minorities in the network, x-axis indicates the homophily within minorities, and y-axis the homophily within majorities. Colors denote the region where the disparity lies in according to our interpretation in Figure 5.2. First, we see that, on average, there is never low global inequality (i.e., regions IV to IX —lighter colors— do not appear). This makes sense because these are scale-free networks. Second, depending on the level of homophily within groups, minorities on average can be under-represented (region I, red), or over-represented (region III, blue), or well-represented (region II, green). For example, when $f_m = 0.1$, minorities are on average under-represented when $h_{MM} \geq 0.7$ and $h_{MM} \geq h_{mm}$.

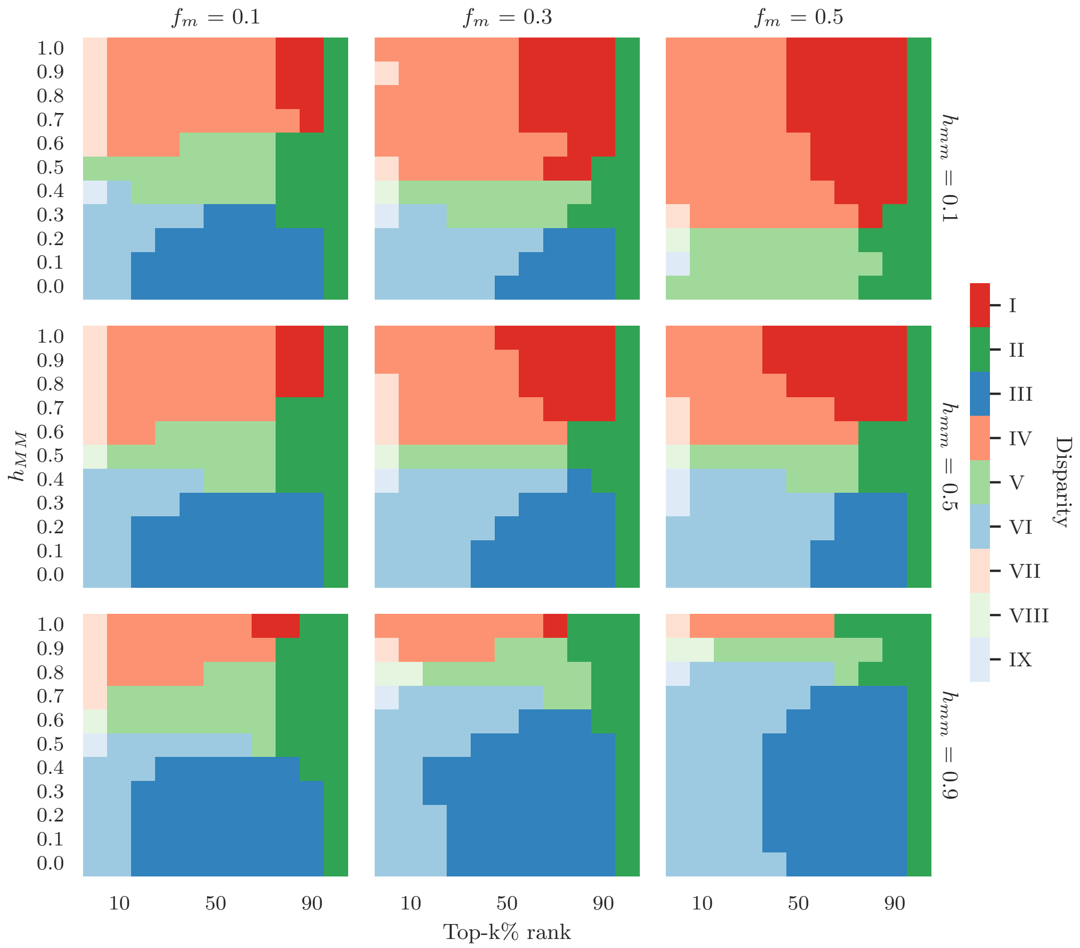


Figure 5.5.: **The effects of homophily and fraction of minorities in the local disparity of PageRank.** Columns represent the fraction of minorities (10%, 30% and 50%) and rows show homophily within minorities (from top to bottom: heterophilic, neutral and homophilic). The x-axis denotes the top-k% rank and the y-axis shows homophily within majorities. Colors refer to the regions of disparity introduced in Figure 5.2. One can see that the minority suffers most (red) when the majority is homophilic and the minority is either heterophilic or neutral. Moreover, inequality is lowest (very light colors) only for a few cases at top-5%. This means that the top best ranked nodes are very similar and their ranks are far from the majority of nodes (i.e., due to preferential attachment). Moreover, inequity remains mostly consistent regardless of top-k%. In other words, if the ranking algorithm favors one group in the top-5% (e.g., red or blue), it will continue to do so until entering the fair regime (green).

Local disparity. We also compute inequality and inequity within each top-k% rank in order to see to what extent they change when k increases. In the case of PageRank, we see in Figure 5.5 that inequality varies (i.e., from light to dark colors) in different regimes mainly due to the size of k (x-axis), and inequity due to the interplay between homophily within groups, h_{MM} and h_{mm}. In particular: (i) Only at the top-5% of the rank we see a few cases of low inequality (regions VII, VIII and IX; very light colors), this means that nodes at the very top possess very similar ranking scores, but they are very far from the rest of the population, i.e., the larger the top-k%, the higher the Gini (darker colors). This holds for WTF up to roughly the top-30% (see Supplementary Figure 5.9). Overall, PageRank converges to high inequality faster than WTF. (ii) Inequity (regions: red, blue, green) is consistent across all top-k% ranks for both algorithms. In other words, if the ranking algorithm favors or harms one group in the top-5%, it will continue to do so until converging to the fair regime (regions II, V, VIII; green). With a few exceptions, this fair regime is only reached when k is very large. For example, if a minority group is under-represented at the top-5%, it will remain under-represented at the top-80% (see $h_{mm} = 0.1$ and $h_{MM} \geq 0.7$ in Figure 5.5 for PageRank, and Supplementary Figure 5.9 for WTF). (iii) Minorities are often over-represented when majorities are heterophilic $h_{MM} < 0.5$; (regions III, VI, IX; blue). In contrast, minorities are often under-represented when majorities are homophilic $h_{MM} > 0.5$ (regions I, IV, VII; red). This is consistent up to $\approx$ top-80% for both algorithms.

In summary, our results suggest that the size of k does not have an influence on inequity. This means that if the algorithm amplifies inequity at the top-5%, it will also amplify inequity at larger top-k%'s. Therefore, increasing the selection pool (larger k) does not improve the representation of minorities. This can be explained by the fact that the preferential attachment mechanism disproportionately affects nodes ranking [97].

Correlation and feature importance. We compute the Spearman correlation between inequality and inequity, and conduct a random forest regression to measure the importance of each network property on both inequality and inequity values (see Supplementary Section 5.6.3 for more details). Results are shown in Table 5.1 for PageRank and Supplementary Table 5.5 for WTF. We find that inequality and inequity are positively correlated in both global and local regimes. In other words, the more skewed the rank distribution (i.e., high Gini), the more unfair with either group (i.e., mean error far from zero), and vice versa. This correlation is stronger and more significant in PageRank than in WTF. In terms of feature importance, we find that global inequality ($Gini$) is mainly explained by both homophily values, whereas global inequity (ME) is mainly driven by homophily within majorities. Local inequality ($Gini_k$), however, is mainly explained by the top-k% rank, and local inequity (ME_k) by the homophily within the majority group. Notice that we added the variable ϵ to verify whether the

Table 5.1.: **10-fold cross-validation for PageRank.** We use a Random Forest Regressor to assess feature importance and report the mean and standard deviation of the out-of-sample R^2. Features are ranked in descending order based on their mean importance (from left to right) and highlighted if their importance represents at least 50% of the total importance. Corr shows the Spearman correlation between inequality and inequity scores (p-values ≈ 0). ϵ represents random chance.

Type	Outcome	Corr	R^2	Feature	Importance
Global	$Gini$	0.41	0.91 (0.009)	$\boldsymbol{h_{MM}}, \boldsymbol{h_{mm}}, f_m, \epsilon$	0.43, 0.31, 0.21, 0.05
	ME		0.99 (0.001)	$\boldsymbol{h_{MM}}, h_{mm}, f_m, \epsilon$	0.61, 0.31, 0.08, 0.0
Local	$Gini_k$	0.21	0.95 (0.002)	$\boldsymbol{k}, h_{MM}, h_{mm}, f_m, \epsilon$	0.73, 0.11, 0.07, 0.06, 0.03
	ME_k		0.99 (0.001)	$\boldsymbol{h_{MM}}, h_{mm}, k, f_m, \epsilon$	0.51, 0.27, 0.14, 0.08, 0.01

network-based features are better than random (see Supplementary Section 5.6.3 for more details). In the case of PageRank, all network-based features perform better than by chance. However, randomness seems to be more relevant for explaining rank inequality ($Gini$) in WTF.

These results are in agreement with what we see in previous figures; even though majority nodes produce most of the inequality and inequity in the rank, their interplay with minority nodes can change or intensify the direction of bias. In fact, both homophily values can explain 75% (49%) of $Gini$, the global inequality in PageRank (WTF), 92% (88%) of ME, the global inequity, and 78% (74%) of ME_k, the local inequity. However, the top-k% rank together with the homophily within majority nodes explain 84% (86%) of $Gini_k$, the local inequality.

5.2.4. How do different social mechanisms of edge formation contribute to disparity?

So far, we show that PageRank and WTF on our network model produce high inequality and a wide-range of possible inequity outcomes. How much of that inequality or inequity was a product of homophily or preferential attachment? To see the effects of these two mechanisms alone, we generate new networks by turning on and off the homophily and preferential attachment features (see Section 5.5 for the details of the models).

Figure 5.6 shows the inequality and inequity produced by PageRank on a variety of models: DPA (Directed Preferential Attachment), DH (Directed Homophily), Random, and DPAH (see Supplementary Figure 5.10 for WTF). Results from both algorithms show that networks whose nodes connect through preferential attachment (DPA) produce on average higher inequality compared to DH and Random. However, when preferential attachment is combined with homophily (DPAH), this inequality increases even further. Additionally, we see that WTF produces higher inequality compared to PageRank (see Supplementary Section 5.6.4 for more details). Inequity, on the other hand, is mainly driven

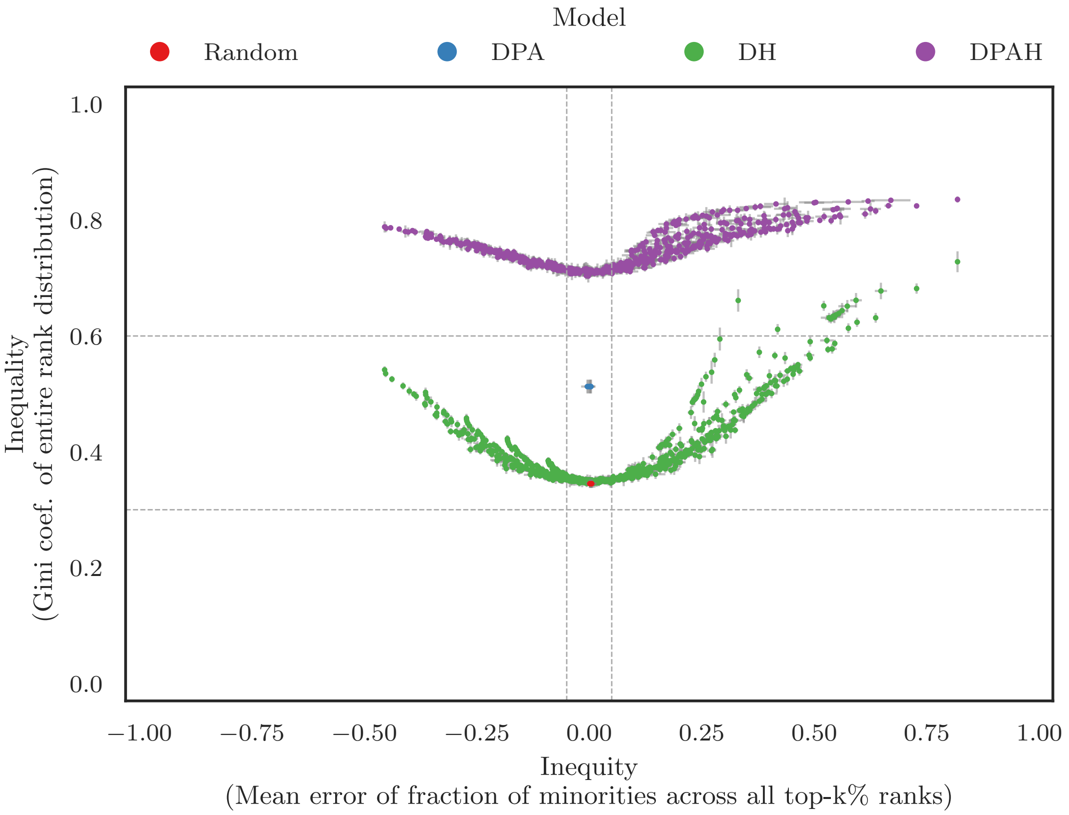


Figure 5.6.: **The effects of homophily and preferential attachment in the global disparity of PageRank.** We generated directed networks using four different models of edge formation. DPA: only preferential attachment. DH: only homophily. DPAH: our proposed model that combines DPA and DH. Random: a baseline where nodes are connected randomly. We see the following patterns: (i) Homophily (DH) produces a moderate-to-high level of inequality ($0.3 < Gini < 0.8$), while preferential attachment (DPA) produces a consistent moderate inequality ($Gini \approx 0.5$). When both mechanisms are combined (DPAH), the rank inequality increases even further ($0.7 < Gini < 0.9$). (ii) Random and preferential attachment (DPA) are always fair ($ME = 0$ or $|ME| \leq \beta$), while in the cases where homophily is involved (DH and DPAH) inequity is often high ($|ME| > \beta$). Thus, in general preferential attachment is the main driver of inequality, while homophily influences both inequality and inequity. Vertical and horizontal error bars represent the standard deviation over 10 runs of the Gini and ME, respectively.

by homophily. This means that, homophily (DPAH and DH) influences both, inequality and inequity in both algorithms.

Note that in Figure 5.6, we fixed the activity of nodes to $\gamma_M = \gamma_m = 3.0$. However, when we set these parameters to $\gamma_M = \gamma_m < 3.0$ (more active nodes or lower values of γ as found in several scale-free networks [4]), inequality decreases, see Supplementary Figure 5.12. This behavior holds even if the minority group is the only one increasing its activity ($\gamma_m = 1.5 < \gamma_M = 3.0$) which in turn increases inequity against the majority, see Supplementary Figure 5.13. Additionally, in Supplementary Figure 5.11, we see that edge density also plays a role in the inequality produced by PageRank and WTF. This means that, by further adjusting these two parameters (node activity and edge density), we would expect changes only to inequality since inequity is mainly affected by homophily as we saw before.

5.2.5. Disparities on empirical networks

First, we fit the DPA, DH, and DPAH models to each of the empirical networks in order to find the mechanism that best explains the inequality and inequity found in the rank. The parameters passed to these models are inferred from the real networks and described in Table 5.3. Second, we rank nodes in the empirical and fitted networks using PageRank and WTF, and compute the disparities (inequality vs. inequity) found in their rank distribution. Results are shown in Figure 5.7 for PageRank and Supplementary Figure 5.16 for WTF. Disparity values from the real-world networks are labeled as *empirical* (black dot), and disparity values from the fitted networks are labeled according to the model (x marks). We see that each network tells a different story. This can be explained by the nature or domain of these networks. For instance, APS and Hate are best explained by the DPA model. This means that scientists tend to cite authors that have already many citations, and users in Twitter tend to retweet content posted by popular users (i.e., popular in terms of the number of retweets they get). Blogs and Wikipedia on the other hand, are best explained by our DPAH model. Notice that both are hyper-link networks. In other words, people tend to add not only popular references to their Web pages, but also related to their topics (i.e., political leaning in Blogs, and gender in Wikipedia). Note that the Hate network shows the lowest (empirical) inequality. This is due to the fact that it possesses low out-degree exponents ($\gamma_M = 2.2$, $\gamma_m = 1.7$).

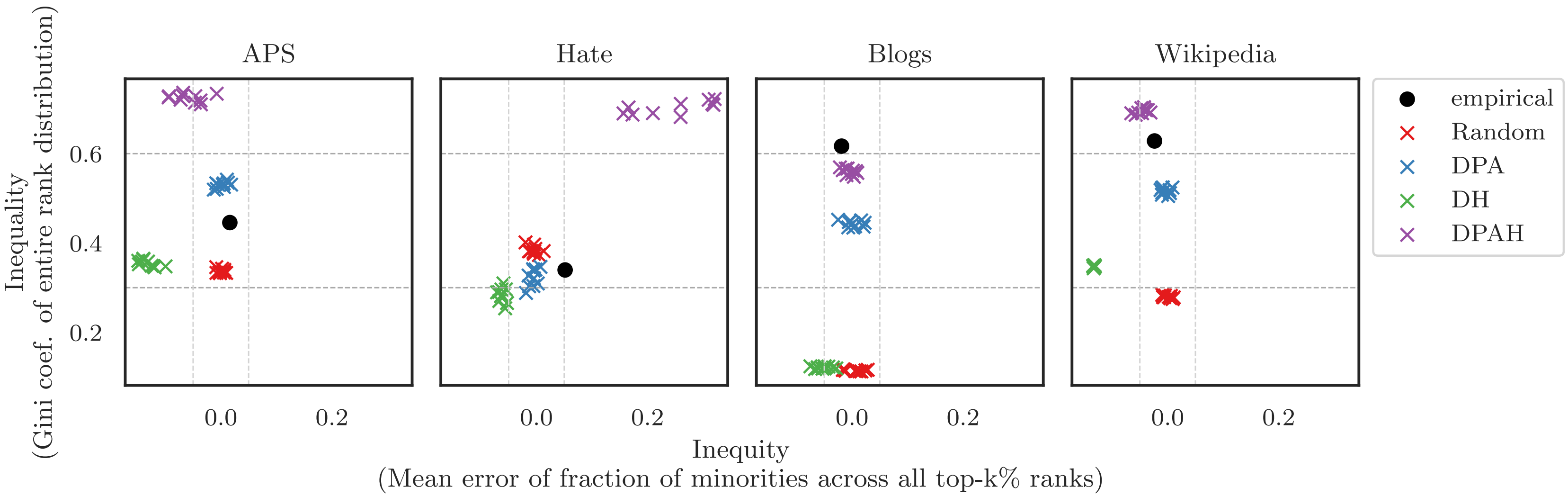


Figure 5.7.: **Global disparity in PageRank on empirical networks.** Each column represents an empirical network. Citation/retweet networks (APS and Hate) and Hyper-link networks (Blogs and Wikipedia). Inequality and inequity are shown in the y- and x-axis, respectively. The disparity in ranking that we see in empirical networks are best explained as follows: (i) citation/retweet networks by preferential attachment PA, and (ii) hyper-link networks by preferential attachment and homophily DPAH.

5.2.6. Strategies towards a fair ranking

Results from both algorithms show that while the homophily within majorities is the main driver for inequality and inequity, minorities may overcome unfair rankings by connecting strategically in the network. For instance, when both groups are equally active, minorities should adjust their homophily based on the homophily of the majority. (i) When majorities are homophilic $h_{MM} > 0.5$, minorities should increase their homophily such that $h_{mm} > h_{MM}$. (ii) When majorities are (somewhat) neutral ($h_{MM} = 0.5 \pm 0.1$), minorities may connect arbitrarily with any group without being too homophilic, otherwise they will become over-represented in the rank. (iii) When majorities are heterophilic $h_{MM} < 0.5$, one solution to achieve a fair rank is to increase the size of the minority group, and make sure that both groups behave similarly in terms of homophily ($h_{MM} \approx h_{mm}$). Otherwise, minorities will be over-represented regardless of their in-class homophily. On the other hand, when one group is more active than the other, achieving a fair rank becomes challenging. Nevertheless, if the objective is to increase the visibility of minorities in the rank, then the minorities themselves should be more active in the network by creating more connections to increase their out-degree. Note that these "strategies" without algorithmic intervention may work in scenarios such as a citation or collaboration networks, but they might not work in other scenarios. In such cases, we need additional recommender systems to help under-represented groups discover those "strategic" links that will help them climb to higher ranks.

5.3. Discussion and Future Work

In this work we have proposed a systematic study to measure the inequality and inequity produced by PageRank and Who-To-Follow (WTF). Our approach disentangles the effect of network structure on the rank distributions of these two algorithms by using synthetic networks. By doing so, we control for the properties of the network and measure how these changes affect the rankings. In particular, we studied six prominent structural properties of social networks: homophily, preferential attachment, fraction of minorities, edge density, node activity and the directionality of links. We found that the systemic bias produced by these algorithms in the rank is mainly due to *homophily imbalance* ($h_{MM} \gg h_{mm}$ or $h_{mm} \gg h_{MM}$) for inequity, and the interplay between our six properties of interest for inequality. Consequently, our systematic study makes PageRank and Who-To-Follow interpretable and explainable since our results show the necessary structural conditions to achieve a fair rank. A potential avenue to reduce inequity is to then create synthetic connections before the ranking as it is done for correcting the class imbalance problem in supervised learning [44]. Alternatively, these conditions or strategic connections may be added into the network to change its structure as a collective fairness intervention. For instance, recommender systems could suggest relevant articles not only based on popularity and (keyword) similarity but also based on fairness by fulfilling diversity constraints.

Notice that our model simplifies the role of homophily and minorities. First, it assumes that all nodes of the same group have the same in-class and between-class homophily. This means that rich mixing patterns might get ignored since some nodes can exhibit local differences [200]. Second, while there exist multiple definitions of minorities [236] we adopted the one by Italian jurist Francesco Capotorti: "*a group numerically inferior to the rest of the population of a State, in a non-dominant position...*" [39, 110]. However, we constrained this population to all nodes in a given network (neither the State nor the world population). Notice as well that the ranking amplifies the representation of minorities by reducing the representation of majorities in top ranks (e.g., when the majority is heterophilic, see Figure 5.1). This aligns with the definition of minorities by Wirth [265] that implies that "*minorities objectively occupy disadvantageous positions in society*". This means that "*a minority may actually, from a numerical standpoint, be the majority*" (e.g., people living in poverty in under-developed countries). In other words, under these two definitions, being in disadvantage in the top-k% (inequity) is not a group size issue only, but a combination of group size and homophily as we have previously shown. More complex definitions of minorities are out of the scope of this paper. A further limitation is that we focus on a single binary attribute (e.g., color $\in \{\text{Black}, \text{White}\}$), this means that multiple sources of inequality and inequity (e.g., intersections of disadvantage such as being poor and of color) cannot be captured at once [125]. Addressing these issues is beyond the scope of this paper and we leave them for future work.

Finally, we disentangled the individual effects of preferential attachment and homophily in the rank by comparing the disparities of our proposed DPAH model with two variants and a baseline: networks with preferential attachment only (DPA), networks with homophily only (DH), and directed Erdös-Rényi (Random) [71] graphs. Further research can investigate other topologies and social mechanisms of edge formation such as clustering [56], transitivity [27], and reciprocity [65]. Similarly, other structural properties such as monophily [6] and second order homophily [82] can be studied to measure their influence on ranking.

5.4. Conclusions

In this work we have investigated under which conditions PageRank and Who-To-Follow (WTF) *reduce*, *replicate* or *amplify* the representation of minorities in top ranks. In particular, given the rank distribution produced by these algorithms, we computed *inequality* as the dispersion among individuals in terms of ranking scores, and *inequity* as whether minorities are over-, under- or well-represented in top ranks compared to their representation in the network. We studied these two metrics separately and in combination to better understand the mechanisms that can explain them.

To that end, we proposed DPAH, a growth network model that allows to generate realistic scale-free directed networks with different levels of homophily, fraction of minorities, node activity, and edge density. In these networks, we found that both inequality and inequity are positively correlated and mainly driven by the homophily within majorities. This means that, when the majority group is highly homophilic, the minority group is under-represented in top ranks. Also, when the majority is highly heterophilic, the minority benefits tremendously since it is over-represented in the top-k%. However, minorities can overcome these disparities by connecting strategically with others. Thus, equity in ranking is a trade-off between homophily and the fraction of minorities.

Our systematic study makes PageRank and Who-to-Follow explainable and interpretable to help data scientists understand and estimate the disparity that these algorithms produce given the structure of networks, which is key for proposing targeted interventions. We hope our results create awareness among majority and minority groups about these disparities since they may replicate and even amplify the biases found in social networks.

5.5. Data and Methods

5.5.1. Synthetic networks

Network models have been proposed with various social mechanisms. For instance, the classic *stochastic-block model* [116] which allows for homophily between and across groups, and the *configuration model* [189] which generates links among nodes by preserving a given degree distribution. On the other hand, the *preferential-attachment* model [18] produces scale-free networks due to cumulative advantage [176]. Although these models can reproduce certain properties of real-world networks such as degree or homophily, they fail at guaranteeing similar visibility of minorities as their empirical counterpart. In this direction, Karimi et al.[133] and Fabbri et al. [83] devise social network models with preferential attachment, adjustable homophily and fraction of minorities. They demonstrate how the degree rank of the minority group in a network is a function of the relative group sizes and the presence or absence of homophily. However, the former models undirected networks, and the latter did not control for edge density and node activity (i.e., power-law out-degree distributions) as we do in this work for minority and majority groups.

Directed network

We define a directed network as: Let $G = (V, E, C)$ be a node-attributed unweighted graph with $V = \{v_1, ..., v_n\}$ being a set of n nodes, $E \subseteq V \times V$ a set of e directed edges, and $C = \{c_1, ..., c_n\}$ a list of binary class labels where each element

c_i represents the class membership of node v_i. The fraction of minorities f_m captures the relative size of the minority class—with respect to C—in the network. We refer to the minority group as m, and to the majority group as M. A network is *balanced* when all class labels have the same number of nodes ($f_m = 0.5$), otherwise it is *unbalanced* ($f_m < 0.5$). Networks fulfill a predefined edge density level d. Since n and d are given, networks stop growing when $e = dn(n-1)$.

In order to generate directed links, inspired by the activity-driven network model [201], we assign an activity score to each node that determines with what probability the existing node becomes active and creates additional links to other nodes. It has been shown that in empirical networks the activity of the nodes follows a power-law distribution [201]. Therefore, we assign an activity to each node drawn from a power-law distribution $\rho(\gamma) = X^{-\gamma}$. Note that each group possesses its own activity distribution and they are defined by the power-law exponent γ_M and γ_m for majority and minority nodes, respectively. The level of activity of a group is inversely proportional to γ. That is, groups with higher out-degree produce lower γ (more skewed).

Then, the probability of connecting a source (active) node v_i to a target node v_j (or in other words the probability of connecting to v_j given the source node v_i) is explained by any of the following three mechanisms of edge formation.

Preferential attachment (DPA)

Also known as the *rich-get-richer* effect or *cumulative advantage* in social networks [18, 176]. It indicates that nodes tend to connect to popular nodes. We define popularity as the in-degree of the node. Therefore, the probability that a source node v_i connects to a target node v_j is proportional to the *in-degree* of the target node v_j.

$$P(i \to j) = P(j|i) = \frac{k_j^{in}}{\sum_{l=1}^{N} k_l^{in}} \tag{5.1}$$

Homophily (DH)

It is the tendency of individuals to connect (or interact) with similar others [171, 189]. Thus, the probability that a source node v_i connects to a target node v_j is driven by the homophily between their classes c_i and c_j. We assign a homophily value to each dyad based on pre-defined homophily parameters within majorities and minorities, h_{MM} and h_{mm}, respectively. Homophily values range from 0.0 to 1.0. If the homophily value is high, that means that nodes of the same class are attracted to each other more often than nodes of different attributes. Following the definitions from previous work [83, 133, 214], nodes of the same class with homophily $h_{aa} = 0.5$ are referred to as *neutral* (i.e., they connect randomly to either class), otherwise they are *heterophilic* if $h_{aa} < 0.5$ (i.e., more likely to connect to the other class), or *homophilic* when $h_{aa} > 0.5$ (i.e., more likely to connect to the same class). Note that in- and between-class homophily values are complementary: $h_{mm} = 1 - h_{mM}$ and $h_{MM} = 1 - h_{Mm}$.

$$P(i \to j) = P(j|i) = \frac{h_{ij}}{\sum_{l=1}^{N} h_{il}} \tag{5.2}$$

Preferential attachment with homophily (DPAH)

We propose DPAH, a directed growth network model with adjustable homophily and fraction of minorities. DPAH stands for **D**irected network with **P**referential **A**ttachment and **H**omophily. This mechanism combines DPA and DH, and is an extension of the BA-Homophily model [133].

$$P(i \to j) = P(j|i) = \frac{h_{ij} k_j^{in}}{\sum_{l=1}^{N} h_{il} k_l^{in}} \tag{5.3}$$

Note that DPA and DH are especial cases of DPAH where only the in-degree mechanism varies. This means that, the out-degree distribution remains the same as in DPAH: it is driven by the activity model. Additionally, we include a random model where both source and target nodes are chosen at random (i.e., directed Erdös-Rényi model [71]). Table 5.2 shows the parameters adjusted in each model. Number of nodes n and edge density d are arbitrary in the sense that they are not part of the edge formation mechanism. Thus, we fix them to make a fair comparison across all models.

Table 5.2.: **Model parameters.** Check marks denote that a given model (column) requires a particular parameter (row): number of nodes n, fraction of minorities f_m, edge density d, in-class homophily h_{aa}, and the power-law exponent of the activity distribution γ. Sub-indices M and m refer to the majority and minority groups, respectively. The difference between DH and DPAH is the preferential attachment (in-degree) mechanism. All models produce directed networks.

	Random	**DPA**	**DH**	**DPAH**
n	✓	✓	✓	✓
f_m	✓	✓	✓	✓
d	✓	✓	✓	✓
h_{MM}	-	-	✓	✓
h_{mm}	-	-	✓	✓
γ_M	-	✓	✓	✓
γ_m	-	✓	✓	✓

5.5.2. Empirical networks

We inspect four networks from different domains and compute the inequalities and inequities produced by PageRank and WTF. Table 5.3 shows the most important properties of these networks.

- **APS:** The American Physical Society citation network whose nodes represent articles, and edges represent citations. The binary class of each node is pacs and encodes two different Physics sub-fields where 05.20.-y (Classical statistical mechanics) is the minority.

- **Hate:** A retweet network [209] where nodes denote users, and edges represent retweets among them. Users are labeled as either hateful or normal depending on the sentiment of their tweets. Hateful users represent the minority.

- **Blogs:** An hyper-link network from political blog posts about the 2004 U.S. election [3]. Nodes represent blog pages, and edges hyper-links among them. Each blog is labeled as either right- or left-leaning. The latter represents the minorities.

- **Wikipedia:** A Wikipedia hyper-link network where nodes represent U.S. politicians [95, 256] labeled as either male or female. Female politicians represent the minorities.

5.5.3. Ranking and recommendation algorithms

There exist a variety of ranking and recommendation algorithms that follow different strategies depending on the nature of the problem. For instance, in information systems, items such as content, Web pages, and products are ranked to recommend users what to read or buy [163]. In social networks, however, people are ranked to identify their hierarchy or importance [59, 101, 228], and recommended to other users in order to establish new connections [19, 181, 270, 275]. These rankings and recommendations are based on algorithms that often rely on whom we are already connected with. In this work, we focus on two such algorithms widely used in practice [99]: PageRank [193] and Who-to-Follow (WTF) [106]. While PageRank determines the global ranking of nodes in comparison with all other nodes, WTF deals with ranking nodes in a node level and thus remains a local measure. For that reason, we focus on these two algorithms to capture both dimensions.

PageRank

It was invented to rank all web pages in the Web [193], and has been used in several applications [99]. For example, to study citation and co-authorship networks

Table 5.3.: **Empirical networks.** APS, a scientific citation network. Hate, a retweet network. Blogs, a political blog hyper-link network. Wikipedia, a hyper-link network of politicians. Each row represents a property of the network. E_{**} represents the fraction of edges within and across groups, and h_{**} homophily values inferred by the DPAH model (see Supplementary Section 5.6 for derivations).

dataset	APS	Hate	Blogs	Wikipedia
n	1853	4971	1224	3159
class	pacs	hate	leaning	gender
M	05.30.-d	normal	right	male
m	05.20.-y	hateful	left	female
f_m	0.37561	0.10943	0.48039	0.15226
d	0.00106	0.00061	0.01271	0.00149
γ_M	3.22246	2.23026	4.88733	4.22425
γ_m	8.93993	1.73445	3.22464	6.16567
E_{MM}	0.64981	0.56898	0.47070	0.78469
E_{Mm}	0.02859	0.10244	0.04741	0.07824
E_{mM}	0.02721	0.07886	0.04105	0.10685
E_{mm}	0.29439	0.24972	0.44084	0.03022
h_{MM}	0.94000	0.58000	0.92000	0.59000
h_{mm}	0.96000	0.95000	0.90000	0.62000

[85, 130, 162]. PageRank assigns an importance score to every single node in a network. This score takes into account the number and quality of incoming links of each node. The PageRank of node i is defined as follows:

$$PR(i) = (1 - \alpha) + \alpha \sum_{j \in N_i} \frac{PR(j)}{k_j^{out}} \tag{5.4}$$

where $i \in V$, N_i represents all neighbors of node v_i (e.g., all nodes v_i points to), and k_j^{out} the out-degree of node v_j. The damping factor α, or probability of following links using a Random Walker, is set to 0.85 as suggested by Brin and Page [33]. We use the `fast-pagerank` [217] python package to compute the PageRank score of all nodes using sparse adjacency matrices.

Who-To-Follow (WTF)

This recommendation algorithm was created and used by Twitter to suggest new people to follow [106]. It is based on SALSA [154] which in turn is based on Personalized PageRank [128]. In a nutshell, for each user u (or node $v_i \in V$), the algorithm looks for its *circle of trust*, which is the result of an egocentric random walk (similar to personalized PageRank) [106]. Then, based on this circle-of-trust, the algorithm ranks all users that are not yet friends with u but are connected through the circle of trust. Then, we take the top-k of these (recommended) users, and add up the counter of being selected as a recommendation to each of them. This is done for every node u in the network. At the end, the rank of each node encodes the *number of times a user was suggested as a recommendation* across all nodes in the network. Thus, the WTF score for each node is defined as follows:

$$WTF(i) = \sum_{j \in V} \mathbb{1}_{SALSA(j)}(i) \tag{5.5}$$

where $SALSA(j)$ refers to the top-k users the SALSA algorithm recommends to node j. In this work we select the top-10 users as recommendations. $\mathbb{1}_A(x)$ denotes the indicator function or boolean predicate function to test set inclusion (i.e., whether $x \in A$).

5.5.4. Gini coefficient

The Gini coefficient was developed by the Italian Statistician Corrado Gini [98] to measure the income inequality of a society. It is defined as the mean of absolute differences between all pairs of individuals for some measure. In our setup this measure is the score given to every node by PageRank and Who-To-Follow. The minimum value is 0 when all individuals' scores are equal, and its maximum value is 1 when there is a big gap or discrepancy between scores [42].

We define the Gini coefficient of the rank distribution X as follows. For more details see [242]:

$$Gini(X) = \frac{\sum_{i=1}^{\hat{n}}(2i - \hat{n} - 1)x_i}{n \sum_{i=1}^{\hat{n}} x_i} \tag{5.6}$$

where $x \in X$ is an observed value in the rank distribution, $\hat{n} = |X|$ is the number of values observed, and i is the rank of values in ascending order.

5.6. Appendix

5.6.1. Derivation of the probability of having an internal link

Let $K_a^{in}(t)$ and $K_a^{out}(t)$ be the sum of the in- and out-degrees of nodes from group a at time t. The overall growth of the network follows a DPAH process. Thus, the evolution of in-degree and out-degree follows:

$$\begin{cases} K_a^{in}(t) + K_b^{in}(t) = K^{in}(t) = mt \\ K_a^{out}(t) + K_b^{out}(t) = K^{out}(t) = mt \end{cases} \tag{5.7}$$

where m is the number of new links in the network at each time step t. In each time step, a node v_i is chosen. That results in m new out-going links from v_i. We set $m = 1$. Thus, in each time step only one edge is created from v_i to v_j.

Let us denote the relative fraction of group size for each group as f_a and f_b and their respective activity parameters γ_a and γ_b that represent the exponents of the activity distribution. Thus, the behavior of the network is similar to what we have shown before [133]; only the total number of links is different. Let us also define $\hat{\gamma_a}$ as an average value drawn from the activity distribution of group a, $\rho(\gamma_a) = X^{-\gamma_a}$ using mean field approximation. Similarly, $\hat{\gamma_b}$ as an average value drawn from the activity distribution of group b, $\rho(\gamma_b) = X^{-\gamma_b}$. We can show that in the limit of $\Delta t \to 0$, for each group, the in-degree growth function follows:

$$\frac{dK_a^{in}}{dt} = m\left(f_a \times \hat{\gamma_a}\left(1 + \frac{h_{aa}K_a^{in}(t)}{h_{aa}K_a^{in}(t) + h_{ab}K_b^{in}(t)}\right) + f_b\hat{\gamma_b}\left(\frac{h_{ba}K_a^{in}(t)}{h_{bb}K_b^{in}(t) + h_{ba}K_a^{in}(t)}\right)\right) \tag{5.8}$$

$$\frac{dK_b^{in}}{dt} = m\left(f_b \times \hat{\gamma_b}\left(1 + \frac{h_{bb}K_b^{in}(t)}{h_{bb}K_b^{in}(t) + h_{ba}K_a^{in}(t)}\right) + f_a\hat{\gamma_a}\left(\frac{h_{ab}K_b^{in}(t)}{h_{aa}K_a^{in}(t) + h_{ab}K_b^{in}(t)}\right)\right) \tag{5.9}$$

Next, we focus on the case of links within group a. The same analysis applies for group b.

Let p_{aa} be the probability to establish a link between two nodes of group a. The probability for an incoming or existing node from group a to link to a node of the same group is given by:

$$p_{aa}(t) = f_a \frac{h_{aa}K_a^{in}(t)}{h_{aa}K_a^{in}(t) + h_{ab}K_b^{in}(t)} \tag{5.10}$$

In the simple network growth model, the total degree of the groups increases linearly over time.

$$\begin{cases} K_a^{in}(t) = Cm\hat{\gamma_a}\hat{\gamma_b}t \\ K_b^{in}(t) = (2 - C)m\hat{\gamma_a}\hat{\gamma_b}t \\ K_a^{out}(t) = m\hat{\gamma_a}t \\ K_b^{out}(t) = m\hat{\gamma_b}t \end{cases} \tag{5.11}$$

Denoting C as the in-degree growth factor of the minority group.

5.6.2. Calculating homophily from empirical network

We can calculate homophily in empirical networks using the information about in-group links. First, the total number of edges in a directed network follows:

$$e = e_{aa} + e_{ab} + e_{ba} + e_{bb} \tag{5.12}$$

To calculate e_{aa}, the number of links within class a, we can simply argue that it depends on p_{aa}, the probability of connecting two nodes belonging to class a, multiplied by the probability of the arrival or source node to be of class a, denoted by f_a, the fraction of nodes in class a, as shown in Equation (5.10).

Our network model grows linearly in time. That means, the in-degree growth for each group is linear. Let us assume that the in-degree growth rate of group a is denoted by C_a:

$$K_a^{in}(t) = C_a K^{in}(t) \tag{5.13}$$

Since the in-degree growth remains constant over time, we can calculate C_a in the empirical network by summing all in-degrees of the group

$$C_a(empirical) = \frac{K_a^{in}}{K^{in}} \tag{5.14}$$

Equation (5.10) can be rewritten as

$$p_{aa} = f_a \frac{h_{aa} C_a}{h_{aa} C_a + h_{ab}(1 - C_a)} \tag{5.15}$$

In empirical networks, p_{aa} represents the probability of a directed edge from class a to class a. This probability is proportional to the number of edges from a to a, normalized by the total number of directed edges that start from a:

$$p_{aa} = \frac{e_{aa}}{e_{aa} + e_{ab}} \tag{5.16}$$

We can then calculate Equation (5.16) in the empirical network. Finally we use maximum-likelihood estimate to find the best values for h_{aa} and h_{bb} in Equation (5.15).

Note that the in-degree growth rate C has an sub-linear relationship to the exponent of the in-degree distribution σ and the exponent of the in-degree growth θ [133]. Thus, another method to retrieve empirical homophily is to first estimate the exponents of the in-degree distributions for minority and majority groups (σ_a and σ_b) and plug that into the equation.

$$p_{aa} = \frac{f_a^2 h_{aa}(1 - \theta_b)}{f_a h_{aa}(1 - \theta_b) + f_b h_{ab}(1 - \theta_a)} \tag{5.17}$$

$$p_{bb} = \frac{f_b^2 h_{bb}(1 - \theta_a)}{f_b h_{bb}(1 - \theta_a) + f_a h_{ba}(1 - \theta_b)} \tag{5.18}$$

where $\sigma_a = -(\frac{1}{\theta_a} + 1)$ and $\sigma_b = -(\frac{1}{\theta_b} + 1)$.

5.6.3. Regression model

We build a `RandomForestRegressor` [198, 225] model to explain rank inequality and inequity given the structure of networks. Features (or independent variables) are transformed by scaling them between zero and one. During training, the model uses $n\_estimators = 100$ and all default values from the `Python` package. We use R^2 scores to evaluate the performance of the 10-fold cross-validated model on the test set. As shown in Table 5.4, the global model takes into account the overall behavior or trend regardless of top-k ranks, while the local model includes the top-k ranks. We add ϵ as a dummy variable, with randomly generated values between 0 and 1, to compare the importance of each feature to random chance.

We report the importance of features given by the `feature_importances_` property of the `RandomForestRegressor` model. The higher the value, the more important the feature. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance [225].

5.6.4. Who-To-Follow produces higher inequality compared to PageRank

In the main manuscript we see that Who-To-Follow (WTF) produces more skewed rank distributions compared to PageRank. To understand this behavior, we need to first understand how the algorithms work. PageRank scores reflect the *global* importance of nodes in the network, and this global importance is mostly determined by in-degree [88] and the age of nodes [167]. On the other hand, the WTF score of a node is the number of times the node appears in the top-10 recommendation across all nodes in the network. This top-10 is determined by the circle-of-trust of each node, similar to a Personalized PageRank. This means that this top-10 contains the most visited nodes by a random walker that always starts at the node who is getting the recommendation. Thus, that (local) top-10 will be highly influenced by in-degree too. However, since the WTF score counts the number of times a node appears as a recommendation, it is likely that the highest WTF scores refer to high degree nodes due to preferential attachment. Therefore, the high inequality produced by WTF can be explained by the fact that WTF combines a local random walk with a global count.

5.7. Additional tables and figures

Table 5.4.: **Regression models.** Dependent and independent variables of the four models of interest: Global/Local inequality (Gini) and inequity (ME). We add the dummy variable ϵ (with randomly generated values between 0 and 1) to verify whether the network-based features are better than random or not.

Type	**Dependent variable (Y)**	**Independent variable (Xs)**
Global	$Gini$ ME	f_m, h_{MM}, h_{mm}, ϵ
Local	$Gini_k$ ME_k	f_m, h_{MM}, h_{mm}, k, ϵ

Table 5.5.: **10-fold cross-validation for WTF.** We use a Random Forest Regressor to assess feature importance and report the mean and standard deviation of the out-of-sample R^2. Features are ranked in descending order based on their mean importance (from left to right) and highlighted if their importance represents at least 50% of the total importance. Features with a mark (*) are less important than random ϵ. Corr shows the disparity as the Spearman correlation between inequality and inequity scores (p-values ≈ 0).

Type	**Outcome**	**Corr.**	R^2	**Feature**	**Importance**
Global	$Gini$	0.29	0.35 (0.03)	ϵ, h^*_{MM}, h^*_{mm}, f^*_m	0.37, 0.27, 0.22, 0.14
	ME		0.92 (0.01)	$\boldsymbol{h_{MM}}, h_{mm}, f_m, \epsilon$	0.51, 0.37, 0.07, 0.05
Local	$Gini_k$	0.06	0.86 (0.00)	$\boldsymbol{k}, \epsilon, h^*_{MM}, h^*_{mm}, f^*_m$	0.86, 0.08, 0.02, 0.02, 0.01
	ME_k		0.85 (0.01)	$\boldsymbol{h_{MM}}, \boldsymbol{h_{mm}}, k, \epsilon, f^*_m$	0.43, 0.31, 0.11, 0.08, 0.07

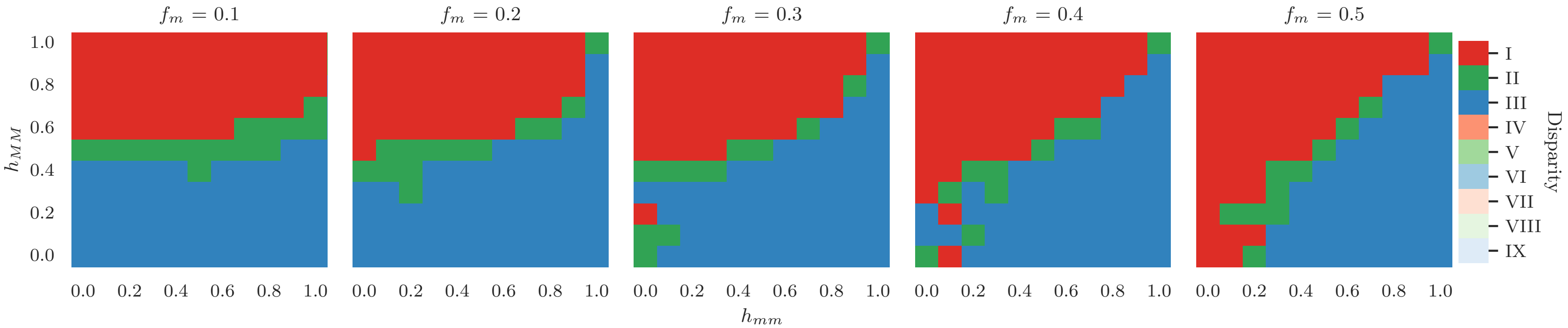


Figure 5.8.: **The effects of homophily and fraction of minorities in the global disparity of WTF.** Columns represent the fraction of minorities in the network, x-axis indicates the homophily within minorities, and y-axis the homophily within majorities. Colors denote the region where the disparity lies according to our interpretation (see Figure 5.2). As in the case of PageRank (cf. Figure 5.4), we see that, on average, there is never low global inequality. Also, depending on the level of homophily within groups, minorities on average can be under-represented (region I, red), or over-represented (region III, blue). Note that the fair case (region II, green) rarely occurs.

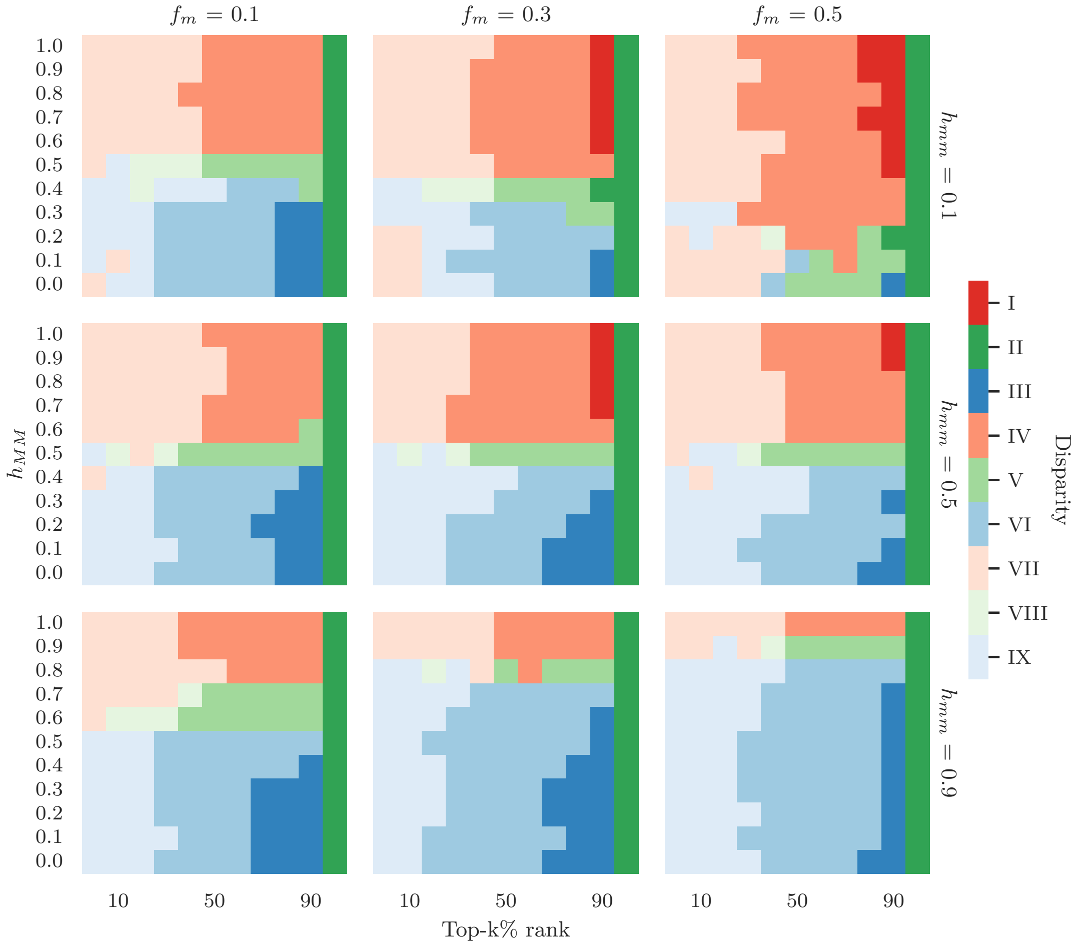


Figure 5.9.: **The effects of homophily and fraction of minorities in the local disparity of WTF.** Columns represent the fraction of minorities (10%, 30% and 50%) and rows show homophily within minorities (from top to bottom: heterophilic, neutral and homophilic). The x-axis denotes the top-k% rank and the y-axis shows homophily within majorities. Colors refer to the regions of disparity (see Figure 5.2). As in the case of PageRank (cf. Figure 5.5), we see that the minority suffers most (red) when the majority is homophilic and the minority is either heterophilic or neutral. Also, inequity remains mostly consistent regardless of top-k%. In contrast to PageRank (up to top-5%), WTF manages to capture nodes with very similar ranking scores (roughly) up to the top-30% (i.e., Gini is low, regions VII, VIII, IX).

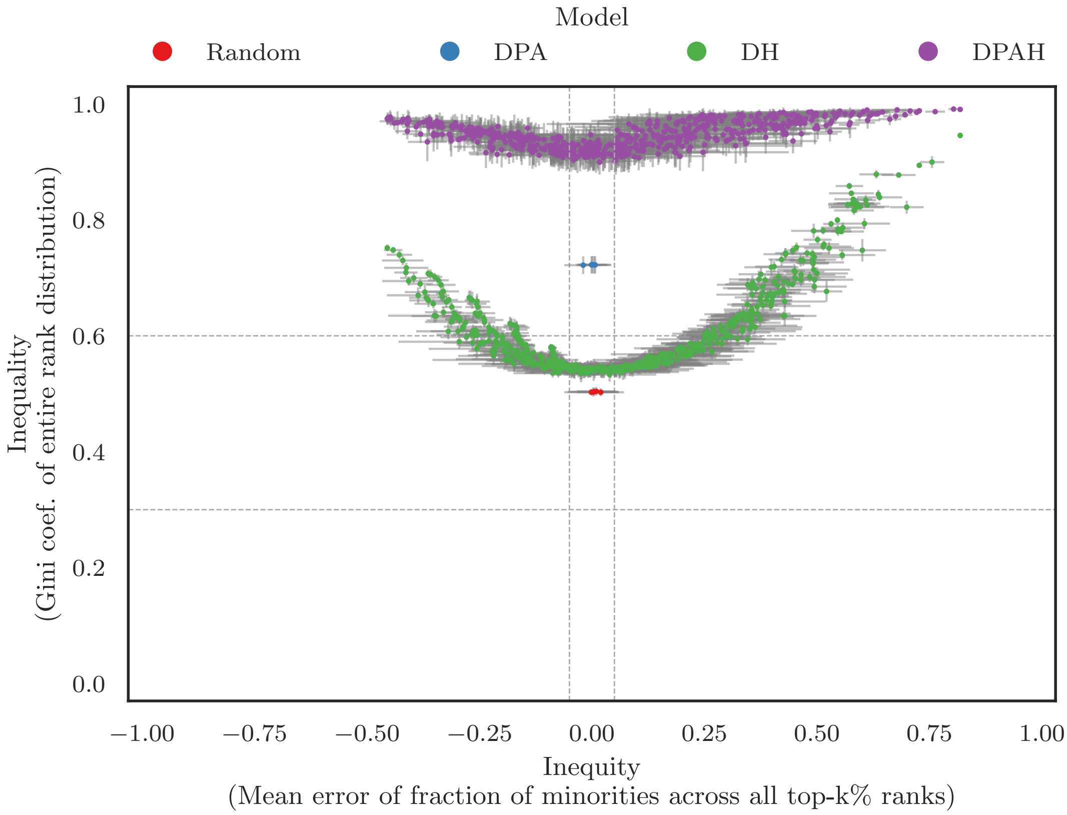


Figure 5.10.: **The effects of homophily and preferential attachment in the global disparity of WTF.** We generated directed networks using four different models of edge formation. DPA: only preferential attachment. DH: only homophily. DPAH: our proposed model that combines DPA and DH. Random: a baseline where nodes are connected randomly. Compared to PageRank (cf. Figure 5.6), all models generate higher inequality (y-axis), whereas inequity remains similar. Vertical and horizontal error bars represent the standard deviation over 10 runs of the Gini and ME, respectively.

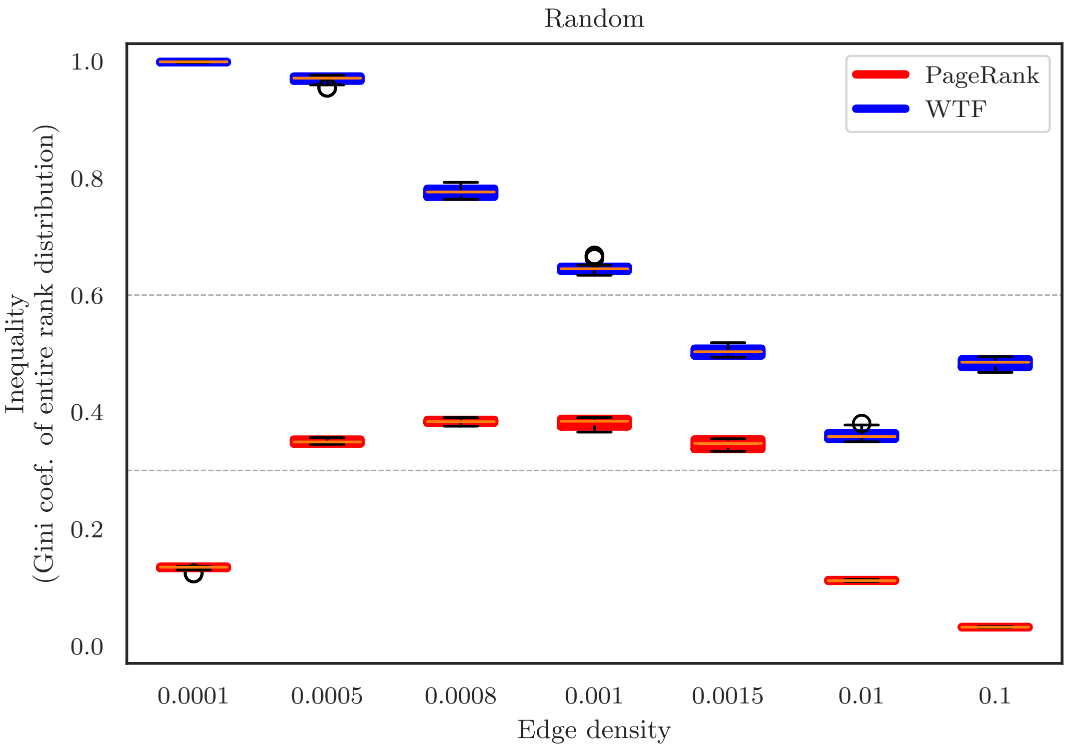


Figure 5.11.: **Global inequality on Random networks as a function of edge density.** We generate directed Erdős-Rényi networks to demonstrate how the global inequality (y-axis) varies with respect to the edge density (x-axis) of the network. For each density value we generate networks with different fractions of minorities and 10 epochs. Note that $d = 0.0015$ corresponds to the Random networks used in the main experiments. Inequality computed on the PageRank distribution is shown in red, while the inequality on WTF is shown in blue. We see different trends for each algorithm. First, the inequality (Gini coefficient) of PageRank is very low when the edge density is extreme (i.e., either too low or too high). This means that in these regimes most nodes are similarly important regardless of the magnitude of their degrees. Second, the inequality of WTF is in general negatively correlated with density (i.e., the lower the density, the higher the inequality [104]). However, in the extreme case of denser networks (i.e., $d = 0.1$), inequality raises. Recall that ranking *inequity* is very close to zero ($ME \approx 0$) in random networks. Further studies are required to analytically understand the limits of inequality with respect to density.

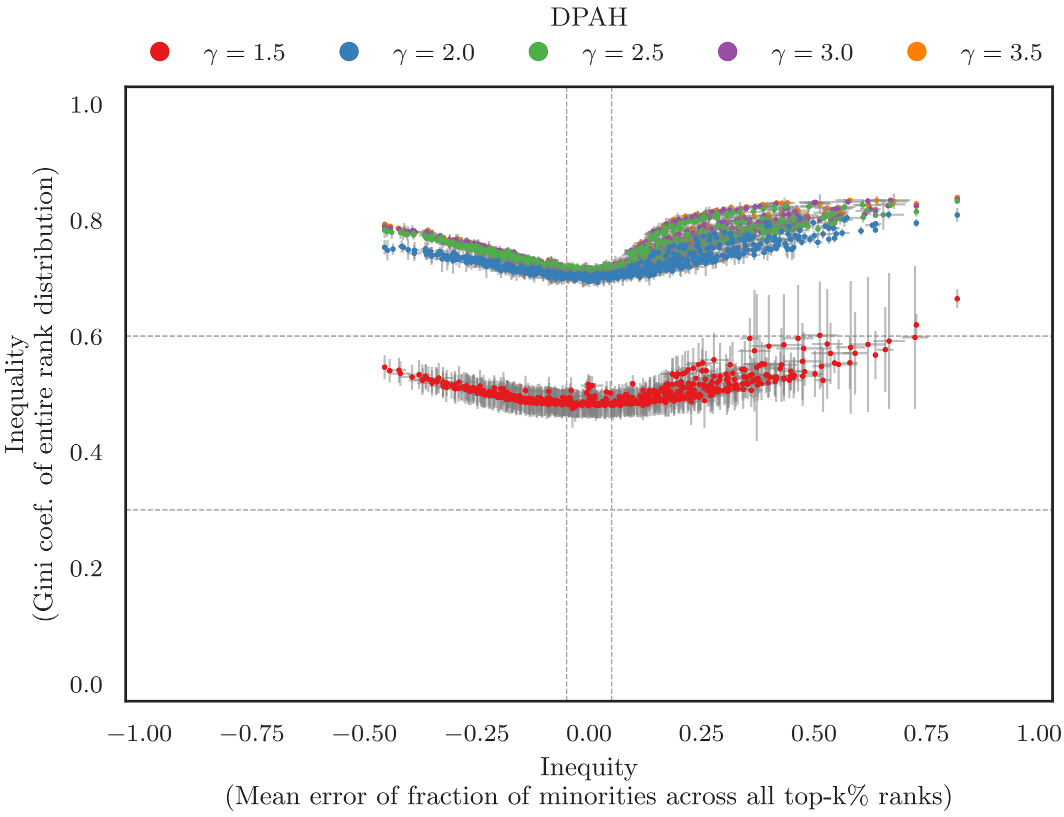


Figure 5.12.: **The effects of symmetric node activity in the global disparity of PageRank.** We generate DPAH networks by varying f_m, h_{MM}, h_{mm}, and fixing number of nodes $n = 2000$ and edge density $d = 0.0015$. Each color represents the activity of nodes as the out-degree exponents of the networks $\gamma = \gamma_M = \gamma_m \in \{1.5, 2.0, 2.5, 3.0, 3.5\}$. We see that by reducing the out-degree exponent in the DPAH networks (from $\gamma = 3.5$ to $\gamma = 1.5$), we reduce inequality (vertical axis), and inequity remains stable. Vertical and horizontal error bars represent the standard deviation over 10 runs of the Gini and ME, respectively.

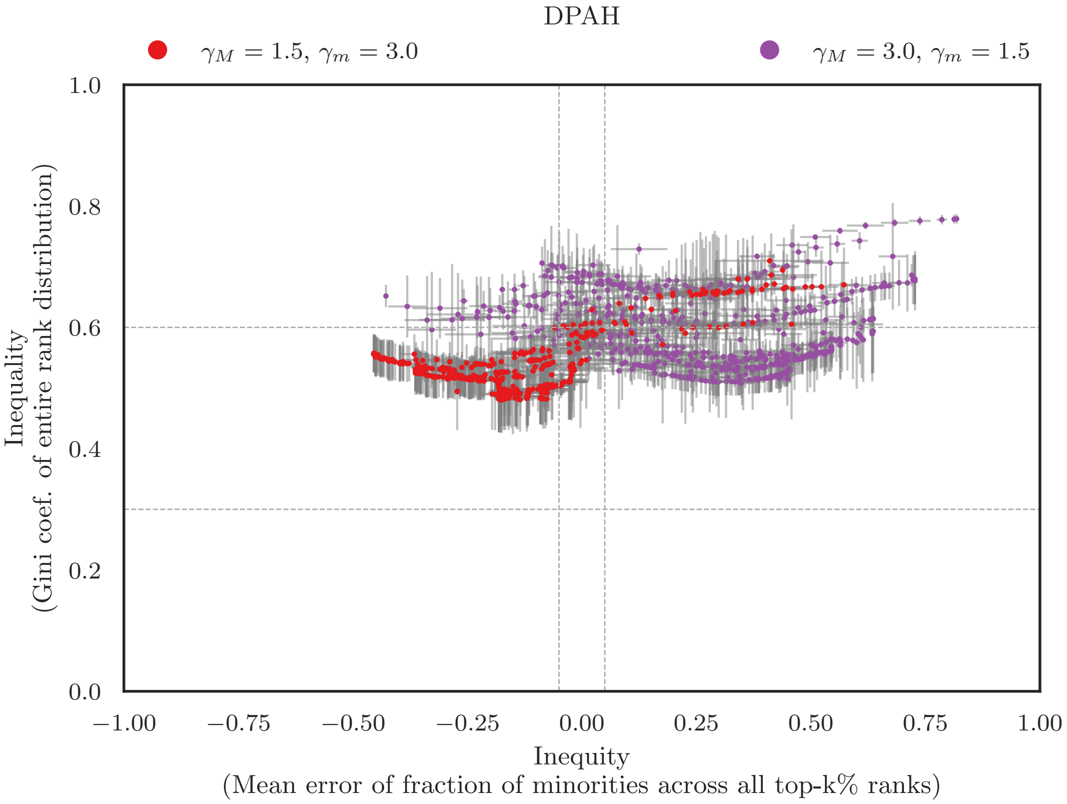


Figure 5.13.: **The effects of asymmetric node activity in the global disparity of PageRank.** Similar to Supplementary Figure 5.12, we generate DPAH networks by varying f_m, h_{MM}, h_{mm} and fixing number of nodes $n = 2000$ and edge density $d = 0.0015$. Additionally, we adjust γ_M and γ_m, the activity of majority and minority groups, respectively. Red represents networks where the majority is more active (higher out-degree) than the minority. In contrast, purple represents networks where the minority is more active than the majority. In comparison with their counterpart $\gamma = \gamma_M = \gamma_m = 3.0$ in Supplementary Figure 5.12, , we see that a more active minority ($\gamma_M = 3.0$, $\gamma_m = 1.5$), reduces inequality (lower Gini) by amplifying its visibility in the rank (positive ME). Conversely, a more active majority ($\gamma_M = 1.5$, $\gamma_m = 3.0$) can amplify minority representation at the cost of increasing inequality. However, in general, active majorities benefit themselves in the rank. Vertical and horizontal error bars represent the standard deviation over 10 runs of the Gini and ME, respectively.

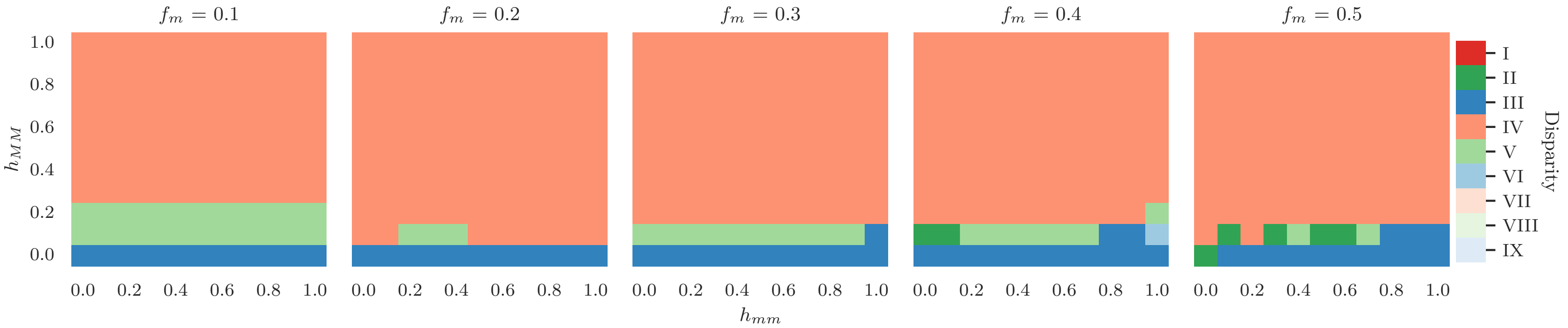


Figure 5.14.: **Disparities caused by an active majority group on PageRank as a function of homophily and group size.** Columns represent the fraction of minorities in the network, x-axis indicates the homophily within minorities, and y-axis the homophily within majorities. Colors denote the regions where the disparity lies according to our interpretation (see Figure 5.2). These results correspond to the red data points in Supplementary Figure 5.13, $\gamma_M = 1.5 < \gamma_m = 3.0$. We see that when the majority is more active than the minority, the algorithm reduces the representation of minorities in almost all combinations of homophily and group size. The exception lies in extreme heterophilic majorities. In this case, the ranking may amplify or replicate the visibility of minorities in the rank.

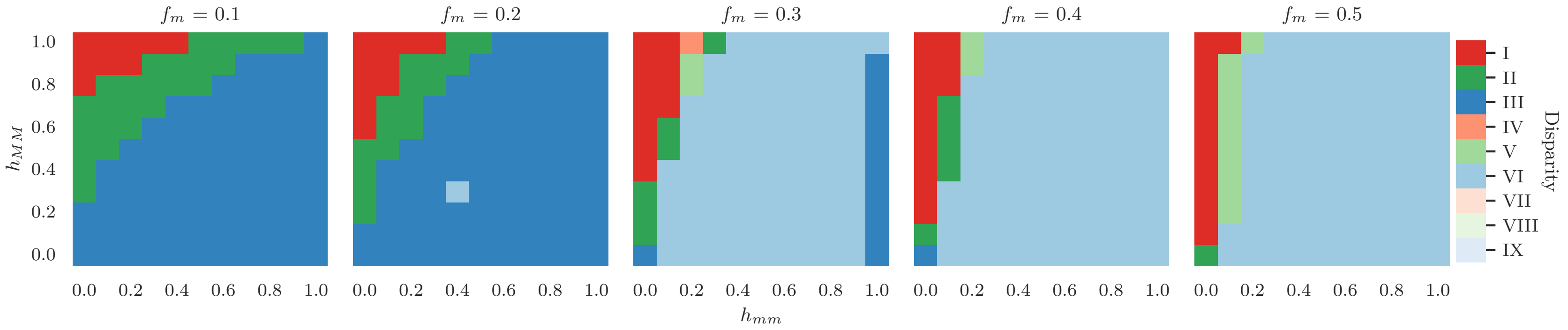


Figure 5.15.: **Disparities caused by an active minority group on PageRank as a function of homophily and group size.** Columns represent the fraction of minorities in the network, x-axis indicates the homophily within minorities, and y-axis the homophily within majorities. Colors denote the regions where the disparity lies according to our interpretation (see Figure 5.2). These results correspond to the purple data points in Supplementary Figure 5.13, $\gamma_M = 3.0 > \gamma_m = 1.5$. We see that when the minority is more active than the majority, the algorithm amplifies the representation of minorities in the rank in almost all combinations of homophily and group size. The exception is an interplay between the fraction of minorities and the homophily of the majority group when minorities are heterophilic.

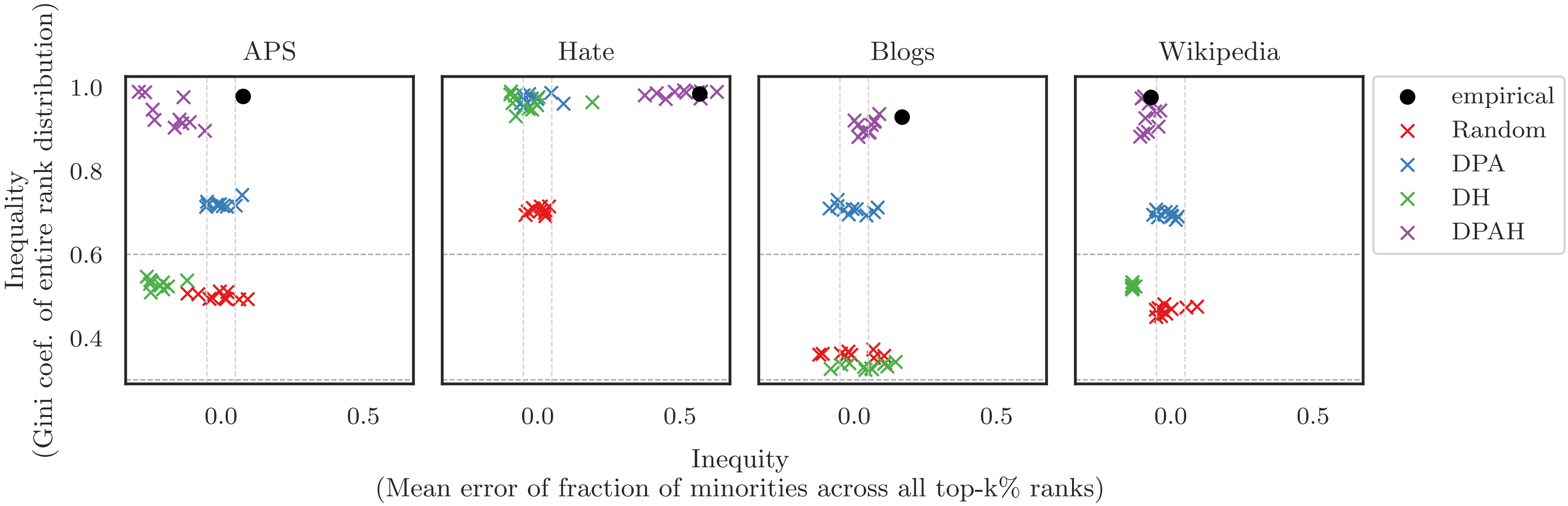


Figure 5.16.: **Global disparity in WTF on empirical networks.** Each column represents an empirical network. Citation/retweet networks (APS and Hate) and Hyper-link networks (Blogs and Wikipedia). Inequality and inequity are shown in the y- and x-axis, respectively. The disparities in ranking that we see in these empirical networks are best explained by preferential attachment and homophily DPAH.

6. Conclusions

For decades, social network analysis has been an interesting and very challenging field of study. Traditional social science research has demonstrated the existence of laws governing *edge formation* in social networks, and put a name to what we know today as cumulative advantage, small-world phenomena, and homophily [171, 177, 179]. Nowadays, social networks can be studied at large scale due to the availability of new kinds of data from social media, mobile phones, and sensors [131, 139, 191]. A common presumption is that more volume and variety of data often makes *machine learning* algorithms more accurate at inference and prediction [149]. However, these kinds of data can also introduce biases into the machine learning pipeline and therefore reinforce societal issues such as the glass ceiling effect, political polarization, urban segregation and denigration [70, 169, 231, 243]. While most studies have focused on calibrating the algorithms to adjust their outcomes, quantifying the impact of biased networked data on machine learning outcomes, has remained relatively neglected. This thesis aims to fill that gap.

One of the advantages of working with networks is that they possess a *structure* defined by the mechanisms of edge formation. Many societal issues can be explained by such mechanisms. For instance, cumulative advantage—also known as preferential attachment or the rich-get-richer effect—is the causal mechanism for many skewed distributions such as the income inequality growth [60]. Therefore, in this thesis I systematically evaluate to what extent the structure of a network reinforces such biases in machine learning outcomes. In particular, outcomes from classification, ranking and recommendation algorithms. The goal is to make existing network-based algorithms *interpretable* and *explainable*. Thereby, data scientists can anticipate erroneous and unfair results, and thus propose interventions accordingly to break the feedback loop of reinserting those biases into the network.

Note that not all algorithms must be completely unbiased. For example, if a user is looking for a relationship in Tinder and her preferences are men in their 30's, there is no reason why the application should show her profiles of younger and older men, or women. However, the application could still do a fair job within that sub-population, based on other characteristics such as religion, profession, etc. When biased results are an issue, a step towards mitigation is to first identify and understand the main source of bias [246]. Since in this thesis I focus on social networks, the main source of bias is its structure, i.e., mechanisms of edge formation. In chapter 2 and Chapter 3, I propose two techniques to characterize these mechanisms which help to understand how edges form in given real-world networks. An additional advantage of analyzing social networks is that we can control for its structure through generative network models. In Chapter 4 and Chapter 5, I leverage the benefits of network modeling to then quantify the influence of different network structures on classification, ranking and recommendation algorithms.

6.1. Results and Contributions

In this section I summarize the findings and contributions of this thesis by addressing each research question.

RQ1 How can we characterize the underlying mechanisms of edge formation of a given network?

RQ2 How do the mechanisms of edge formation influence classification performance and the direction of bias in relational classification?

RQ3 How do the mechanisms of edge formation affect the distribution of opportunities across individuals and minorities in ranking and recommendation algorithms?

First, I start addressing **RQ1** by presenting HopRank and Janus, two methods that help us to understand how edges form in social networks. The former leverages the distance between nodes to explain human navigation on networks, and the latter ranks hypotheses based on homophily and preferential attachment. Second, I address **RQ2** and **RQ3** by showing how homophily, preferential attachment, fraction of minorities, and edge density affect minorities in network-based classification, ranking and recommendation algorithms. Third, besides these methods and benchmarks, I show my support of digital scholarship by making all the code and data openly available on GitHub.

1. **HopRank:** I propose HopRank a method for characterizing edge formation in *non-attributed* networks. HopRank models transitions (from one node to another) across k-hop neighborhoods on semantic networks (e.g., memory and ontologies). This approach is motivated by *information foraging* [203] and *decentralized search* [113] where people, in principle, tend to acquire information from two stages: exploration (breadth-low-cost navigation) and exploitation (depth-high-cost navigation). Specifically, HopRank extends PageRank where instead of having only one teleportation or damping factor (i.e., probability of following links), it relies on the HopPortation vector which defines the probabilities of transitioning to any node at a k-hop distance. I applied this method to model human navigation on BioPortal—a repository of biomedical ontologies—and found that semantic structure does influence navigation on networks. In particular, users tend to prefer certain k-hop neighborhoods more often than others depending on the type

of navigation (e.g., search, menu expand, direct link, etc.). For instance, when manually browsing the tree-like explorer on BioPortal, users tend to hop to nearby concepts, whereas far-away concepts are more likely to be reached by non-browsing types such as external links. These results advance the understanding of how ontologies are actually navigated and consumed (i.e., neither fully random nor only by following links), and help to develop and improve user interfaces for ontology exploration. Note that, while in this work I focus on navigation on BioPortal, HopRank can be used to model navigation on any network. The main conditions are that nodes must be non-attributed and people must have some prior knowledge about the network they are navigating (e.g., they know it or can see it). The latter pre-condition helps people navigate the network by optimizing resources (e.g., time). For instance, another use case could be navigating a network of places (e.g., a city), where people would often go to the closest supermarket and to a not so close-by workplace. When nodes are attributed, however, I leverage that extra information to create more informative hypotheses and rank them using Janus.

2. **Janus:** I propose Janus, the counterpart of HopRank, a method for characterizing edge formation in *node-attributed* networks and multiplex networks. Janus computes Bayes factors with marginal likelihood estimations to compare the relative plausibility of expressed hypotheses as they are specifically sensitive to the priors. These hypotheses are built using the attributes of the nodes, and are based on *homophily* and *preferential attachment*. In terms of scalability and interpretability, Janus outperforms similar methods such as statistical regression methods based on QAP [122] or mixed-effects models [230]. I applied Janus on synthetic networks to demonstrate the robustness of the hypothesis ranking, and on real-world social networks to understand how social interactions develop. In particular, interactions among (i) members of five different households [240] and (ii) Twitter users on a particular topic [237]. Results show that (i) among the members of the five households, having a similar age or the same sex does not influence the amount of times they interact with each other, and (ii) the number of replies one user gives to another in Twitter is not random and correlates with the number of mentions. Janus is then relevant for data scientists interested in studying the mechanisms that can explain edge formation in networks (with node attributes or multi-layer networks) from both empirical and methodological perspectives.

3. **Evaluation benchmarks for relational classification:** The way we connect to other people often carries meaningful information about ourselves. For instance, the political leaning of your friends can serve as a proxy of your own political preferences because we tend to be friends with people who are similar to us. This principle is leveraged by *relational classification*

[43], a machine learning technique that is trained from a known portion of the graph, and then transfers this knowledge to the rest of the unseen nodes. Naturally, there are multiple ways of learning such relational model [165] and previous work shows how certain network properties affect the outcomes of these algorithms [84, 227, 274]. However, to the best of my knowledge there is no clear consensus on how the performance and bias of relational classification is affected by co-existing characteristics of the network such as *preferential attachment*, *homophily*, *fraction of minorities*, and *edge density*. To this end, in this work I conduct a systematic study on synthetic and real-world networks to show when and why classification performance and the direction of bias occur as a function of network structure, sampling technique and sample size. My findings show that homophily, fraction of minorities and the sampling method make a considerable difference when training samples are small. In this case, if the goal is to achieve *high accuracy* (i.e., when most nodes are correctly classified), then heterophilic networks (i.e., when most connections happen between nodes that are not alike) should be sampled by random edge sampling or partial crawls [12], and homophilic networks should be sampled by degree or random edge sampling. However, if the goal is to achieve *high quality of model parameters* (i.e., when the proportions of nodes and edges per type in the training sample are similar to the ones in the whole network), then balanced networks (i.e., when groups have the same size) must be sampled by degree, and unbalanced networks by partial crawls. Additionally, we find that edge density improves accuracy only when networks are close to neutral (i.e., when they are slightly heterophilic or homophilic) and training samples are small. These findings suggest that performance and bias in relational classification can be anticipated and therefore mitigated before relying on erroneous results. Consequently, this thesis extends the methodological toolbox for studying transparency in machine learning models by proposing a first step towards an interpretable and explainable relational model.

4. **Evaluation benchmarks for network-based ranking:** Similar to the classification problem, ranking algorithms that learn how to rank nodes merely on network structure also amplify biases found in data. For instance, it has been shown that recommender systems such as *Who-to-Follow* [106] tend to increase the popularity of users who are already popular [244]. However, it is unclear how exactly the ranking is affected by the properties of the network, either individually or in combination. Understanding these effects is crucial when such algorithms are used in decision making processes since they could amplify discrimination against under-represented groups. This motivated me to study when *inequality* and *inequity* arise in PageRank and Who-To-Follow, two well-known network-based ranking and recommendation algorithms, respectively. Inequality measures the skewness

of the rank distribution across all individuals, and inequity measures how much the proportional representation of minorities in top ranks deviates from a given diversity constraint or quota. Note that this quota is flexible and can be adjusted depending on the context of the application. In this study, the diversity quota reflects the representation of minorities in the whole network. Therefore, if the inequity or deviation is very small, then the ranking is fair. Depending on the sign of the deviation the ranking may under-represent or over-represent minorities in top ranks. I also propose a growth model for directed networks which allows to adjust certain properties of the network such as homophily, node activity, edge density and the fraction of minorities. This model generates power-law in-degree and out-degree distributions, two characteristics found in many social networks [201, 255]. By fitting this model to real-world networks I show how edges are formed and thus explain and quantify the inequality and inequity that these algorithms produce. Additionally, by using synthetic networks, I demonstrate that homophily is the main driver of inequity, while inequality is driven by the interplay between homophily, preferential attachment, node activity, and edge density. This study shows that although majorities dominate the general behavior of the network—not only in group size but also in number of connections—minorities can help to mitigate their under-representation in top ranks. For instance, if the minority is very small, and people from the majority connect mostly with each other, then the minority should also connect with each other to surpass the homophily of the majority and achieve a fair rank. While this strategy works at reducing inequity, it may split the network into two isolated groups. To avoid such a split, a better strategy is to involve both groups, and increase the number of connections from majority to minority. For individuals, on the other hand, being more active (i.e., more out-going links), and in general having more connections, helps reducing the skewness of the ranking distributions. This work adds to the methodological toolbox for studying transparency in machine learning models by proposing first steps towards interpretable and explainable network-based ranking and recommendation algorithms.

5. **Code and data:** All code used to generate intermediate and final results and plots for each of these projects are publicly available through GitHub repositories. I followed best coding practices and used object-oriented programming in Python. Code for generating synthetic networks and intermediate results is run through multiple parallel batches on the command line, and code to generate the plots that appear in the final manuscripts is often instantiated in Jupyter Notebooks: [74], [73], [75], [76].

6.2. Implications and Applications

The contributions mentioned before have important implications for society as well as for the design and execution of machine learning methods that rely on network structure. Some of these are briefly discussed below.

Recommender systems: The purpose of a recommender system is to suggest relevant items to users. In Chapter 5, I studied PageRank and Who-To-Follow, two algorithms which have been the core of many recommender systems. PageRank was used to originally *rank* all web pages in the Web by the Google search engine. Since its conception it has also been used to rank authors, scientific papers and baseball players [99]. Who-To-Follow was implemented by Twitter to *recommend* users new people to follow. These algorithms, however, exacerbate bubble and popularity effects [64] because the fairness component is often dismissed. In this thesis, a *fair rank* is such that its top-k preserves the proportional representation of groups in the whole network. Based on this definition, my findings suggest that the magnitude and direction of bias in top ranks mainly depend on the level of homophily. Interestingly, this bias is not always against the minority group, especially in cases such as *celebrity-fan* or *politician-voter* where the majority group has sparse connections among each other, but they mostly connect to the popular minorities. These results, demonstrate under which conditions ranking and recommender algorithms may ensure fairness or reinforce inequity in society. For example, in a citation network where men represent 90% of all nodes and women only 10%, if men mostly cite other men (i.e., high homophily), then women will be under-represented in the top of the rank. Conversely, if men mostly cite women (i.e., high heterophily), then men will be under-represented in the top of the rank compared to their actual representation in the network. If we want the rank to be fair, men should be neutral (i.e., equally likely to cite men or women) and women should not be extremely homophilic. Therefore, based on these conditions, we can recommend the important missing links that balance the importance of nodes and groups in the network.

Handling missing values: Often, applied machine learning practitioners are faced with real-world data that has multiple missing values. Reasons for having missing data range from corrupted records and outdated datasets to historical biases. While there are several ways of handling missing values [5, 69, 161], I would like to focus on the ones using machine learning to predict missing attributes of individuals. If the data is a social network, and some people are lacking attributes such as gender, my work in Chapter 4 can help to not only infer those missing attributes, but also to evaluate the results. The evaluation benchmarks that I propose give the analyst the confidence to trust (or not) the classification outcomes. For instance, given the homophily of a social network, its fraction of minorities and the sample size, the analyst can be informed in advanced of the performance and the direction of bias of the results.

Interface design: In Chapter 2, I study human navigation on semantic networks. While I particularly focused on BioPortal, a repository of biomedical ontologies, the study is suitable for the analysis of human navigation on any network. The main conditions are that nodes must be non-attributed and people must have some prior knowledge about the network they are navigating (e.g., they know it or can see it). In the particular case of digital networks, often they are presented as tree-like structures including hyper-links to related nodes (e.g., parent and children nodes). Some platforms offer a search functionality to jump directly to specific nodes that are not visible in the near neighborhood of the current node. Therefore, by understanding the dynamics of navigation in such systems, we could identify signals for possible interfaces adjustments to facilitate navigation and enhance the user experience. For instance, besides having a visible direct access to only 1-hop neighbors, an improved interface may show nodes that other users also visited after visiting the current node. Another potential improvement is to highlight or show only relevant nodes by feeding the information about popular visited paths into the fisheye layout [92, 220].

6.3. Limitations and Future Work

Finally, this thesis represents an early attempt to understand and explain the effects of biased networked data on machine learning outcomes. More work is necessary to mitigate the harms that machine learning algorithms are causing to society.

- **Generalization of results.** First, the empirical findings using real-world networks in this thesis are not generalizable to all networks. For instance, the ranking of hypotheses to explain the contact network in Section 3.5.3 is only valid for this particular network; it does not generalize to all contact networks in the world. Second, the proposed benchmarks in Chapters 4 and 5 (based on synthetic networks) are generalizable to some extent to networks that possess certain properties that are frequently observed in social networks. The main conditions are that such networks must exhibit preferential attachment (i.e., skewed degree distributions), homophily, and connections triggered by humans. Nevertheless, the code for all methodologies proposed in this thesis is openly available for reproducibility and reusability. Thus, other types of networks could be used to discover new knowledge in different domains.
- **Multivariate analysis.** For simplicity, this thesis focuses on networks whose nodes possess only a single binary attribute. In the context of classification (see Chapter 4), this decision was made to facilitate the explanation of results. Future research should incorporate protected and unprotected attributes to capture more complex structures in explainable and interpretable relational models. For instance, to investigate the role of protected

attributes such as gender in the approval of credit loans that are based on the credit score of friends. This means that, protected and non-protected attributes should be studied in combination and go beyond the binary domain. Similarly, in the context of ranking (see Chapter 5), I also focused on a single binary attribute to define minority and majority groups. A possible avenue for future studies should incorporate categorical attributes such as nationality, social status, and religion. Possible research questions in this direction include: "What are the structural conditions that all groups in a network need to fulfill in order to achieve a fair top-k rank?", and "How do the different groups of people compete for attention in the top of the rank?".

- **Network structure.** In this thesis I mainly focus on four prominent properties of social networks: preferential attachment, homophily, fraction of minorities and edge density. In Chapter 4, we see that one real-world dataset (i.e., GitHub) was not fully captured by the benchmarks. One explanation is that the synthetic networks were generated with symmetric homophily (i.e., $h_{min} = h_{maj} = H$), although the real-world networks exhibit asymmetric homophily. While the model fitting was not perfect, the general behavior was captured and could explain high and low classification performance given the type of network and sample size. Therefore, future directions should include asymmetric homophily in undirected networks, directed networks, and many other properties such as monophily [6] and triadic closure. Note that in the case of monophily, it could be used as a network property to control for in the network generation, or in the learning phase to learn the model parameters. For instance, instead of learning conditional probabilities using 1-hop neighbors, we could learn them using 2-hops.

- **Beyond selected algorithms.** Similar to the ranking problem, modeling and forecasting the spread of infectious diseases or information in social networks may suffer from fairness issues. In the context of disease outbreaks, it has been shown that poverty and social determinants of health create conditions to further contribute to unequal burdens of morbidity and mortality [206]. However, to the best of my knowledge, most of the algorithms that have been proposed to model this phenomena [16, 137, 181, 279] ignore the fairness component. Most spreading processes are optimized to reach multiple people (or contain a virus) very fast by inferring influential nodes, also known as "super spreaders". However, this optimization does not guarantee a fair treatment to all groups in a network. In fact, under-represented groups are often ignored due to the unequal allocation of resources suggested by these algorithms. Towards this goal, Vijayshankar and Roy [254] analytically demonstrated the cost of fairness in disease spread control in contact network models. However, it remains unclear how each mechanism of edge formation (or in combination) affects the selection of influential nodes, and to what extent they contribute to the unequal allocation of resources. I believe that a systematic study—such as the one I conducted in Chapter 5 to rank people in a network—can answer these questions.

- **Adversarial behavior.** A valid concern for explainability is that the more knowledge a person has about the model, the easier it is to attack or trick the system into making a decision favorable to the adversary [115]. In fact, in Chapter 5 one of the conclusions is that our results can help to reduce inequalities without algorithmic intervention by connecting strategically in the network. This means that, a minority group in disadvantage could strategically change its behavior to gain more visibility in the rank. What if this group is an adversarial attacker? In this case, minorities will unfairly increase their visibility in top ranks at the expense of the majority. For example, articles from predatory journals trying to appear in top search results, or bots in social networks trying to trick the recommendation algorithm to appear as friend suggestions. On the other hand, if the adversarial attacker is the majority group, we can expect more harm in the network since the minority group will always be in disadvantage. Fortunately, the machine learning community has already proposed techniques to reduce or prevent adversarial attacks [21, 63, 184]. However, more work needs to be done in the context of classification, ranking and recommendation algorithms applied to networked data.

- **Computational power.** In general, the code generated in this thesis optimizes RAM by representing graphs as sparse matrices in Python (e.g., `scipy.sparse.csr_matrix`) for HopRank [74], Janus [73], and [76] to compute PageRank and the Who-To-Follow scores. However, that might not be enough for very large networks especially for calculating all possible k-hops in HopRank, since sparsity might be reduced for some k's. Recall that k goes from 1 to the diameter of the network. Roughly, I required 1TB of RAM to allocate 10.175M cells for calculating the k-hops of LOINC, the largest ontology from our datasets with 101K nodes. In the case of Janus, as I mentioned in Section 3.6, it outperforms its counterpart QAP thanks to the sparsity of the networks. However, the local model of Janus could speed up the computation of posterior probabilities by using parallelization, similar to [23]. The work on relational classification [75] could also benefit from the sparsity of networks to not only use less RAM but also to make the computation of posterior probabilities faster. The current version represents networks as `networkx.Graph` and computes the posterior probabilities for each node sequentially.

Bibliography

[1] Joshua T Abbott, Joseph L Austerweil, and Thomas L Griffiths. "Random walks on semantic networks can resemble optimal foraging." In: 122.3 (2015), page 558.

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. "The unfairness of popularity bias in recommendation." In: *CEUR Workshop Proceedings* 2440 (2019). [Online; accessed 02-June-2021].

[3] Lada A Adamic and Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog." In: *Proceedings of the 3rd int. workshop on Link discovery.* ACM. 2005, pages 36–43.

[4] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks." In: *Reviews of modern physics* 74.1 (2002), page 47.

[5] Paul D Allison. *Missing data.* Volume 136. Sage publications, 2001.

[6] Kristen M Altenburger and Johan Ugander. "Monophily in social networks introduces similarity among friends-of-friends." In: *Nature Human Behaviour* 2.4 (2018), page 284.

[7] Ian Anderson, Santiago Gil, Clay Gibson, Scott Wolf, Will Shapiro, Oguz Semerci, and David M Greenberg. ""Just the Way You Are": Linking Music Listening on Spotify and Personality." In: *Social Psychological and Personality Science* 12.4 (2021), pages 561–572.

[8] Aili Asikainen, Gerardo Iñiguez, Javier Ureña-Carrión, Kimmo Kaski, and Mikko Kivelä. "Cumulative effects of triadic closure and homophily in social networks." In: *Science Advances* 6.19 (2020), eaax7310.

[9] Abolfazl Asudeh, HV Jagadish, Julia Stoyanovich, and Gautam Das. "Designing fair ranking schemes." In: *Proceedings of the 2019 International Conference on Management of Data.* 2019, pages 1259–1276.

[10] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. "Dbpedia: A nucleus for a web of open data." In: *The semantic web* (2007), pages 722–735.

[11] Chen Avin, Barbara Keller, Zvi Lotker, Claire Mathieu, David Peleg, and Yvonne-Anne Pignolet. "Homophily and the glass ceiling effect in social networks." In: *Proceedings of the 2015 conference on innovations in theoretical computer science.* 2015, pages 41–50.

[12] Konstantin Avrachenkov, Bruno Ribeiro, and Jithin K Sreedharan. "Inference in OSNs via Lightweight Partial Crawls." In: *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*. ACM. 2016, pages 165–177.

[13] Alexander Bachmann, Alexander Becker, Daniel Buerckner, Michel Hilker, Frank Kock, Mark Lehmann, Phillip Tiburtius, and Burkhardt Funk. "Online peer-to-peer lending-a literature review." In: *Journal of Internet Banking and Commerce* 16.2 (2011), page 1.

[14] Lars Backstrom and Jon Kleinberg. "Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook." In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM. 2014, pages 831–841.

[15] Lars Backstrom and Jure Leskovec. "Supervised random walks: predicting and recommending links in social networks." In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM. 2011, pages 635–644.

[16] Norman TJ Bailey et al. *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

[17] Albert-László Barabási. "Scale-free networks: a decade and beyond." In: *science* 325.5939 (2009), pages 412–413.

[18] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks." In: *science* 286.5439 (1999), pages 509–512.

[19] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. "Who to follow and why: link prediction with explanations." In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014, pages 1266–1275.

[20] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. "The problem with bias: from allocative to representational harms in machine learning. Special Interest Group for Computing." In: *Information and Society (SIGCIS)* 2 (2017).

[21] Solon Barocas, Andrew D Selbst, and Manish Raghavan. "The hidden assumptions behind counterfactual explanations and principal reasons." In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pages 80–89.

[22] Marcia J Bates. "The design of browsing and berrypicking techniques for the online search interface." In: *Online review* 13.5 (1989), pages 407–424.

[23] Martin Becker, Hauke Mewes, Andreas Hotho, Dimitar Dimitrov, Florian Lemmerich, and Markus Strohmaier. “Sparktrails: A mapreduce implementation of hyptrails for comparing hypotheses about human trails.” In: *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee. 2016, pages 17–18.

[24] Alejandro Bellogín, Pablo Castells, and Iván Cantador. “Statistical biases in Information Retrieval metrics for recommender systems.” In: *Information Retrieval Journal* 20.6 (2017), pages 606–634.

[25] Reuben Binns. “On the apparent conflict between individual and group fairness.” In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pages 514–524.

[26] *BioPortal*. Accessed: 2019-02-21. 2011. URL: `https://bioportal.bioontology.org/`.

[27] Per Block. “Reciprocity, transitivity, and the mysterious three-cycle.” In: *Social Networks* 40 (2015), pages 163–173.

[28] Michał Bojanowski and Rense Corten. “Measuring segregation in social networks.” In: *Social Networks* 39 (2014), pages 14–32.

[29] S. P. Borgatti, K. Carley, and D. Krackhardt. “Robustness of Centrality Measures under Conditions of Imperfect Data.” In: *Social Networks* 28.1 (2006), pages 124–136.

[30] Stephen P Borgatti, Martin G Everett, and Jeffrey C Johnson. *Analyzing social networks*. Sage, 2018.

[31] Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. “Network analysis in the social sciences.” In: *science* 323.5916 (2009), pages 892–895.

[32] Wendy Bottero and Nick Crossley. “Worlds, fields and networks: Becker, Bourdieu and the structures of social relations.” In: *Cultural sociology* 5.1 (2011), pages 99–119.

[33] Sergey Brin and Lawrence Page. “The anatomy of a large-scale hypertextual web search engine.” In: *Computer networks and ISDN systems* 30.1-7 (1998), pages 107–117.

[34] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. “The balanced accuracy and its posterior distribution.” In: *2010 20th International Conference on Pattern Recognition*. IEEE. 2010, pages 3121–3124.

[35] Michael Brownstein. “Implicit bias.” In: (2015).

[36] Taina Bucher. “Want to be on the top? Algorithmic power and the threat of invisibility on Facebook.” In: *New media & society* 14.7 (2012), pages 1164–1180.

[37] Ronald S Burt. “Positions in networks.” In: *Social forces* 55.1 (1976), pages 93–122.

[38] Ronald S Burt. “The social structure of competition.” In: *Networks in the knowledge economy* 13 (2003), pages 57–91.

[39] Francesco Capotorti. *Study on the rights of persons belonging to ethnic, religious and linguistic minorities.* Volume 384. New York: United Nations, 1979.

[40] Stuart K Card, Peter Pirolli, Mija Van Der Wege, Julie B Morrison, Robert W Reeder, Pamela K Schraedley, and Jenea Boshart. “Information scent as a driver of Web behavior graphs: results of a protocol analysis method for Web usability.” In: *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM. 2001, pages 498–505.

[41] Giona Casiraghi, Vahan Nanumyan, Ingo Scholtes, and Frank Schweitzer. “Generalized Hypergeometric Ensembles: Statistical Hypothesis Testing in Complex Networks.” In: *arXiv:1607.02441* (2016).

[42] Lidia Ceriani and Paolo Verme. “The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini.” In: *The Journal of Economic Inequality* 10.3 (2012), pages 421–443.

[43] Soumen Chakrabarti, Byron Dom, and Piotr Indyk. “Enhanced hypertext categorization using hyperlinks.” In: *ACM SIGMOD Record.* Volume 27. 2. ACM. 1998, pages 307–318.

[44] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. “SMOTE: synthetic minority over-sampling technique.” In: *Journal of artificial intelligence research* 16 (2002), pages 321–357.

[45] Irene Chen, Fredrik D Johansson, and David Sontag. “Why is my classifier discriminatory?” In: *Advances in Neural Information Processing Systems.* 2018, pages 3539–3550.

[46] Ed H Chi, Peter Pirolli, Kim Chen, and James Pitkow. “Using information scent to model user information needs and actions and the Web.” In: *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM. 2001, pages 490–497.

[47] Ed H Chi, Peter Pirolli, and James Pitkow. “The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site.” In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems.* ACM. 2000, pages 161–168.

[48] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. “A fair classifier using mutual information.” In: *2020 IEEE International Symposium on Information Theory (ISIT).* IEEE. 2020, pages 2521–2526.

[49] David A Cieslak and Nitesh V Chawla. “A framework for monitoring classifiers’ performance: when and why failure occurs?” In: *Knowledge and Information Systems* 18.1 (2009), pages 83–108.

[50] James S Coleman. “Social capital in the creation of human capital.” In: *American journal of sociology* 94 (1988), S95–S120.

[51] Michele Coscia and Luca Rossi. “Benchmarking API costs of network sampling strategies.” In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE. 2018, pages 663–672.

[52] E. Costenbader and T. W. Valente. “The stability of centrality measures when networks are sampled.” In: *Social Networks* 25.4 (Oct. 2003), pages 283–307. ISSN: 0378-8733. DOI: 10.1016/s0378-8733(03)00012-1.

[53] David A. Cotter, Joan M. Hermsen, Seth Ovadia, and Reeve Vanneman. “The Glass Ceiling Effect.” In: *Social Forces* 80.2 (Dec. 2001), pages 655–681. ISSN: 0037-7732. DOI: 10.1353/sof.2001.0091.

[54] Bo Cowgill, Fabrizio Dell’Acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. “Biased programmers? or biased data? a field experiment in operationalizing ai ethics.” In: *Proceedings of the 21st ACM Conference on Economics and Computation*. 2020, pages 679–681.

[55] Kate Crawford. “The trouble with bias.” In: *Conference on Neural Information Processing Systems (NIPS)*. [Online; accessed 26-May-2021]. 2017.

[56] James A Davis. “Clustering and hierarchy in interpersonal relations: Testing two graph theoretical models on 742 sociomatrices.” In: *American Sociological Review* (1970), pages 843–851.

[57] Manlio De Domenico, Antonio Lima, Paul Mougel, and Mirco Musolesi. “The anatomy of a scientific rumor.” In: *Scientific reports* 3 (2013).

[58] Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. “Echo chambers: Emotional contagion and group polarization on facebook.” In: *Scientific reports* 6 (2016), page 37825.

[59] Ying Ding, Erjia Yan, Arthur Frazho, and James Caverlee. “PageRank for ranking authors in co-citation networks.” In: *Journal of the American Society for Information Science and Technology* 60.11 (2009), pages 2229–2243.

[60] Thomas A DiPrete and Gregory M Eirich. “Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments.” In: *Annu. Rev. Sociol.* 32 (2006), pages 271–297.

[61] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. “Measuring and mitigating unintended bias in text classification.” In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pages 67–73.

[62] Sa Dong, Dayou Liu, Ruochuan Ouyang, Yungang Zhu, Lina Li, Tingting Li, and Jie Liu. "Second-Order Markov Assumption Based Bayes Classifier for Networked Data with Heterophily." In: *IEEE Access* (2019).

[63] Georgios Douzas and Fernando Bacao. "Effective data generation for imbalanced learning using conditional generative adversarial networks." In: *Expert Systems with applications* 91 (2018), pages 464–471.

[64] Marina Drosou, HV Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. "Diversity in big data: A review." In: *Big data* 5.2 (2017), pages 73–84.

[65] Martin Dufwenberg and Amrish Patel. "Reciprocity networks and the participation problem." In: *Games and Economic Behavior* 101 (2017), pages 260–272.

[66] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through awareness." In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012, pages 214–226.

[67] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. "Decoupled classifiers for group-fair and efficient machine learning." In: *Conference on Fairness, Accountability and Transparency*. 2018, pages 119–133.

[68] David Easley, Jon Kleinberg, et al. "Power laws and rich-get-richer phenomena." In: *Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press* (2010).

[69] Craig K Enders. *Applied missing data analysis*. Guilford press, 2010.

[70] Robert Epstein and Ronald E Robertson. "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections." In: *Proceedings of the National Academy of Sciences* 112.33 (2015), E4512–E4521.

[71] Erdös, Paul, and Alfréd, Rényi. "On random graphs." In: *Publicationes Mathematicae* 6 (1959), pages 290–297.

[72] Lisette Espín Noboa, Florian Lemmerich, Philipp Singer, and Markus Strohmaier. "Discovering and characterizing mobility patterns in urban spaces: A study of manhattan taxi data." In: *Proceedings of the 25th International Conference Companion on World Wide Web*. 2016, pages 537–542. DOI: `10.1145/2872518.2890468`.

[73] Lisette Espín-Noboa. *Janus GitHub repository*. `https://github.com/lisette-espin/JANUS`. Accessed: 2017-03-10. 2017.

[74] Lisette Espín-Noboa. *HopRank GitHub repository*. `https://github.com/gesiscss/HopRank`. Accessed: 2019-01-23. 2019.

[75] Lisette Espín-Noboa. *Biases in relational classification GitHub repository.* `https://github.com/gesiscss/Discrimination-in-Relational-Classification`. 2021.

[76] Lisette Espín-Noboa. *Homophilic Directed scale-free Networks GitHub repository.* `https://github.com/gesiscss/Homophilic_Directed_ScaleFree_Networks`. 2021.

[77] Lisette Espín-Noboa, Fariba Karimi, Bruno Ribeiro, Kristina Lerman, and Claudia Wagner. "Explaining Classification Performance and Bias via Network Structure and Sampling Technique." In: *Applied Network Science* 6.1 (2021), page 78. DOI: `10.1007/s41109-021-00394-3`.

[78] Lisette Espín-Noboa, Florian Lemmerich, Markus Strohmaier, and Philipp Singer. "JANUS: A hypothesis-driven Bayesian approach for understanding edge formation in attributed multigraphs." In: *Applied Network Science* 2.1 (2017), page 16.

[79] Lisette Espín-Noboa, Florian Lemmerich, Simon Walk, Markus Strohmaier, and Mark Musen. "HopRank: How Semantic Structure Influences Teleportation in PageRank (A Case Study on BioPortal)." In: *The World Wide Web Conference*. WWW '19. San Francisco, CA, USA: Association for Computing Machinery, 2019, pages 2708–2714. ISBN: 9781450366748. DOI: `10.1145/3308558.3313487`.

[80] Lisette Espín-Noboa, Claudia Wagner, Fariba Karimi, and Kristina Lerman. "Towards Quantifying Sampling Bias in Network Inference." In: *Companion Proceedings of the The Web Conference 2018*. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pages 1277–1285. ISBN: 9781450356404. DOI: `10.1145/3184558.3191567`.

[81] Lisette Espín-Noboa, Claudia Wagner, Markus Strohmaier, and Fariba Karimi. "Inequality and inequity in network-based ranking and recommendation algorithms." In: *Scientific reports* 12.1 (2022), pages 1–14. DOI: `10.1038/s41598-022-05434-1`.

[82] Anna Evtushenko and Jon Kleinberg. "The paradox of second-order homophily in networks." In: *Scientific Reports* 11.1 (2021), pages 1–10. DOI: `10.1038/s41598-021-92719-6`.

[83] Francesco Fabbri, Francesco Bonchi, Ludovico Boratto, and Carlos Castillo. "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*. Volume 14. 2020, pages 165–175.

[84] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. "Fairness in relational domains." In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pages 108–114.

[85] Dalibor Fiala, François Rousselot, and Karel Ježek. "PageRank for bibliographic networks." In: *Scientometrics* 76.1 (2008), pages 135–158. DOI: `10.1007/s11192-007-1908-4`.

[86] Seth Flaxman, Sharad Goel, and Justin M Rao. "Filter bubbles, echo chambers, and online news consumption." In: *Public opinion quarterly* 80.S1 (2016), pages 298–320.

[87] Santo Fortunato, Marián Boguná, Alessandro Flammini, and Filippo Menczer. "On local estimations of PageRank: a mean field approach." In: *Internet Mathematics* 4.2-3 (2007), pages 245–266.

[88] Santo Fortunato, Marián Boguñá, Alessandro Flammini, and Filippo Menczer. "Approximating PageRank from in-degree." In: *International workshop on algorithms and models for the web-graph*. Springer. 2006, pages 59–71.

[89] Anderson J Franklin and Nancy Boyd-Franklin. "Invisibility syndrome: A clinical model of the effects of racism on African-American males." In: *American Journal of Orthopsychiatry* 70.1 (2000), pages 33–41.

[90] Jonathan L Freedman and David O Sears. "Selective exposure." In: *Advances in experimental social psychology*. Volume 2. Elsevier, 1965, pages 57–97.

[91] Linton Freeman. "The development of social network analysis." In: *A Study in the Sociology of Science* 1 (2004), page 687.

[92] George W Furnas. *Generalized fisheye views*. Volume 17. 4. ACM New York, NY, USA, 1986, pages 16–23.

[93] Joseph Galaskiewicz. "Estimating point centrality using different network sampling techniques." In: *Social Networks* 13.4 (Dec. 1991), pages 347–386. DOI: `10.1016/0378-8733(91)90002-B`.

[94] Florian Geigl, Daniel Lamprecht, Rainer Hofmann-Wellenhof, Simon Walk, Markus Strohmaier, and Denis Helic. "Random surfers on a web encyclopedia." In: *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*. ACM. 2015, page 5.

[95] GESIS. *Temporal Network of Politicians on Wikipedia*. `https://github.com/gesiscss/Wikipedia-Politician-Network`. Accessed: 2020-10-07. 2018.

[96] Lise Getoor and Ben Taskar. *Introduction to statistical relational learning*. MIT press, 2007.

[97] Gourab Ghoshal and Albert-László Barabási. "Ranking stability and super-stable nodes in complex networks." In: *Nature communications* 2 (2011), page 394.

[98] Corrado Gini. "Variabilità e mutabilità." In: *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E* (1912).

[99] David F Gleich. "PageRank beyond the Web." In: *SIAM Review* 57.3 (2015), pages 321–363.

[100] Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. "A survey of statistical network models." In: *Foundations and Trends® in Machine Learning* 2.2 (2010), pages 129–233.

[101] Sujatha Das Gollapalli, Prasenjit Mitra, and C Lee Giles. "Ranking authors in digital libraries." In: *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*. 2011, pages 251–254.

[102] Bruno Gonçalves and Nicola Perra. *Social phenomena: From data analysis to models*. Springer, 2015.

[103] Michael Gorman. "Google and God's mind." In: *Los Angeles Times* 17 (2004).

[104] Swati Goswami, CA Murthy, and Asit K Das. "Sparsity measure of a network graph: Gini index." In: *Information Sciences* 462 (2018), pages 16–39.

[105] Sander Greenland, Mohammad Ali Mansournia, and Douglas G Altman. "Sparse data bias: a problem hiding in plain sight." In: *bmj* 352 (2016), page i1981.

[106] Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. "Wtf: The who to follow service at twitter." In: *Proceedings of the 22nd international conference on World Wide Web*. 2013, pages 505–514.

[107] Branka Hadji Misheva, Alessandro Spelta, and Paolo Giudici. "Network based scoring models to improve credit risk management in peer to peer lending platforms." In: *Frontiers in Artificial Intelligence* 2 (2019), page 3.

[108] Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky, Skye Bender-deMoll, and Martina Morris. *statnet: Software Tools for the Statistical Analysis of Network Data*. R package version 2016.4. The Statnet Project (`https://www.statnet.org`). 2016. URL: `https://CRAN.R-project.org/package=statnet`.

[109] Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. "statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data." In: *Journal of Statistical Software* 24.1 (2008), pages 1–11. URL: `https://www.jstatsoft.org/v24/i01`.

[110] Hurst Hannum. "The concept and definition of minorities." In: *Universal Minority Rights* (2007), page 49.

[111] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." In: *Advances in neural information processing systems*. 2016, pages 3315–3323.

[112] Taher H Haveliwala. "Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search." In: *IEEE transactions on knowledge and data engineering* 15.4 (2003), pages 784–796.

[113] Denis Helic, Markus Strohmaier, Michael Granitzer, and Reinhold Scherer. "Models of human navigation in information networks based on decentralized search." In: *Proceedings of the 24th ACM Conference on Hypertext and Social Media.* ACM. 2013, pages 89–98.

[114] Thomas T Hills, Michael N Jones, and Peter M Todd. "Optimal foraging in semantic memory." In: *Psychological review* 119.2 (2012), page 431.

[115] Michael Hind. "Explaining explainable AI." In: *XRDS: Crossroads, The ACM Magazine for Students* 25.3 (2019), pages 16–19.

[116] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. "Stochastic blockmodels: First steps." In: *Social networks* 5.2 (1983), pages 109–137.

[117] Paul W Holland and Samuel Leinhardt. "An exponential family of probability distributions for directed graphs." In: *Journal of the american Statistical association* 76.373 (1981), pages 33–50.

[118] Petter Holme and Beom Jun Kim. "Growing scale-free networks with tunable clustering." In: *Physical review E* 65.2 (2002), page 026107.

[119] Harald Holone. "The filter bubble and its effect on online personal health information." In: *Croatian medical journal* 57.3 (2016), page 298.

[120] Kaizhu Huang, Haiqin Yang, Irwin King, and Michael R Lyu. "Learning classifiers from imbalanced data based on biased minimax probability machine." In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* Volume 2. IEEE. 2004, pages II–II.

[121] Lingxiao Huang and Nisheeth Vishnoi. "Stable and Fair Classification." In: *International Conference on Machine Learning.* 2019, pages 2879–2890.

[122] Lawrence Hubert and James Schultz. "Quadratic assignment as a general data analysis strategy." In: *British journal of mathematical and statistical psychology* 29.2 (1976), pages 190–241.

[123] David John Hughes, Moss Rowe, Mark Batey, and Andrew Lee. "A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage." In: *Computers in Human Behavior* 28.2 (2012), pages 561–569.

[124] Mark Huisman. "Imputation of missing network data: some simple procedures." In: *Social Structure* 10.1 (2009), pages 1–29.

[125] Aída Hurtado. "Intersectional understandings of inequality." In: *The Oxford handbook of social psychology and social justice* (2018), pages 157–172.

[126] Lihi Idan and Joan Feigenbaum. “Show me your friends, and I will tell you whom you vote for: Predicting voting behavior in social networks.” In: *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE. 2019, pages 816–824.

[127] Abigail Z Jacobs and Hanna Wallach. “Measurement and fairness.” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pages 375–385.

[128] Glen Jeh and Jennifer Widom. “Scaling personalized web search.” In: *Proceedings of the 12th international conference on World Wide Web*. 2003, pages 271–279.

[129] David Jensen, Jennifer Neville, and Brian Gallagher. “Why collective inference improves relational classification.” In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2004, pages 593–598.

[130] Karel Jezek, Dalibor Fiala, and Josef Steinberger. “Exploration and Evaluation of Citation Networks.” In: *ELPUB*. 2008, pages 351–362.

[131] Shan Jiang, Joseph Ferreira, and Marta C González. “Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore.” In: *IEEE Transactions on Big Data* 3.2 (2017), pages 208–219.

[132] Nathan Kallus, Xiaojie Mao, and Angela Zhou. “Assessing algorithmic fairness with unobserved protected class using data combination.” In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, 2020, page 110.

[133] Fariba Karimi, Mathieu Génois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. “Homophily influences ranking of minorities in social networks.” In: *Scientific reports* 8 (2018). DOI: 10.1038/s41598-018-29405-7.

[134] Brian Karrer and Mark EJ Newman. “Stochastic blockmodels and community structure in networks.” In: *Physical Review E* 83.1 (2011), page 016107.

[135] Robert E Kass and Adrian E Raftery. “Bayes factors.” In: *Journal of the American Statistical Association* 90.430 (1995), pages 773–795.

[136] Matt Keeling. “The implications of network structure for epidemic dynamics.” In: *Theoretical population biology* 67.1 (2005), pages 1–8.

[137] William Ogilvy Kermack and Anderson G McKendrick. “A contribution to the mathematical theory of epidemics.” In: *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115.772 (1927), pages 700–721.

[138] Myunghwan Kim and Jure Leskovec. "Modeling Social Networks with Node Attributes using the Multiplicative Attribute Graph Model." In: *UAI 2011, Barcelona, Spain, July 14-17, 2011*. 2011, pages 400–409.

[139] Moses C Kiti, Michele Tizzoni, Timothy M Kinyanjui, Dorothy C Koech, Patrick K Munywoki, Milosch Meriac, Luca Cappa, André Panisson, Alain Barrat, Ciro Cattuto, et al. "Quantifying social contacts in a household setting of rural Kenya using wearable proximity sensors." In: *EPJ Data Science* 5.1 (2016), page 1.

[140] Jon Kleinberg and Manish Raghavan. "Selection problems in the presence of implicit bias." In: *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Volume 94. Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018, 33:1–33:17.

[141] Jon M Kleinberg. "Small-world phenomena and the dynamics of information." In: *Advances in neural information processing systems*. 2002, pages 431–438.

[142] Kaj-Kolja Kleineberg, Marián Boguñá, M ángeles Serrano, and Fragkiskos Papadopoulos. "Hidden geometric correlations in real multiplex networks." In: *Nature Physics* (2016).

[143] Hyunsik Kong, Samuel Martin-Gutierrez, and Fariba Karimi. "First-mover advantage explains gender disparities in physics citations." In: *arXiv preprint arXiv:2110.02815* (2021).

[144] Gueorgi Kossinets. "Effects of missing data in social networks." In: *Social Networks* 28 (2006), pages 247–268.

[145] David Krackhardt. "Predicting with networks: Nonparametric multiple regression analysis of dyadic data." In: *Social networks* 10.4 (1988), pages 359–381.

[146] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. "Adaptive sensitive reweighting to mitigate bias in fairness-aware classification." In: *Proceedings of the 2018 World Wide Web Conference*. 2018, pages 853–862.

[147] John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* Academic Press, 2014.

[148] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. "Counterfactual fairness." In: *Advances in neural information processing systems*. 2017, pages 4066–4076.

[149] Alexandra L'heureux, Katarina Grolinger, Hany F Elyamany, and Miriam AM Capretz. "Machine learning with big data: Challenges and approaches." In: *Ieee Access* 5 (2017), pages 7776–7797.

[150] Renaud Lambiotte and Michal Kosinski. "Tracking the digital footprints of personality." In: *Proceedings of the IEEE* 102.12 (2014), pages 1934–1939.

[151] Laura Larrimore, Li Jiang, Jeff Larrimore, David Markowitz, and Scott Gorski. "Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success." In: *Journal of Applied Communication Research* 39.1 (2011), pages 19–37.

[152] Ju-Sung Lee and Jürgen Pfeffer. "Estimating Centrality Statistics for Complete and Sampled Networks: Some Approaches and Complications." In: *48th Hawaii International Conference on System Sciences, HICSS 2015, Kauai, Hawaii, USA, January 5-8, 2015.* 2015, pages 1686–1695. DOI: 10.1109/HICSS.2015.203.

[153] Florian Lemmerich, Philipp Singer, Martin Becker, Lisette Espín-Noboa, Dimitar Dimitrov, Denis Helic, Andreas Hotho, and Markus Strohmaier. "Comparing hypotheses about sequential data: A bayesian approach and its applications." In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer. 2017, pages 354–357.

[154] Ronny Lempel and Shlomo Moran. "SALSA: the stochastic approach for link-structure analysis." In: *ACM Transactions on Information Systems (TOIS)* 19.2 (2001), pages 131–160.

[155] Jure Leskovec and Christos Faloutsos. "Sampling from large graphs." In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM. 2006, pages 631–636.

[156] Jhao-Yin Li and Mi-Yen Yeh. "On Sampling Type Distribution from Heterogeneous Social Networks." In: *Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part II.* PAKDD'11. Shenzhen, China: Springer-Verlag, 2011, pages 111–122. ISBN: 978-3-642-20846-1. URL: http://dl.acm.org/citation.cfm?id=2022850.2022860.

[157] Yanying Li, Yue Ning, Rong Liu, Ying Wu, and Wendy Hui Wang. "Fairness of Classification Using Users' Social Relationships in Online Peer-To-Peer Lending." In: *Companion Proceedings of the Web Conference 2020.* WWW '20. Taipei, Taiwan: Association for Computing Machinery, 2020, pages 733–742. ISBN: 9781450370240. DOI: 10.1145/3366424.3383557.

[158] David Liben-Nowell and Jon Kleinberg. "The link-prediction problem for social networks." In: *journal of the Association for Information Science and Technology* 58.7 (2007), pages 1019–1031.

[159] Frank Lin and William W Cohen. "Semi-supervised classification of network data using very few labels." In: *2010 International Conference on Advances in Social Networks Analysis and Mining.* IEEE. 2010, pages 192–199.

[160] Mingfeng Lin, Nagpurnanand R Prabhala, and Siva Viswanathan. “Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending.” In: *Management Science* 59.1 (2013), pages 17–35.

[161] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. Volume 793. John Wiley & Sons, 2019.

[162] Xiaoming Liu, Johan Bollen, Michael L Nelson, and Herbert Van de Sompel. “Co-authorship networks in the digital library research community.” In: *Information processing & management* 41.6 (2005), pages 1462–1480.

[163] Peter Lofgren, Siddhartha Banerjee, and Ashish Goel. “Personalized pagerank estimation and search: A bidirectional approach.” In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. 2016, pages 163–172.

[164] Kristian Lum and William Isaac. “To predict and serve?” In: *Significance* 13.5 (2016), pages 14–19.

[165] Sofus A Macskassy and Foster Provost. “Classification in networked data: A toolkit and a univariate case study.” In: *Journal of machine learning research* 8.5 (2007), pages 935–983. URL: https://dl.acm.org/doi/10.5555/1248659.1248693.

[166] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. “Feedback loop and bias amplification in recommender systems.” In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pages 2145–2148.

[167] Manuel Sebastian Mariani, Matúš Medo, and Yi-Cheng Zhang. “Ranking nodes in growing networks: When PageRank fails.” In: *Scientific reports* 5 (2015).

[168] Leandro Balby Marinho, Christine Preisach, Lars Schmidt-Thieme, et al. “Relational classification for personalized tag recommendation.” In: *ECML PKDD Discovery Challenge 2009 (DC09)* (2009), page 7.

[169] Eduardo Marques. “Urban poverty, segregation and social networks in São Paulo and Salvador, Brazil.” In: *International Journal of Urban and Regional Research* 39.6 (2015), pages 1067–1083.

[170] Travis Martin, Brian Ball, Brian Karrer, and MEJ Newman. “Coauthorship and citation patterns in the Physical Review.” In: *Physical Review E* 88.1 (2013), page 012814.

[171] Miller McPherson, Lynn Smith-Lovin, and James M Cook. “Birds of a Feather: Homophily in Social Networks.” In: *Annual Review of Sociology* 27.1 (2001), pages 415–444. DOI: 10.1146/annurev.soc.27.1.415.

[172] *MEDDRA*. Accessed: 2019-02-21. 2011. URL: `https://bioportal.bioontology.org/ontologies/MEDDRA`.

[173] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A survey on bias and fairness in machine learning." In: *ACM Computing Surveys (CSUR)* 54.6 (2021), pages 1–35.

[174] Aditya Krishna Menon and Robert C Williamson. "The cost of fairness in binary classification." In: *Conference on Fairness, Accountability and Transparency*. 2018, pages 107–118.

[175] Robert K Merton. "The Matthew effect in science: The reward and communication systems of science are considered." In: *Science* 159.3810 (1968), pages 56–63.

[176] Robert K Merton. "The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property." In: *isis* 79.4 (1988), pages 606–623.

[177] Stanley Milgram. "The small world problem." In: *Psychology today* 2.1 (1967), pages 60–67.

[178] Jacob L Moreno and Helen H Jennings. "Statistics of social configurations." In: *Sociometry* (1938), pages 342–374.

[179] Jacob Levy Moreno. "Who shall survive?: A new approach to the problem of human interrelations." In: (1934).

[180] Sebastian Moreno and Jennifer Neville. "Network hypothesis testing using mixed Kronecker product graph models." In: *Data Mining (ICDM)*. IEEE. 2013, pages 1163–1168.

[181] Flaviano Morone and Hernán A Makse. "Influence maximization in complex networks through optimal percolation." In: *Nature* 524.7563 (2015), page 65.

[182] Carlo Morselli. "Career opportunities and network-based privileges in the Cosa Nostra." In: *Crime, Law and Social Change* 39.4 (2003), pages 383–418.

[183] Adam Mosseri. *Bringing People Closer Together*. `https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together`. [Online; accessed 23-Oct-2020]. 2018.

[184] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. "Generative adversarial minority oversampling." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pages 1695–1704.

[185] Mark A Musen, Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Christopher G Chute, Margaret-Anne Story, Barry Smith, and NCBO team. "The national center for biomedical ontology." In: *Journal of the American Medical Informatics Association* 19.2 (2011), pages 190–195.

[186] *National Center for Biomedical Ontology (NCBO).* Accessed: 2019-02-21. 2011. URL: `https://www.bioontology.org/`.

[187] Jennifer Neville and David Jensen. "Iterative classification in relational data." In: *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data.* 2000, pages 13–20.

[188] Mark EJ Newman. "The structure of scientific collaboration networks." In: *Proceedings of the national academy of sciences* 98.2 (2001), pages 404–409.

[189] Mark EJ Newman. "The structure and function of complex networks." In: *SIAM review* 45.2 (2003), pages 167–256.

[190] Hoai-Tuong Nguyen. "Multiple hypothesis testing on edges of graph: a case study of Bayesian networks." In: (2012).

[191] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. "Mining user mobility features for next place prediction in location-based services." In: *2012 IEEE 12th international conference on data mining.* IEEE. 2012, pages 1038–1043.

[192] Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute, et al. "BioPortal: ontologies and integrated data resources at the click of a mouse." In: *Nucleic acids research* (2009), gkp440.

[193] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. *The PageRank citation ranking: Bringing order to the web.* Technical report. Stanford InfoLab, 1999.

[194] Gopal Pandurangan, Prabhakar Raghavan, and Eli Upfal. "Using pagerank to characterize web structure." In: *International computing and combinatorics conference.* Springer. 2002, pages 330–339.

[195] Fragkiskos Papadopoulos, Maksim Kitsak, M ángeles Serrano, Marián Boguná, and Dmitri Krioukov. "Popularity versus similarity in growing networks." In: *Nature* 489.7417 (2012), pages 537–540.

[196] Orestis Papakyriakopoulos, Juan Carlos Medina Serrano, and Simon Hegelich. "Political communication on social media: A tale of hyperactive users and bias in recommender systems." In: *Online Social Networks and Media* 15 (2020), page 100058.

[197] Eli Pariser. *The filter bubble: What the Internet is hiding from you.* Penguin UK, 2011.

[198] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pages 2825–2830.

[199] Leto Peel. "Graph-based semi-supervised learning for relational networks." In: *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM. 2017, pages 435–443. URL: `http://hdl.handle.net/2078.1/182929`.

[200] Leto Peel, Jean-Charles Delvenne, and Renaud Lambiotte. "Multiscale mixing patterns in networks." In: *Proceedings of the National Academy of Sciences* 115.16 (2018), pages 4057–4062.

[201] Nicola Perra, Bruno Gonçalves, Romualdo Pastor-Satorras, and Alessandro Vespignani. "Activity driven modeling of time varying networks." In: *Scientific reports* 2 (2012), page 469.

[202] Joseph J Pfeiffer III, Sebastian Moreno, Timothy La Fond, Jennifer Neville, and Brian Gallagher. "Attributed graph models: Modeling network structure with correlated attributes." In: *WWW*. ACM. 2014, pages 831–842.

[203] Peter Pirolli and Stuart Card. "Information foraging." In: *Psychological review* 106.4 (1999), page 643.

[204] Peter LT Pirolli and James E Pitkow. "Distributions of surfers' paths through the World Wide Web: Empirical characterizations." In: *World Wide Web* 2.1 (1999), pages 29–45.

[205] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. "Echo chambers on Facebook." In: *Available at SSRN 2795110* (2016).

[206] Sandra Crouse Quinn and Supriya Kumar. "Health inequalities and infectious disease epidemics: a challenge for global health security." In: *Biosecurity and bioterrorism: biodefense strategy, practice, and science* 12.5 (2014), pages 263–273.

[207] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2016. URL: `https://www.R-project.org/`.

[208] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. "Mitigating bias in algorithmic hiring: Evaluating claims and practices." In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pages 469–481.

[209] Manoel Ribeiro. *Hateful Users on Twitter: Detecting hate speech with context*. `https://www.kaggle.com/manoelribeiro/hateful-users-on-twitter`. Accessed: 2020-10-07. 2018.

[210] Ronald E Robertson, David Lazer, and Christo Wilson. "Auditing the personalization and composition of politically-related search engine results pages." In: *Proceedings of the 2018 World Wide Web Conference*. 2018, pages 955–965.

[211] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. "An introduction to exponential random graph (p*) models for social networks." In: *Social networks* 29.2 (2007), pages 173–191.

[212] Luis E. C. Rocha, Fredrik Liljeros, and Petter Holme. "Information dynamics shape the sexual networks of Internet-mediated prostitution." In: *Proceedings of the National Academy of Sciences* 107.13 (2010), pages 5706–5711.

[213] Simon Rodan and Charles Galunic. "More than network structure: How knowledge heterogeneity influences managerial performance and innovativeness." In: *Strategic management journal* 25.6 (2004), pages 541–562.

[214] Everett M Rogers and Dilip K Bhowmik. "Homophily-heterophily: Relational concepts for communication research." In: *Public opinion quarterly* 34.4 (1970), pages 523–538.

[215] Cristòfol Rovira, Lluís Codina, Frederic Guerrero-Solé, and Carlos Lopezosa. "Ranking by relevance and citation counts, a comparative study: Google Scholar, Microsoft Academic, WoS and Scopus." In: *Future Internet* 11.9 (2019), page 202.

[216] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. "Multi-scale attributed node embedding." In: *Journal of Complex Networks* 9.2 (2021), cnab014.

[217] Armin Sajadi. *Fast Personalized PageRank Implementation.* `https://github.com/asajadi/fast-pagerank`. Accessed: 2021-03-30. 2019.

[218] Samuel F Sampson. *A novitiate in a period of change: An experimental and case study of social relationships.* Cornell University, 1968.

[219] Stephanie van de Sandt, Artemis Lavasa, Sünje Dallmeier-Tiessen, and Vivien Petras. "submitter: The Definition of Reuse." In: *Data Sci. J.* 18 (2019), page 22.

[220] Manojit Sarkar and Marc H Brown. "Graphical fisheye views of graphs." In: *Proceedings of the SIGCHI conference on Human factors in computing systems.* 1992, pages 83–91.

[221] Michael Scharkow, Frank Mangold, Sebastian Stier, and Johannes Breuer. "How social network sites and other online intermediaries increase exposure to news." In: *Proceedings of the National Academy of Sciences* 117.6 (2020), pages 2761–2763.

[222] *Wiederholende Forschung in den digitalen Geisteswissenschaften.* Zenodo, Feb. 2017. DOI: `10.5281/zenodo.277113`.

[223] Gideon Schwarz et al. "Estimating the dimension of a model." In: *The annals of statistics* 6.2 (1978), pages 461–464.

[224] Loren Schwiebert, Sandeep KS Gupta, and Jennifer Weinmann. "Research challenges in wireless networks of biomedical sensors." In: *Proceedings of the 7th annual international conference on Mobile computing and networking*. ACM. 2001, pages 151–165.

[225] Scikit-learn. *sklearn.ensemble.RandomForestRegressor*. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html`. Accessed: 2020-11-09.

[226] John Scott. "Social network analysis." In: *Sociology* 22.1 (1988), pages 109–127.

[227] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. "Collective classification in network data." In: *AI magazine* 29.3 (2008), pages 93–106. DOI: `10.1609/aimag.v29i3.2157`.

[228] Upul Senanayake, Mahendra Piraveenan, and Albert Zomaya. "The pagerank-index: Going beyond citation counts in quantifying scientific impact of researchers." In: *PloS one* 10.8 (2015), e0134794.

[229] Young-Duk Seo, Young-Gab Kim, Euijong Lee, and Doo-Kwon Baik. "Personalized recommender system based on friendship strength in social network services." In: *Expert Systems with Applications* 69 (2017), pages 135–148.

[230] Kirti R. Shah and Bikas K. Sinha. "Mixed Effects Models." In: *Theory of Optimal Designs*. Springer New York, 1989, pages 85–96.

[231] Tom Simonite. "When it comes to gorillas, google photos remains blind." In: *Wired, January* 11 (2018). [Online; accessed 01-June-2021].

[232] Philipp Singer, Denis Helic, Andreas Hotho, and Markus Strohmaier. "HypTrails: A Bayesian Approach for Comparing Hypotheses About Human Trails on the Web." In: *Proceedings of the 24th International Conference on World Wide Web*. WWW. Florence, Italy: ACM, 2015, pages 1003–1013. ISBN: 978-1-4503-3469-3.

[233] Philipp Singer, Denis Helic, Behnam Taraghi, and Markus Strohmaier. "Detecting memory and structure in human navigation patterns using markov chain models of varying order." In: *PloS one* 9.7 (2014), e102070.

[234] Ashudeep Singh and Thorsten Joachims. "Fairness of exposure in rankings." In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM. 2018, pages 2219–2228.

[235] Daniel Smilkov, Cesar A Hidalgo, and Ljupco Kocarev. "Beyond network structure: How heterogeneous susceptibility modulates the spread of epidemics." In: *Scientific reports* 4.1 (2014), pages 1–7.

[236] MG Smith. "Some problems with minority concepts and a solution." In: *Ethnic and Racial Studies* 10.4 (1987), pages 341–362.

[237] *SNAP Higgs Twitter datasets.* `https://snap.stanford.edu/data/higgs-twitter.html`. Accessed: 2016-08-15.

[238] Tom Snijders, Marinus Spreen, and Ronald Zwaagstra. "The use of multilevel modeling for analysing personal networks: Networks of cocaine users in an urban area." In: *Journal of quantitative anthropology* 5.2 (1995), pages 85–105.

[239] Tom AB Snijders. "Statistical models for social networks." In: *Review of Sociology* 37 (2011), pages 131–153.

[240] *Sociopatterns.* `http://www.sociopatterns.org/datasets/kenyan-households-contact-network/`. Accessed: 2016-08-26.

[241] Clemens Stachl, Quay Au, Ramona Schoedel, Samuel D Gosling, Gabriella M Harari, Daniel Buschek, Sarah Theres Völkel, Tobias Schuwerk, Michelle Oldemeier, Theresa Ullmann, et al. "Predicting personality from patterns of behavior collected with smartphones." In: *Proceedings of the National Academy of Sciences* 117.30 (2020), pages 17680–17687.

[242] StatsDirect. *Gini Coefficient of Inequality.* `https://www.statsdirect.com/help/default.htm#nonparametric_methods/gini.htm`. Accessed: 2020-11-09.

[243] Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. "Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity." In: *Proceedings of the 2018 World Wide Web Conference.* 2018, pages 923–932.

[244] Jessica Su, Aneesh Sharma, and Sharad Goel. "The effect of recommendations on network structure." In: *Proceedings of the 25th international conference on World Wide Web.* 2016, pages 1157–1167.

[245] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. "Yago: a core of semantic knowledge." In: *Proceedings of the 16th international conference on World Wide Web.* ACM. 2007, pages 697–706.

[246] Harini Suresh and John V Guttag. "A framework for understanding unintended consequences of machine learning." In: *arXiv preprint arXiv:1901.10002* (2019).

[247] Troy Tassier and Filippo Menczer. "Social network structure, segregation, and equality in a labor market with referral hiring." In: *Journal of Economic Behavior & Organization* 66.3-4 (2008), pages 514–528.

[248] Gergő Tóth, Johannes Wachs, Riccardo Di Clemente, ákos Jakobi, Bence Ságvári, János Kertész, and Balázs Lengyel. "Inequality is rising where social network segregation interacts with urban topology." In: *Nature communications* 12.1 (2021), pages 1–9.

[249] Amanda L Traud, Peter J Mucha, and Mason A Porter. "Social structure of Facebook networks." In: *Physica A: Statistical Mechanics and its Applications* 391.16 (2012), pages 4165–4180.

[250] Stephen Tu. "The dirichlet-multinomial and dirichlet-categorical models for bayesian inference." In: *Computer Science Division, UC Berkeley* (2014).

[251] Tania Tudorache, Jennifer Vendetti, and Natalya Fridman Noy. "Web-Protege: A Lightweight OWL Ontology Editor for the Web." In: *OWLED*. Volume 432. 2008.

[252] Frank Van Ham and Adam Perer. ""Search, show context, expand on demand": supporting large graph exploration with degree-of-interest." In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009).

[253] Sahil Verma and Julia Rubin. "Fairness definitions explained." In: *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE. 2018, pages 1–7.

[254] Arun Vijayshankar and Sandip Roy. "Cost of fairness in disease spread control." In: *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. IEEE. 2012, pages 4930–4935.

[255] Ivan Voitalov, Pim van der Hoorn, Remco van der Hofstad, and Dmitri Krioukov. "Scale-free networks well done." In: *Physical Review Research* 1.3 (2019), page 033034.

[256] Claudia Wagner. *Politicians on Wikipedia and DBpedia (Version: 1.0.0)*. GESIS - Leibniz-Institute for the Social Sciences, 2017. DOI: 10.7802/1515.

[257] Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. "Women through the glass ceiling: gender asymmetries in Wikipedia." In: *EPJ Data Science* 5.5 (2016). DOI: 10.1140/epjds/s13688-016-0066-4.

[258] Simon Walk, Lisette Esín-Noboa, Denis Helic, Markus Strohmaier, and Mark A Musen. "How users explore ontologies on the Web: A study of NCBO's BioPortal usage logs." In: *Proceedings of the 26th International Conference on World Wide Web*. 2017, pages 775–784. ISBN: 9781450349130. DOI: 10.1145/3038912.3052606.

[259] Simon Walk, Philipp Singer, Lisette Espín-Noboa, Tania Tudorache, Mark A Musen, and Markus Strohmaier. "Understanding how users edit ontologies: Comparing hypotheses about four real-world projects." In: *International Semantic Web Conference*. Springer. 2015, pages 551–568.

[260] Dan J Wang, Xiaolin Shi, Daniel A McFarland, and Jure Leskovec. "Measurement error in network data: A re-classification." In: *Social Networks* 34.4 (2012), pages 396–409.

[261] Stanley Wasserman, Katherine Faust, et al. “Social network analysis: Methods and applications.” In: (1994).

[262] Duncan J. Watts. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, 1999. URL: `http://www.jstor.org/stable/j.ctv36zr5d`.

[263] Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark A Musen. “BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications.” In: *Nucleic acids research* 39.suppl_2 (2011), W541–W545.

[264] Bodo Winter. “Linear models and linear mixed effects models in R with linguistic applications.” In: *arXiv:1308.5499* (2013).

[265] L Wirth. “The problem of minority groups (pp. 347–72).” In: *Indianapolis, IN: Bobbs-Merrill* (1945).

[266] Rongjing Xiang, Jennifer Neville, and Monica Rogati. “Modeling relationship strength in online social networks.” In: *WWW*. ACM. 2010, pages 981–990.

[267] Wenpu Xing and Ali Ghorbani. “Weighted pagerank algorithm.” In: *Proceedings of the Second Annual Conference on Communication Networks and Services Research, 2004*. IEEE. 2004, pages 305–314.

[268] Jiasen Yang, Bruno Ribeiro, and Jennifer Neville. “Should We Be Confident in Peer Effects Estimated From Social Network Crawls?” In: *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*. 2017, pages 708–711. URL: `https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15696`.

[269] Ke Yang and Julia Stoyanovich. “Measuring fairness in ranked outputs.” In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM. 2017, page 22.

[270] Yan Yu and Xinxin Wang. “Link prediction in directed network and its application in microblog.” In: *Mathematical Problems in Engineering* 2014 (2014).

[271] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. “Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment.” In: *Proceedings of the 26th international conference on world wide web*. 2017, pages 1171–1180.

[272] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. “Fairness constraints: Mechanisms for fair classification.” In: *Artificial Intelligence and Statistics*. 2017, pages 962–970.

[273] Meike Zehlike, Ke Yang, and Julia Stoyanovich. “Fairness in Ranking: A Survey.” In: *arXiv preprint arXiv:2103.14000* (2021).

[274] Giselle Zeno and Jennifer Neville. “Investigating the impact of graph structure and attribute correlation on collective classification performance.” In: (2016). 12th International Workshop on Mining and Learning with Graphs, KDD 2016.

[275] Jian-Xiong Zhang, Duan-Bing Chen, Qiang Dong, and Zhi-Dan Zhao. “Identifying a set of influential spreaders in complex networks.” In: *Scientific reports* 6 (2016), page 27823.

[276] Yue Zhang and Arti Ramesh. “Learning Fairness-aware Relational Structures.” In: *24th European Conference on Artificial Intelligence - ECAI 2020*. Volume 325. IOS Press. 2020, pages 2543–2550.

[277] Tong Zhao, Julian McAuley, and Irwin King. “Leveraging social connections to improve personalized ranking for collaborative filtering.” In: *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*. 2014, pages 261–270.

[278] Elena Zheleva and Lise Getoor. “To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles.” In: *Proceedings of the 18th international conference on World wide web*. 2009, pages 531–540.

[279] Kai Zhu and Lei Ying. “Information source detection in the SIR model: A sample-path-based approach.” In: *IEEE/ACM Transactions on Networking* 24.1 (2014), pages 408–421.

[280] James Zou and Londa Schiebinger. *AI can be sexist and racist—it’s time to make it fair*. 2018. DOI: 10.1038/d41586-018-05707-8.