Thesis Proposal: Structural Analysis and Application of Antagonistic Interactions in Online Social Networks.

Emma Fraxanet[1], Supervisor: Vicenç Gómez[1], and Co-supervisor: David Garcia[2,3]

[1]Pompeu Fabra University
[2]University of Konstanz
[3]Complexity Science Hub (Vienna)

September 12, 2022

Contents

1 Introduction

1.1 Computational Social Sciences and Digital Traces

The presence of online social environments has had an impact on interpersonal connectivity and information access, and while it has clearly increased both it also changed the nature in which individuals adapt them into their daily life. At the same time, changes in individual's behaviour aggregate and give birth to emergent phenomena that shape a new paradigm of social organization and norms. Despite the creation of these platforms opening a great assortment of beneficial social advances, growing polarization, conflict, and hostility online have been increasingly prevalent in the past decade [1] [2] [3]. A substantial amount of work has been and is dedicated to understanding how new online connectivity configurations are affecting human interaction.

The enormous bulk of digital trace data comprised by these interactions takes on a great variety of forms: from the agreeability of a "like" click in Facebook to a long stance defining paragraph in Reddit, and even GPS location history. This data is already used in add recommendations, optimization of commuting routes and many other data analysis pipelines managed by companies and institutions, who understand the value of this data not only for its spontaneity but also because the there's power in numbers: aggregation of individual data can reveal behavioural regularities of the society seen as a whole system. Understanding these patterns is not only relevant for the goal of predicting, which goes first in line for what concerns business interests, but can also provide incredibly insightful knowledge to social sciences concerns. For example, in [4], they study how network structures grown from *Homophily*—our tendency to connect with those similar to us— and *Preferential attachment* (PA)—our tendency to connect with already popular individuals— can harm minorities' connectivity with the majority and consequently their overall visibility in the network. The models and theories developed around the use of this data for such purposes are framed around the field of Computational Social Sciences. Even if the availability of data and computational power for these techniques and methodologies has been around in the last decades, the field feeds on substantially interdisciplinary and established sources (e.g. social psychology and statistical physics).

1.2 Networks and Signed Networks

Expanding the example of statistical physics, to model the behaviour of groups of particles and find emergence of collective phenomena, it is necessary to add the interaction between particles, often in the shape of pair-wise interactions. When modelling online communities, interactions (such as following each other, re-tweets or likes) can be thought of as one of these pair-wise interactions between two users. These constellations of interactions fall naturally in the competence of Network Science frameworks, being this field an optimal tool for the study of large social systems. From the description of local interactions, we achieve to explain macroscopic properties of the system (e.g. social networks resilience [5]) and unveil relevant patterns in the data (e.g. community detection [6]) . To all of these, additional layers of temporal information [7] or multi modal interactions [8] can be considered, given the amount of methodological and theoretical work done on these new types of network structures recently.

Signed Networks

A sub-set in the field of Network Science is the use of Signed Networks, in which each node represents an individual in a community and interactions are bimodal and generate positive (+) or negative (-) edges.
In social media, such interaction attributes can be based on liking, praise, friendships or trust for positive edges and disliking, toxicity, enmity or distrust for negative edges. Even though these parameters follow continuous distributions, and one can have a neutral stance towards a neighbour user, there are proper methodologies to extract statistically significant edge information, and the consideration of the system particularities is of most relevance when designing the extraction of such information. For example, in small systems where users recognize each other, the explicit decision of ignoring their neighbour can be sufficiently significant to generate a negative edge, while in larger systems it can easily mean the two users simply did not happen to find each other in the platform. There exists an established and known amount of datasets obtained from mining signed graphs out of online social systems [9][10][11][12][13][14][15].

Structural Balance

Signed Network analysis is strongly linked to social psychology theories that were tested in small scale, real life social systems. Balance theory supports a human tendency to fall into specific balanced local configurations regarding the sign of the ties (i.e. the enemy of my friend is my enemy [16]) and it has been studied largely from different approaches [17][18][19][20][21]. Structural balance is a macro-scale property derived from a balance theory generalization to large scale networks. In the past decade there have been efforts in defining partial balance, which denotes how far is the network from being completely balanced [19] [22][21].

Other properties

Besides this formalization of previous social psychology theories, the proposal of using antagonistic interactions as well serves as an extension to a large amount of work done on only positive interactions. For example, previous research failed to fill the gap in analysing signed and dynamic interactions in online communities simultaneously [23], mostly due to data unavailability. This also indicates there is a plethora of methodologies that, even if completely established and commonly used on positive networks, needs to be assessed and adapted to other structures such as signed networks. Moreover, the use of these methodologies on systems that contain new information can lead to relevant, non-trivial outcomes. For example, while the function and importance of weak links in certain social systems [24] is known for positive networks, it is not so clear which roles nodes take in bridging communities detected in signed networks. Other elements to pay attention to could be: the proper definition of null models, the overview of community detection methodologies or the extraction of latent space models.

1.3 Study of Polarization with Networks

One of the online (and offline) phenomena that is in the focus of attention of academics from several disciplines currently is polarization. Polarization is a property of social systems in which different groups of people hold opposing views on a specific matter with an antagonistic nature. While the population can be polarized in one given topic, the correlation of those extreme opinions (i.e. issue alignment) in an aggregated bulk of topics is central to obtain a global polarized structure [25], in which, for example, the stance of a user regarding abortion will be highly correlated with their stance on gun control. On the other side, while the aforementioned phenomena would be labelled as ideological polarization, the addition of in-group positive feelings and out-group hate or dislike is defined as affective polarization.

Polarization is visible in real life recent contexts (either in the shape of demonstrations and social unrest and conflict or election results), but some growing effects can be easily traced back to the online realm, especially when putting it side to side with click-bait news, social media feed design and anonymous extremist forums. These increasing cases have important social consequences, given that it directly influences public debates, violence and formation of governments.

Polarization study with networks

The study of such systems is strongly based on detecting and analysing opinion or attitude distributions with respect to a given topic. This can be done traditionally (with surveys or polls), but is increasingly done with automated tools and social media data, such as using Sentiment analysis (Supervised Machine Learning Classification models) on Tweets [26][27]. These types of methods require proper data labelling and are dependent on language.

Other approaches, instead, are based on network structures only (such as following or retweet networks), in which the goal is to locate the position of users in a Latent Ideology Space to retrieve the ideology distribution [28] [3] [29] [10].

Polarization study with signed networks

The description of polarization, and more specifically affective polarization, organically calls for the addition of antagonistic interaction. Together with all the other approaches, the use of signed networks can provide a new radiography of polarization for certain communities. Even though some high-quality work has been done in the intersection of signed networks and polarization [19] [30] [31] [10] [32], there is no settled pipeline or methodology to obtain clear polarization measures or indicators from signed networks.

Elite minorities and polarization

Finally, another point of interest in this thesis is the presence of elite minorities in the communities and their relevance in global phenomena (such as polarization). As hinted in [32], sub-communities of a system can be the drivers of large-scale polarized structures and can be responsible for the dissemination of extreme ideologies. This could be an apparent effect even if there exists a moderate majority in online environments, that is somehow underrepresented in the opinion distributions obtained as the outcomes of these methods. In any normally distributed opinion there exists two tails of small strongly conflicting minorities. If these minorities gain control of the discourse, we could be talking of a "spiral of noise" or "elite capture" more than a "global polarization". In which the latter does not mean it is not an existing problem but that it might be a problem rooted in or exacerbated by such mechanisms.

1.4 Summary: the intersection. Motivation and challenges.

In summary, there exists a gap in the assortment of points of views from which we can describe how opinions spread and evolve online. Our social circles and dynamics are clearly defined by those who we see opposite to us as well as our friends, and differentiation processes are a key aspect in conflict resolution, but can also be a source of extremism or violence when done in unfortunate environments. Therefore, the addition of antagonistic interactions to the study of social systems is of most relevance, and needs to be done under robust and carefully designed methodologies.

2 Research objectives

2.1 Thesis goals

The main goal of this Thesis will be to understand and describe the differences perceived in our knowledge, methodology and conclusions related to online social systems analysis when considering antagonistic structures as well as the prevailing positive network analysis, with a special focus in its effect for phenomena related to opinion dynamics, group identity and polarization. The consideration of antagonistic structures is flexible and can be obtained from different approaches. For example, the addition of negative edges in positive networks or the design of competing minorities in network creation models tuned with homophily parameters.

To achieve this global goal, we set the following specific objectives with the corresponding project information:

1. Develop models based on group and opinion dynamics that can explain phenomena such as unequal representation in opinion distributions in online platforms. *(Publication 2.1, Publication 3.1)*

2. Improve our tools and unify the framework to describe the addition of negative interactions in networks of online social interactions, with a special focus on polarization

   - Develop methodology for the mining of signed networks and their polarization study (based on balance theory) from social media environments. *(Publication 1.1)*
   - Combine temporal and multilayer structures with the aforementioned methodology in a robust manner. *(Possible extension)*
   - Characterize the needed differences in methodology and approach between denser layers of the networks and periphery. *(Publication 1.2)*
   - Adapt Latent Ideology Space models to Signed Networks. *(Possible extension)*

3. Application of this methodology to available data

   - Apply the designed methodologies to a variety of datasets and extract regular behavioural patterns to characterize online user behaviour. *(Publication 1.1)*
   - Clean and prepare suitable datasets for the analysis of signed networks. *(Publication 1.1)*

2.2 Thesis Plan: Projects

- **Project 1: Signed Networks and Polarization**
  - **Keywords:** Signed networks, Partitioning, Structural Balance, Polarization, Latent Ideology Space, Multilayer networks, Temporal Network Analysis
  - **Type of research work:** Data driven
- **Project 2: Opinion Dynamics, homophily and competing minorities**
  - **Keywords:** Generative Network Model, Opinion Dynamics Model, Homophily, Preferential Attachment, Social media, Affective Polarization, Elite Capture
  - **Type of research work:** Theoretical modelling
- **Project 3: Polarization and Rakings model**
  - **Keywords:** Ranking Algorithm, Polarization, Opinion Dynamics, Misinformation, Experimental Design
  - **Type of research work:** Experimental

3 Research plan and projects

3.1 Project 1

This project serves as the main line of research of the Thesis, since it brings together all concepts mentioned above: Social media platforms, Signed networks and Polarization.

Currently, the project is at an intermediate stage at which the state-of-the-art methodologies and the data availability have been already explored. The first tangible outcome of this project is the preprint draft for the first publication (see Publication 1 below for more details). After this sub-project, we plan to extend and generalize the used methodology to obtain a framework easily adaptable to different data sources. The next step will be, therefore, to mature the robustness of the methods and enhance the analysis with the addition of new features in the analysis pipeline. The use of additional available datasets is crucial for the success of these next steps.

3.1.1 Publication 1.1

Collaborators: Max Pellert, Vicenç Gómez, Simon Schweighofer, David Garcia.
Advice and feedback: Samin Aref

Status: 70 % complete (Preprint draft in final stages). Preliminary results accepted as a talk in several conferences (IC2S2, NetSci and SunBelt). Also presented as a poster at WebSci. From the reviewer feedback of NetSci some extensions are considered as a contingency plan: considering the use of different null models (see Publication 1.2) and the application of the methodology to different datasets (see Possible extensions).

Abstract:

Online media are widely held responsible for the rise of political polarization throughout the Western world. But popular narratives of 'filter bubbles' and 'echo chambers' have recently been heavily criticized, because users clearly do communicate with political opponents online [28]. So how do users of online media polarize? We approach this question using a novel dataset, derived from the discussion forums of a major German-speaking news platform. This dataset contains over 94,000 users and 46 million interactions between them. Crucially, the interactions comprise up-votes as well as down-votes, and can thus be represented as edges with both signed and temporal information. This unique combination of features allow us to investigate the formation of political alliances and the emergence of political conflict in real time. We focus on debates surrounding the highly contentious European refugee crisis (2015-16), a notoriously turbulent year regarding corruption scandals which led to the Austrian government collapsing (2019) and the months comprising the start of the COVID-19 pandemic (2020).

We theorize political polarization as an increase of structural balance between actors with opposing ideological positions (see [33]). We test this model by quantifying the trajectory of structural polarization in the network of signed interactions as a longitudinal user analysis across years (similarly done in [20]). We do so by finding an optimal bi-partition of the signed network and defining a normalized upper bound of balance based on the amount of frustrated edges in that partition (i.e. edges that violate the assumptions of our partition model) [21].

Our results are congruent with the political developments over the same period: Overall, the level of structural polarization is increasing, and moments of acute political crisis and conflict coincide with peaks in structural polarization. Following a start of the migration crisis where humanitarian help and unity prevailed, several controversial government decisions increased the social tensions from the start of 2016. This led to abrupt alternations of power by very different political parties in the small interval of five years (2016-2020). We examine the context of the peaks in polarization through a detailed analysis of the social and political circumstances of each time period provided by the news articles texts. Moreover, we zoom in the mechanisms from which polarization increases in moments of crisis and find that, while divisiveness between partitions stays constant and high throughout the time period, cohesiveness within groups is what drives the changes in our polarization score.

More details in appendix A.

3.1.2 Publication 1.2

Collaborators: Max Pellert, Vicenç Gómez, Simon Schweighofer, David Garcia.
Advice and feedback: Samin Aref

Status: 30 % complete.

Description:
When designing the appropriate normalization factor for our Signed Polarization Score (see Appendix A), we encountered differences in the balance encoded in the underlying unsigned network depending on the density level: deeper, more connected, layers of the networks display larger baselines of polarization procured by the unsigned structure. We believe this can be related to the fact that users can recognize each other in inner layers and therefore positive interactions already encode most of the structural information.

These findings provide a strong point for differentiating between network structures when performing these types of analysis. Moreover, it shows a relevant pattern of human behaviour that is exhibited in real life systems. For this reason, we believe there is value in approaching these results more formally, and compare it across platforms. The collection of results from these exploration would be gathered in Publication 2 from Project 1.

3.1.3 Possible extensions

As mentioned above, the following steps of this project that serve as possible extensions to *Publication 1.1*, are:

- Methodology extension: Use of *Stochastic Degree Sequence Model* (SDSM) [34] for the mining of signed networks instead of the use of arbitrary thresholds.
- Methodology extension: use of temporal information dynamically. Use of rolling window measures instead of fixed arbitrary time windows.
- Methodology extension: Adaptation of *Latent Ideology Space Models* [28] [29] [35] to signed networks.
- Data extensions: application of the methodology to some of the datasets below (in order of relevance).
  - Inferring signed networks from interactions on user-generated content in Wikipedia, similar to [11],[12]. Can also compare between different languages.

- Debagreement [9] : 42,894 comment-reply pairs from the popular discussion website Reddit, annotated with agree, neutral or disagree labels.
- BirdwatchSG [10]: signed edge-attributed, multi-edge, directed graph with $441,896$ edges between $2,987$ Birdwatch participants (nodes) based in the USA, and spanning 1,020 diverse topics prone to misleading content and/or partisanship.
- TwitterSG [10]: signed edge-attributed, multi-edge, directed graph with $12,848,093$ edges between $753,944$ Twitter users (nodes), spanning 200 sports-related topics: teams, sports, players, managers, and events.
- Epinions [13]: Dataset based on the $841,372$ ratings of helpfulness of reviews from $131,828$ users in a product rating website.
- Slashdot [14]: Dataset based on $516,575$ friend and foe links network between $77,357$ users of a technology news platform.
- BitcoinOTC and Bitcoin Alpha [15]: who-trusts-whom graphs of $3,783$-$5,881$ users with $24,186$-$35,592$ edges who trade using Bitcoin on online platforms, in which reputation is important due to anonymity.

These extensions are not strict parts of the plan but serve as plausible topics towards which to re-direct the main project if needed.

3.2 Project 2. Publication 2.1

Collaborators: Elise Koskelo, Adam Finnemann, Ben Genta, Rachel Freedman.

Advisors/mentors: Mirta Galesic, Henrik Olsson, Fariba Karimi, Tamara van Der Does, Maria del Río Chanona, Jonas Dalege

Status: 50 % complete. Literature Review, Design of methodology and Code preparation and testing done. Experiments pending. Objective: preprint draft by January 2023

This project was conceived during the Complexity-GAINs International Summer School, organized by the Complexity Science Hub (Vienna) and the Santa Fe Institute.

It is based on an extension of an already existing model, by [36], in which the creation of a social network can be tuned according to different connection preferences between different groups. In our case, the project is restrained with the case of one large majority and two minorities that oppose each other. The main goal is to assess the effect of these network structures on the distributions of opinions, that are associated in a flexible manner with group identity. For that, we design and propose a new opinion dynamics model that fits the project's needs.

More details in appendix C.

3.3 Project 3. Publication 3.1

Collaborators: Vicenç Gómez, Fabrizio Germano.

The third project will analyze polarization from a different perspective than the previous ones. In this case, we will design an experiment to gather data and try to provide empirical evidence of complex phenomena related to polarization as predicted by a mathematical model.

In particular, we will build on the results in [37]. In this work, a model of opinion dynamics is presented where a platform ranks incoming news items, while individuals sequentially access the platform to decide which news items to view and possibly highlight (e.g., like, share or retweet). Some key features of the model are that *(i)* the platform uses an algorithm based on popularity of the news items and personalization of certain key individual characteristics, *(ii)* the platform cares about profits, *(iii)* individuals are driven by behavioral traits such that they have some preference for confirmatory news as well as to read higher ranked news, *(iv)* they highlight news that is sufficiently close to their prior beliefs and the more so the more extreme their prior beliefs are [38].

The authors in [37] showed analytically and via simulations is that a profit-maximizing platform will tend to choose higher than efficient popularity and personalization weights, both of which result in polarization and misinformation with users reading and sharing higher amounts of extreme (and not so accurate) content and lower amounts of less extreme (more accurate) content.

We plan to extend the work by providing empirical evidence of the phenomena captured in the aforementioned (or a simplified version) model and simulations through a designed experiment involving human participants and a simplified platform. The first task will be to design an experiment. For that we will follow a similar methodology as done in a previous work of the thesis supervisor Vicenç Gómez [39].

3.4 Planification timeline

Regarding the time management of the projects, as mentioned above, project one serves as the basis of the Thesis and therefore is comprised of several sub-projects. These projects will be carried out during the full duration of the PhD (see 1). The other two projects have a limited duration and will be taken on sequentially.

In Figure 2 it is shown the activities and time organization of the first year. While some of the activities, like PhD seminars and participation in particular summer schools and conferences, belong singularly to the first year, other activities such as teaching duties and participation in the CSS Lab Austria reading group and seminars will continue throughout the four years.

Concerning location, in order to maintain contact with CSS Lab Austria and my co-supervisor David Garcia, and considering I already did a stay with his group at the Complexity Science Hubin Vienna, I plan to visit the group in University of Konstanz during the third year.

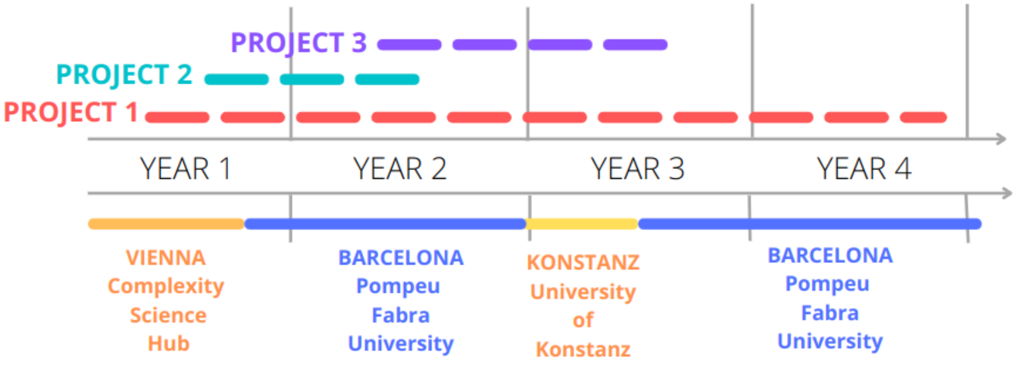
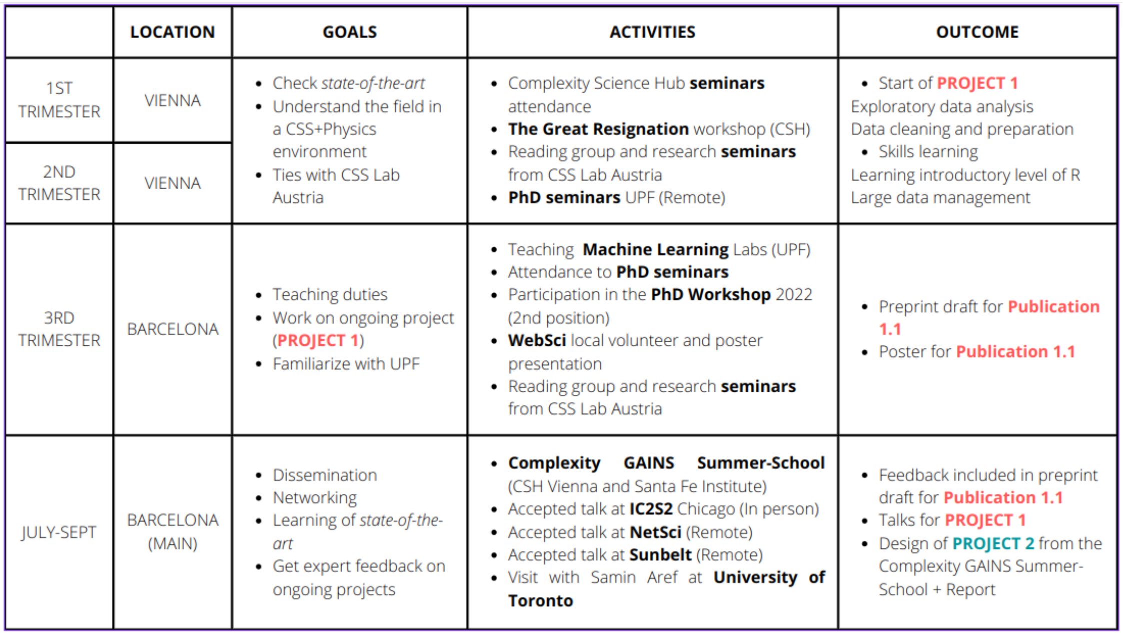

PROJECT 3
PROJECT 2
PROJECT 1
YEAR 1
YEAR 2
YEAR 3
YEAR 4
VIENNA Complexity Science Hub
BARCELONA Pompeu Fabra University
KONSTANZ University of Konstanz
BARCELONA Pompeu Fabra University

Figure 1: Project and location timeline



	LOCATION	GOALS	ACTIVITIES	OUTCOME
1ST TRIMESTER	VIENNA	- Check *state-of-the-art* - Understand the field in a CSS+Physics environment - Ties with CSS Lab Austria	- Complexity Science Hub **seminars** attendance - **The Great Resignation** workshop (CSH) - Reading group and research **seminars** from CSS Lab Austria - **PhD seminars** UPF (Remote)	- Start of **PROJECT 1** Exploratory data analysis Data cleaning and preparation - Skills learning Learning introductory level of R Large data management
2ND TRIMESTER	VIENNA			
3RD TRIMESTER	BARCELONA	- Teaching duties - Work on ongoing project (**PROJECT 1**) - Familiarize with UPF	- Teaching **Machine Learning** Labs (UPF) - Attendance to **PhD seminars** - Participation in the **PhD Workshop** 2022 (2nd position) - **WebSci** local volunteer and poster presentation - Reading group and research **seminars** from CSS Lab Austria	- Preprint draft for **Publication 1.1** - Poster for **Publication 1.1**
JULY-SEPT	BARCELONA (MAIN)	- Dissemination - Networking - Learning of *state-of-the-art* - Get expert feedback on ongoing projects	- **Complexity GAINS Summer-School** (CSH Vienna and Santa Fe Institute) - Accepted talk at **IC2S2** Chicago (In person) - Accepted talk at **NetSci** (Remote) - Accepted talk at **Sunbelt** (Remote) - Visit with Samin Aref at **University of Toronto**	- Feedback included in preprint draft for **Publication 1.1** - Talks for **PROJECT 1** - Design of **PROJECT 2** from the Complexity GAINS Summer-School + Report

Figure 2: First year goals, activities and outcome

References

[1] Vicky Chuqiao Yang et al. "Why Are U.S. Parties So Polarized? A "Satisficing" Dynamical Model". In: *SIAM Review* 62.3 (Jan. 2020), pp. 646–657. DOI: 10.1137/19m1254246. URL: https://doi.org/10.1137/19m1254246.

[2] Shanto Iyengar et al. "The Origins and Consequences of Affective Polarization in the United States". In: *Annual Review of Political Science* 22.1 (May 2019), pp. 129–146. DOI: 10.1146/annurev-polisci-051117-073034. URL: https://doi.org/10.1146/annurev-polisci-051117-073034.

[3] Isaac Waller and Ashton Anderson. "Quantifying social organization and political polarization in online platforms". In: *Nature* 600.7888 (2021), pp. 264–268.

[4] Fariba Karimi et al. "Homophily influences ranking of minorities in social networks". In: *Scientific reports* 8.1 (2018), pp. 1–12.

[5] David Garcia, Pavlin Mavrodiev, and Frank Schweitzer. "Social resilience in online communities: The autopsy of friendster". In: *Proceedings of the first ACM conference on Online social networks*. 2013, pp. 39–50.

[6] Symeon Papadopoulos et al. "Community detection in social media". In: *Data mining and knowledge discovery* 24.3 (2012), pp. 515–554.

[7] Petter Holme and Jari Saramäki. "Temporal networks". In: *Physics reports* 519.3 (2012), pp. 97–125.

[8] Mark E Dickison, Matteo Magnani, and Luca Rossi. *Multilayer social networks*. Cambridge University Press, 2016.

[9] John Pougué-Biyong et al. "DEBAGREEMENT: A comment-reply dataset for (dis) agreement detection in online debates". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021.

[10] John Pougué-Biyong et al. "Learning Stance Embeddings from Signed Social Graphs". In: *arXiv preprint arXiv:2201.11675* (2022).

[11] Robert West et al. "Exploiting social network structure for person-to-person sentiment analysis". In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 297–310.

[12] Silviu Maniu, Bogdan Cautis, and Talel Abdessalem. "Building a Signed Network from Interactions in Wikipedia". In: *Databases and Social Networks on - DBSocial '11*. Athens, Greece: ACM Press, 2011, pp. 19–24. ISBN: 978-1-4503-0650-8. DOI: 10.1145/1996413.1996417.

[13] Ramanthan Guha et al. "Propagation of trust and distrust". In: *Proceedings of the 13th international conference on World Wide Web*. 2004, pp. 403–412.

[14] Jérôme Kunegis, Andreas Lommatzsch, and Christian Bauckhage. "The slashdot zoo: mining a social network with negative edges". In: *Proceedings of the 18th international conference on World wide web*. 2009, pp. 741–750.

[15] Srijan Kumar et al. "Edge weight prediction in weighted signed networks". In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE. 2016, pp. 221–230.

[16] Fritz Heider. "Attitudes and cognitive organization". In: *The Journal of psychology* 21.1 (1946), pp. 107–112.

[17] Tuan M Pham et al. "Balance and fragmentation in societies with homophily and social balance". In: *Scientific reports* 11.1 (2021), pp. 1–7.

[18] Xiaolong Zheng, Daniel Zeng, and Fei-Yue Wang. "Social balance in signed networks". In: *Information Systems Frontiers* 17.5 (2015), pp. 1077–1095.

[19] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. "Signed networks in social media". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2010, pp. 1361–1370.

[20] Ernesto Estrada. "Rethinking structural balance in signed social networks". In: *Discrete Applied Mathematics* 268 (2019), pp. 70–90.

[21] Samin Aref and Mark C Wilson. “Measuring partial balance in signed networks”. In: *Journal of Complex Networks* 6.4 (2018), pp. 566–595.

[22] Ernesto Estrada and Michele Benzi. “Walk-based measure of balance in signed networks: Detecting lack of balance in social networks”. In: *Physical Review E* 90.4 (2014), p. 042802.

[23] Pedro Calais Guerra et al. “A measure of polarization on social media networks based on community boundaries”. In: *Seventh international AAAI conference on weblogs and social media.* 2013.

[24] Mark S Granovetter. “The strength of weak ties”. In: *American journal of sociology* 78.6 (1973), pp. 1360–1380.

[25] Simon Schweighofer, David Garcia, and Frank Schweitzer. “An agent-based model of multi-dimensional opinion dynamics and opinion alignment”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 30.9 (2020), p. 093139.

[26] Loris Belcastro et al. “Learning political polarization on social media using neural networks”. In: *IEEE Access* 8 (2020), pp. 47177–47187.

[27] Vorakit Vorakitphan et al. “Regrexit or not regrexit: Aspect-based sentiment analysis in polarized contexts”. In: *Proceedings of the 28th International Conference on Computational Linguistics.* 2020, pp. 219–224.

[28] Pablo Barberá et al. “Tweeting from left to right: Is online political communication more than an echo chamber?” In: *Psychological science* 26.10 (2015), pp. 1531–1542.

[29] Samuel Martın-Gutiérrez, Juan Carlos Losada, and Rosa M Benito. “Recurrent patterns of user behavior in different electoral campaigns: a twitter analysis of the Spanish general elections of 2015 and 2016”. In: *Complexity* 2018 (2018).

[30] Samin Aref et al. “Multilevel structural evaluation of signed directed social networks based on balance theory”. In: *Scientific reports* 10.1 (2020), pp. 1–12.

[31] Samin Aref and Zachary P Neal. “Identifying hidden coalitions in the US House of Representatives by optimally partitioning signed networks based on generalized balance”. In: *Scientific reports* 11.1 (2021), pp. 1–9.

[32] Francesco Bonchi et al. “Discovering polarized communities in signed networks”. In: *Proceedings of the 28th acm international conference on information and knowledge management.* 2019, pp. 961–970.

[33] Simon Schweighofer, Frank Schweitzer, and David Garcia. “A weighted balance model of opinion hyperpolarization”. In: *Journal of Artificial Societies and Social Simulation* 23.3 (2020), p. 5.

[34] Zachary Neal. “The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors”. In: *Social Networks* 39 (2014), pp. 84–97.

[35] Felix Gaisbauer et al. “Grounding force-directed network layouts with latent space models”. In: *arXiv preprint arXiv:2110.11772* (2021).

[36] Lisette Espın-Noboa et al. “Inequality and inequity in network-based ranking and recommendation algorithms”. In: *Scientific Reports* 12.1 (Feb. 2022). DOI: 10.1038/s41598-022-05434-1. URL: https://doi.org/10.1038/s41598-022-05434-1.

[37] Fabrizio Germano, Vicenç Gómez, and Francesco Sobbrio. “Crowding out the truth? A simple model of online misinformation, polarization and meaningful social interactions”. In: *International Conference on Computational Social Science.* 2022, p. 1.

[38] Eytan Bakshy, Solomon Messing, and Lada A Adamic. “Exposure to ideologically diverse news and opinion on Facebook”. In: *Science* 348.6239 (2015), pp. 1130–1132.

[39] Fabrizio Germano, Vicenç Gómez, and Gaël Le Mens. “The few-get-richer: a surprising consequence of popularity-based rankings”. In: *The World Wide Web Conference.* 2019, pp. 2764–2770.

[40] Dorwin Cartwright and Frank Harary. “Structural balance: a generalization of Heider’s theory.” In: *Psychological review* 63.5 (1956), p. 277.

[41] Samin Aref, Andrew J Mason, and Mark C Wilson. “A modeling and computational study of the frustration index in signed networks”. In: *Networks* 75.1 (2020), pp. 95–110.

[42] Patrick Doreian. “A multiple indicator approach to blockmodeling signed networks”. In: *Social Networks* 30.3 (2008), pp. 247–258.

[43] David Schoch. *signnet: An R package to analyze signed networks.* 2020. URL: https://github.com/schochastics/signnet.

[44] Xindi Wang, Onur Varol, and Tina Eliassi-Rad. “Information access equality on generative models of complex networks”. In: *Applied Network Science* 7.1 (Aug. 2022). DOI: 10.1007/s41109-022-00494-8. URL: https://doi.org/10.1007%2Fs41109-022-00494-8.

[45] Peter Turchin. “Modeling social pressures toward political instability”. In: *Cliodynamics* 4.2 (2013).

[46] Sandra González-Bailón et al. “The Dynamics of Protest Recruitment through an Online Network”. en. In: *Scientific Reports* 1.1 (Dec. 2011), p. 197. ISSN: 2045-2322. DOI: 10.1038/srep00197. URL: http://www.nature.com/articles/srep00197 (visited on 09/01/2022).

[47] Olúfmi O Táıwò. *Elite capture: How the powerful took over identity politics (and everything else).* Haymarket Books, 2022.

A Publication 1.1

A.1 Prior work, background and definitions

Given a signed graph that represents our network of positive and negative interactions, we can assess its structural balance as a binary condition: either it is balanced or not. When generalized to this framework, a signed network is balanced if it can be partitioned into $k \leq 2$ such that all negative edges fall outside the partitions and all positive edges fall within the partitions. Balance can also be defined by the absence of cycles containing an off number of negative edges [40]. For unbalanced graphs, a measure for partial balance, which denotes how far is the network from being completely balanced, can be obtained from different means: signed triangle count[19], walks [22] or frustration [21]. The latter is based on a *Frustration index*, that is obtained from the count of frustrated edges (i.e. edges that violate the assumptions of the optimal partition model).

Following [21] notation, we represent an undirected signed graph with $G = (V, E, \sigma)$, where V are the set of n nodes, E the set of m edges and σ is the sign function that maps the edges to their given sign $\sigma : E \to \{-1, +1\}$. Given a partition $P = \{X, V \setminus X\}$, the frustration count will be the sum of the frustration state of all edges, $f_G = \sum_{(i,j)} f_{ij}$, where f_{ij} equals 1 for frustrated edges and 0 for non-frustrated edges. The problem thus is stated as finding the optimal partition P^* such that the amount of frustrated edges is the minimum possible. This will be the correct description of partial balance of the network. The globally optimal solution then should satisfy $L(G) = \min_{X \subseteq V} f_G(X)$.

The computation of $L(G)$ is NP-hard given its relation to the EDGE-BIPARTIZATION unsigned graph optimization models and the MAXCUT graph optimization problem [41]. For small scale networks, however, it is possible to use an efficient way to compute the frustration index exactly [21]. This method is based on binary linear programming formulation and allows to make use of specific speed up techniques and powerful mathematical programming solvers.

For large scale networks there are several ways to approximate the solution of this optimization problem. For example, Doreian and Mvar apply Blockmodeling to this problem in [42], in which they optimize the criterion function $P(X) = E_{f,p} + E_{f,n}$ via a relocation algorithm, with $E_{f,p}$ defined as the frustrated positive edges and $E_{f,n}$ the frustrated negative edges. This method together with simulated annealing provides partitions that can define a good approximation to $L(G)$. We use the implementation of this model through the library *Signnet* in R [43]. Any approximated value for $L(G)$ will therefore be an upper bound on the minimum number of frustrated edges.

A.2 Methods (contributions)

The intuition is that, given a system defined by a signed network, the minimum amount of frustrated edges will hint the degree up to which this system can be easily separated into groups. Moreover, we understand this level of separation as direct measure of polarization. We approach the idea of polarization from a structural point of view, given that the network structure already encodes bimodal interactions, we can presume it to be intrinsically related to affective polarization.

The next steps are: 1. Providing a formal definition of polarization in this framework, and 2. Examining to what extent this measure is given by the sign distribution and not the unsigned backbone of our networks.

Since we are interested in quantifying polarization in our networks we look for an index that ranges from 0 to 1, such as $1 - \frac{L(G)}{m/2}$, with 1 being the completely balanced case. The $m/2$ term accounts for different network sizes. The relation between these different concepts in our framework, then, is that balance and polarization are equivalent and both grow in an inverse trend compared to frustration in the system. The more frustration there is, the more blended the groups are, and the less polarized the global community is.

We describe the *Signed Polarization Index as*:

$$SPI = 1 - \frac{L(G)}{m/2}$$

Secondly, to assess the variation of this index in the DerStandard networks we normalize the score by extracting the amount of polarization encoded in the unsigned structure. This is done by simply averaging over several measures of $1 - \frac{L(G)}{m/2}$ in the network with randomly re-distributed sign attributes ($G_{shuffled}$). This baseline is constant along different instances of networks for the Derstandard data source.

$$SPI_{global} = SPI(G) - \overline{SPI(G_{shuffled})}$$

A.3 Data

DerStandard is one of the largest newspapers in Austria. Its online community is highly engaged and the platform has not suffered any relevant collapse or large shift of users in the previous decade. For example, the site had almost 57 million visits in November 2020.

Our dataset focuses on the forum nature of the Derstandard site and includes the unique user identifiers, the text of each posting, the timestamp of the posting, the article id under which the posting was published, the thread status of the comment (original or reply) and the list of like and dislike voters for each comment. The combination of these features in a dataset is unique and novel: We can combine explicitly signed (positive and negative) interactions on a large-scale on (political) topics for a long time frame.

DerStandard has a stable and active user base thanks to human-driven moderation schemes to deescalate conflict and promote healthy and insightful discussions. This platform is also established and therefore does not have strong influxes of users or sudden large losses of users. For those reasons, we can take in consideration only users that voted at least once every year in our chosen time period. This allows us to find around 14K users that can be tracked during 7 years.

From this data, we construct signed networks that capture the status of the interaction between pairs of users at a given time window. To have sufficient data to build a signed network, we use a 3-month time window to convert votes data into signed networks based on normalized scores between -1 and 1. We remove a small fraction (less than 1%) of edges with perfect balance, i.e. a score of exactly zero. Since we study 7 years, we obtain 28 signed networks, one for each quarter, describe the relationships between the users in DerStandard.

For each of these networks, we analyze the largest connected component. Note that some users might not be active in a given time window, thus the size of the network changes over time. Networks in our analysis have an average of 7236 nodes ($sd = 612$) and an average number of edges of 506454 ($sd = 169833$), thus having an average density of $\rho = 0.02$. Since we correct for the positivity bias in votes, the proportion of positive edges is centered around an average of 0.51 with a standard deviation of 0.02.

The extensive time frame of our observation period covers events including the highly contentious European refugee crisis (2015-16), a notoriously turbulent year regarding corruption scandals which led to the the dissolution of the Austrian government coalition (2019) and the months comprising the start of the COVID-19 pandemic (2020). The inclusion of the year 2014 allows us to compare possibly turbulent years with more calm times as a baseline.

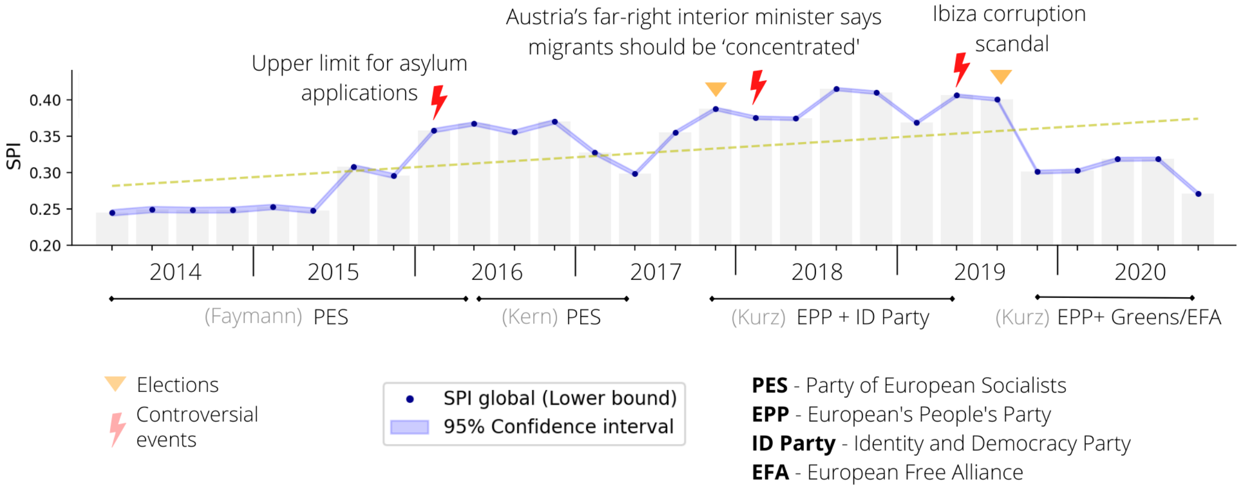


Figure 3: Timeline of events and government coalitions showing the *Global Signed Polarization Index* for each time window network. Confidence intervals are provided based on bootstrapping. Political ideology dominating the government is represented as the alignment with parties at the European level and specific controversial events as well as elections are shown for social and political context.

A.4 Preliminary results

Polarization timeline

We see that an underlying polarized structure is present and that the level of polarization of the community is reactive to societal and political changes, within considerably narrow time-frames.

We find that the *Signed Polarization Index* $SPI(G)$ of the networks with shuffled signs, is stable through the time windows and stays in low values, while $SPI(G)$ for the real networks has larger variability and is centered around higher values (Figure **??**). The fact that this result is consistent for all time windows denotes that the signed structure of the networks encodes relevant information that, even if maybe not understood in absolute balance terms, provides an intuition of an underlying preference towards more balanced states.

Given the corrected measure, SPI_{global}, we examine the fluctuations and behaviour of this measure longitudinally through the timeline. Polarization measures are variable with respect to time and seem to follow an increasing trend. However, this trend is clearly non-linear, it seems to peak at certain time windows and decrease especially towards 2020.

To provide an objective intuition in what are the relevant events influencing the perturbations of polarization values, we perform a simple frequency study of words that are more present in the aggregated article text in each specific quarter compared to the full corpus.

Looking at the variations in more detail we identify a calm baseline start for 2014 and the start of 2015, followed by a large increase that peaks at the first quarter of 2016. This first large increase could be related to the shift in public opinion regarding the migration crisis, which provided a controversial situation that broke the traditional left-right wing scheme and social tension was higher.

After that, a period of strong government instability followed, and we can see polarization peaks coinciding with election periods. For context, two other largely controversial and commented events: 1.The very inadequate wording of the far-right Prime Minister at the time, Herbert Kickl, relating ”concentration camps” and migrants and 2. A largely discussed corruption scandal involving Strache, the Vice-Chancellor at the time, which led to government collapse.

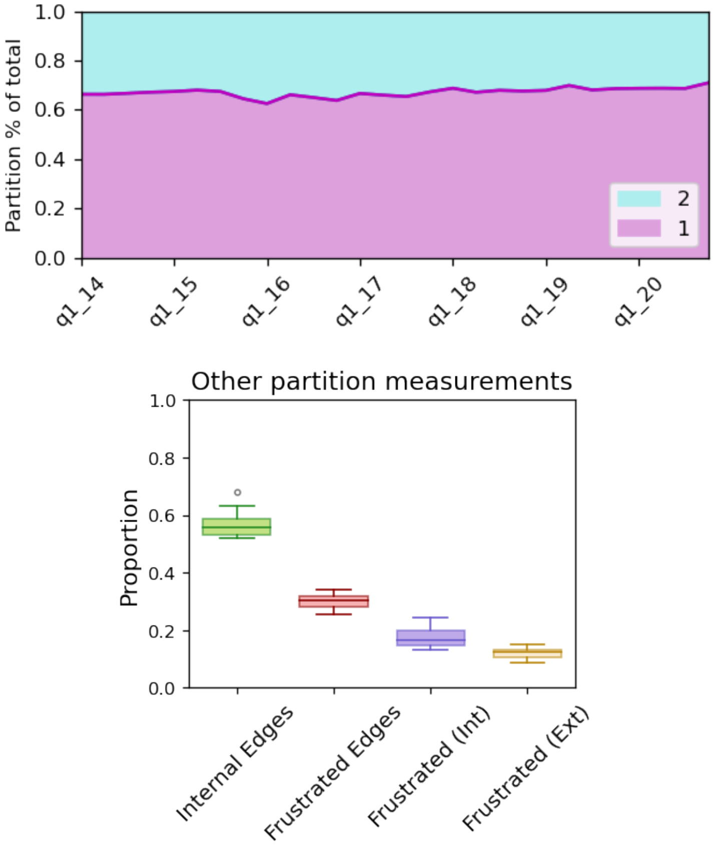


Figure 4: Proportion of active nodes that belong to each partition for each time window (upper figure) and basic measurements of the partitions for all time windows (lower figure): proportion of edges that are internal to one of the partitions, proportion of total frustrated edges, internal frustrated edges (negative) and external frustrated edges (positive).

Meso-measurements: zooming into the polarization mechanisms

By examining the combination of edge characteristics: internal (between nodes of the same partition) or external (between nodes of different partitions), together with the frustration state of the edge, we can describe in more detail the mechanisms underlying the timeline results of 3.

In Figure 5 it is shown that consistently around 60% of the edges are internal. Therefore, the underlying unsigned structure already favours the partitions. We also see that frustrated edges account for around 30% of the edges and are mostly internal (negative inside partitions).

By checking the meso-scale balance measurements we reach an interesting conclusions: First of all, both measures reach high values, but divisiveness is mostly higher or equal to cohesiveness. However, cohesiveness has higher variability and reaches high values for high polarization moments (see Figure 5). Both of these observations point out that our stronger polarization times are given by stronger cohesion inside the partitions, instead of stronger disagreement between the two groups. Also, since internal negative frustration is higher, it would mean that most of the frustration that impedes a clean separation into two groups is given by disagreements within groups, more than unexpected agreements between opposing groups.

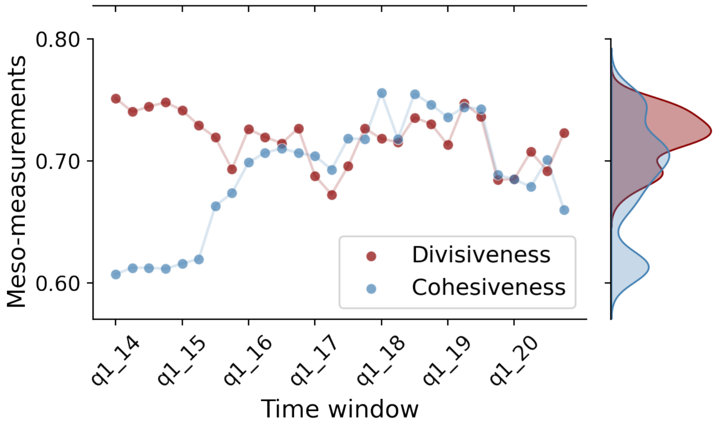


Figure 5: Meso-scale measurements, cohesiveness and divisiveness, for the partitions obtained in each time window with density distributions of each measure. Divisiveness stays at a large value and has one single mode, while cohesiveness fluctuates between two main modes. Note that both measures stay above 0.6, which denotes a significant balance present at all times.

A.5 Conclusions and discussion

Conclusions from the study of DerStandard, related to this particular platform and not directly generalizable:

1. Finding that large scale online political discussions display an underlying polarized structure based on balance.

2. Polarization is a dynamic, reactive phenomenon subject to current political and social events. In this case it affects the community in general, even when only looking at votes regarding any issue or topic.

3. Changes in polarization are driven by stronger cohesion, opposite to what it is speculated (stronger disagreement between communities).

General statements:

1. We provide the first temporal analysis of structural balance in large scale online political discussions.

2. Given the fluctuations of the partitions and distinctive behaviours in specific points: Our SPI is reactive to external social and political context with short-term responses.

Other contributions of this work:

1. Curation and first analysis of a novel dataset, from a platform with good moderation dynamics and extremely loyal user base.

While the results obtained are robust under severa sanity checks, there are some challenges in the use of this methodology. The following are the main limitations that we encountered and the strategy used to tackle them:

1. Arbitrary choice of time windows. To account for the choice of 3-month time windows and the robustness of the results, all measures were compared with those obtained from the aggregated network of 7 years.

2. Assumption of two partitions. Other work has also focused in polarization between more than two groups [29] [31]. In our case we assess the benefit of adding an extra group, which lowers the amount of frustrated edges, balanced with the robustness of the partitions obtained with three groups. The three group partition does not maintain under the robustness check when using the aggregated network.

3. Application to one single platform. Results are not easy to generalize given the unique nature of the dataset. This can be fixed by the analysis and validation of the method on the Wikipedia or Birdwatch data.

B Publication 2.1

B.1 Prior work, background and definitions

In this paper, we start from an important network theoretical result showing how minorities' online visibility is limited by two common principles of human behavior in social environments.

[4] and [36] study network structures based on PA and tuned with different parameters of homophily. They show that *node degree* based recommender systems will not represent minorities adequately given certain states of the parameter space. Google's PageRank algorithm and Twitter's WhoToFollow algorithm are influential examples that suggests information and followers based on node degree[4].

[44] extends this work by studying the consequences of homophily and PA on simple and complex contagion processes such as information spread. They demonstrate a complex relation between network structures and contagion properties and a general 'price of fairness': ensuring information equality comes at the cost of spreading efficiency, and vice versa.

In this paper, we study a related but different scenario. Our first research question asks how homophily and PA influence minorities' ability to overtake the general opinion. Secondly, we are interested in how minorities' opinion dynamics change as we introduce multiple conflicting minorities.

According to [45], conflicts between elite minorities has historically led to social unrest by destabilising the ruling powers. Troubles between minorities is not limited to the powerful classes. In any normally distributed opinion there exists two tails of small strongly conflicting minorities. Secondly, seemingly united minorities break into conflicting sub-communities such as the division of the Reddit based 'anti-work' movement into conflicting 'work reform' and 'anti work' sub-movements.

To study this, we extend the work of [36] with the use of an opinion dynamics model, to see if network structures originated from different homophily and PA settings have an effect on the final representation and distribution of opinions. For this, the group identity label is set as a fixed attribute to each node, while the opinion on a certain issue is allowed to shift given certain rules: each group has an initial and continuous preference for an opinion, but some individuals may be convinced by their surroundings to shift to the other group's opinion. While we expect the outcome of these simulations to resemble the conclusions [36] extract from the study of rankings, the relation between the two is not trivial, as network structure plays a different role in the performance of opinion dynamics models. For example, highly segregated networks in high homophily settings may benefit minorities in the rankings (preservation of degree) but would not allow their opinions to gain control of the discussion.

B.2 Methods (contributions)

Our model consists of two steps. First, we generate a network with different parameters of interests. This model is an extension of the directed preferential attachment with homophily(DPAH) network model developed in [36]. Second, we run an Boltzmann opinion dynamics model on the networks to study how parameters related to the DPAH affects opinion spreading.

B.2.1 Network model

Following [36] we let $G = \langle V, E, C \rangle$ be a directed graph ("digraph") where $V = \{v_1, v_2, \ldots, v_N\}$ is a set of N vertices, $E \subseteq V \times V$ is a set of M edges, and $C \in L^N$ is a list of N class label from the label set L. This graph represents a social network with $|L|$ distinct classes. Each vertex v_i represents an individual belonging to class $c_i \in L$. For example, if the classes are popular American political stances, each vertex will be labeled from the 3-element label set $L = \{\mathit{democrat}, \mathit{moderate}, \mathit{republican}\}$. Finally, each directed edge e_{ij} indicates that individual i has initiated connection with individual j. For example, if we apply this model to Twitter, then an edge e_{ij} indicates that i "follows" j, and if we apply this model to an academic network, then it indicates that i cites j.

The network is grown by first selecting a node i with probability $P(\beta_i)$ where β_i is a node specific fixed activity level. As Twitter activity has been empirically shown to mimic power laws β_i is drawn from such with exponent γ [36]. The next step is to determine who i is going to follow. This decision is shaped by *homophily* and *preferential attachment* which are explained below. The overall process of drawing followers and identify targets is repeated until our stopping condition, $e = dn(n-1)$ with d being a density parameter, is reached

Preferential attachment: That few nodes posses a disproportionate amount of edges is a frequent feature of networks across the the physical, biological, and social world. Preferential attachment states that nodes acquire new edges with a probability proportional to its existing number of edges implying a rich-gets-richer like dynamic for edge generation. Mathematically we say that the probability of node i follows j, $P(i \rightarrow j)$, is proportional to the in-degree of j, k_i

$$P(i \rightarrow j) = \frac{k_j}{\sum_{l=1}^{n} k_l} \tag{1}$$

Homophily: A central function of our class attributes C is to facilitate homophilic edge generation, that is, the tendency for similars to connect. In our case, high homophily parameters will mean edges are more likely to form within classes. The exact probabilities are controlled by a set of homophily parameters. For each class there are three homophily parameters, one for in-group preference and two for out-group preferences. With three classes we have a total of nine homophily parameters (Majority: hMM, hMm_1, hMm_2. Minority one: hm_1m_1, hm_1M, hm_1m_2. Minority two: hm_2m_2, hm_2M, hm_2m_1). We constrain the majority to have equal preferences for each minority ($hMm_1 = hMm_2$) and we assume each class' homophily parameters to sum to one. Because of the symmetry between the minorities we end up with two homophily relations with three parameters free to vary.

$$1 = hMM + 2hMm_i, 1 = hm_im_i + hm_iM + hm_im_j \tag{2}$$

In equation 3 we unite preferential attachment and homophily into our three class directed network with preferential attachment and homophily model.

$$P(i \rightarrow j) = \frac{h_{ij}k_j}{\sum_{l=1}^{n} h_{ij}k_l} \tag{3}$$

A core feature of digital services is recommending films, who to follow, answers to searches, etc. These processes rely on recommender algorithms often based on network properties to rank and identify optimal recommendations for users. For instance, Twitter employs a *Who-to-Follow* (WTF) algorithm to suggest following options. Notably, WTF works locally and obtains different recommendations for different nodes. Thus, to obtain an overall ranking of users we sum how often they are recommended to any nodes.

B.2.2 Opinion Contagion Model

Spreading of ideas is a complex contagion. While exposure to a single infected neighbor may be sufficient to infect an individual with a disease, individuals typically require exposure to an opinion from multiple sources before they adopt it as their own. We therefore develop a novel opinion contagion model governing the spread of opinions across our network.

Our model has two core components: *neighbor agreement* and *class consistency.* We introduce a "softmax" or "Boltzmann" model of contagion, with two key benefits: 1) the probabilistic nature of the model captures the stochasticity in opinion change; and 2) the summation allows our model to generalize to variable numbers of groups.

Neighbor Agreement

Neighbor agreement captures individuals' propensity to adopt opinions that are popular amongst their neighbors. Let $\mathcal{N}_i(\omega)$ be the number of vertices in v_i's neighbor set that hold opinion ω. $\mathcal{N}_i(\omega) = \Sigma_{v_j \in \mathcal{N}_i} I(o_j = \omega)$, where $I(\cdot)$ is the indicator function, which equals 1 if the predicate is true, and 0 otherwise. The likelihood that v_i also holds opinion ω should be correlated with $\mathcal{N}_i(\omega)$, but it is still possible for

v_i to agree with a minority of its neighbors, so we use a softmax neighbor agreement model in which the probability that vertex i takes opinion $\omega \propto \exp(\mathcal{N}_i(\omega))$.

However, v_i may be more influenced by some neighbors than others. For example, v_i may be more likely to agree with neighbors of the same class, or neighbors that are more central to the network. We therefore define a *score* $s_i(v_j)$, which gives the weight that vertex v_i places on neighbor v_j. According to the weighted softmax neighbor agreement model, the likelihood that vertex v_i has opinion ω at the next timestep $P_{na}(o_i^{t+1} = \omega)$ is shown in Equation 4.

$$P_{na}(o_i^{t+1} = \omega) = \frac{1}{Z} \cdot \exp(\sum_{v_j \in \mathcal{N}_i} s_i(v_j) I(o_j^t = \omega)) \tag{4}$$

$Z = \sum_{\omega \in \Omega} P_{na}(o_i^{t+1} = \omega)$ is the normalizing factor. Note that this model is not specific to two or three groups, and can generalization to as many distinct groups as neccessary.

Class Consistency

Class Consistency captures individuals' tendency to adhere to the opinion that is most consistent with their class. For each class, we define a default opinion ω^{def}. An *adherence parameter*, $\alpha_i \in (0, 1)$, modulates the tendency of individual i to adhere to their class-based default opinion ω_i^{def}. Specifically, $o_i^{t+1} = \omega_i^{def}$ with probability α_i, and o_i^{t+1} is governed by Equation 4 with probability $1 - \alpha_i$. In this way, α_i modulates the tradeoff between class consistency and neighbor agreement in opinion updating.

Combining class consistency and neighbor agreement, the probability that node v_i will have opinion ω at time $t + 1$ is specified in Equation B.2.2.

$\mathrm{P}(\mathrm{o}_i^{t+1} = \omega) = \ \alpha_i \cdot I(\omega = \omega_i^{def})$
$+ \ (1 - \alpha_i) \cdot \frac{1}{Z} \cdot \exp(\sum_{v_j \in \mathcal{N}_i} s_i(v_j) I(o_j^t = \omega))$

The opinion of a given individual at the next time-step depends on both the opinions of its neighbors (*neighbor agreement*) and the default opinion of its class (*class consistency*) while its *adherence* parameter modulates the trade-off between these influences.

Experimental configuration

The selection of homophily parameters and minority size will depend on the experiments performed. However, some variables are fixed for all of the following results.

We work with networks of $N = 2000$ nodes and set a density threshold of $d = 0.0015$. We set the activity of nodes to follow a power-law distribution with exponent $\gamma = 3$. All homophily parameters are symmetric with respect to minorities-majority. In these initial simulations, we keep the score of all nodes constant so that $s_i = 1$ for all i.

B.3 Preliminary results

To get a sense of the importance of each parameter in the rankings, we first produce nine simulations to examine relevant combinations of homophily parameters, allowing them to be: very low (0.1 or 0.2), neutral (0.5 or 0.33), or very high (0.8 or 0.9). The two options are selected regarding summation rules from Equation 2. By assessing the proportion of minorities in the top 50 rank of PageRank (Figure 6), we conclude that the nature of the majority is the most relevant: simulations II, V and VIII produce a peak and have a heterophilic majority, while simulations III, VI and IX have a homophilic majority and are lowest. When the majority is neutral, homophilic minorities (simulation I) achieve higher presence in rankings, while neutral and heterophilic minorities are underrepresented in the top 50 (simulations IV and VII).

Replication of results for one minority and expansion of the model

Here, we present our replication of the model presented in [36]. The central challenge here is to make our 2-minority model mimic the original model with just a single minority. Since our model has more

Parameters/simulation	hMM (within majority)	hMm_i (majority to minorities)	hm_im_i (within minorities)	hm_im_j (between minorities)	hm_iM (minorities to majority)
Simulation I: neutral M, homophilic m	1/3	1/3	0.9	0	0.1
Simulation II: heterophilic M, homophilic m	0.2	0.4	0.9	0	0.1
Simulation III: homophilic M, homophilic m	0.8	0.1	0.9	0	0.1
Simulation IV: neutral M, neutral m	1/3	1/3	0.5	0	0.5
Simulation V: heterophilic M, neutral m	0.2	0.4	0.5	0	0.5
Simulation VI: homophilic M, neutral m	0.8	0.1	0.5	0	0.5
Simulation VII: neutral M, heterophilic m	1/3	1/3	0.1	0	0.9
Simulation VIII: heterophilic M, heterophilic m	0.2	0.4	0.1	0	0.9
Simulation IX: homophilic M, heterophilic m	0.8	0.1	0.1	0	0.9

Table 1: Exploration of different states of the parameter space for two minorities.

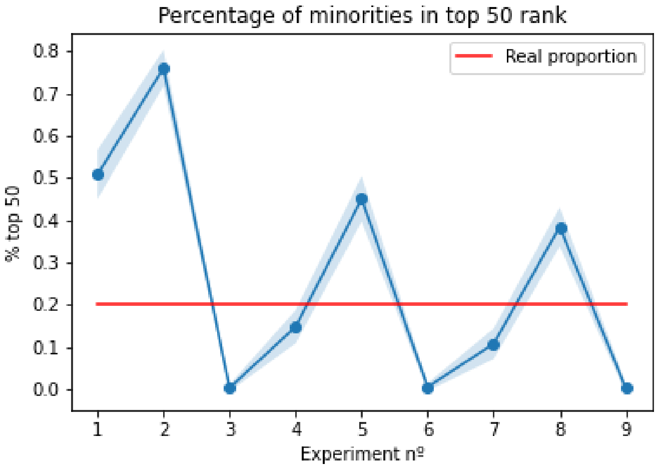


Figure 6: Percentage of minorities in top 50 nodes from PageRank ranking. Each datapoint corresponds to one of the simulations of Table B.3.

homophily parameters the translation is not trivial. For this reason we tried two different ways of expanding the four homophily parameters of [36] to our $3x3$ matrix of parameters in the case of two minorities. Table B.3 provides an explanation for the two cases: Simulation VI satisfies the summation rule (Equation 2) regarding the normalization of the homophily parameters for each group, while Simulation VII does not. By applying these parameters in our $N = 2000$ nodes networks, we find that Simulation VII replicates the target results best (pink trajectories in Figure 7).We find good overall agreement between our results and the original model, however, with one central differences. In our model we see a consistent dip at top-k% Pagerang of 90%.

Despite slight differences we are generally able to reproduce the findings of [36] using the VII parameter setup. Thus, we proceed with our novel two-minority experiments assuming the VII parameters. In all experiments we fix the total minority size to a 20% of the nodes. We then study two double minority cases, one where the initial minority fragments into halves (10% + 10%). In the second condition an additional minority appears and doubles the overall proportion of minority nodes in the network (20% + 20%). In both conditions we set the homophily parameter between minorities to $h_{m_im_j} = 0$, as we assume them to be conflicting.

Figure 8 depicts the ranking curves for our three cases: of a single minority (VII), a fragmented minority ($divided$), and an additional minority ($double$). We see the largest differences between the groups in experiment b (homophilic minorities and heterophilic majorities). We see the divided minority loses PageRank representation relative to the others. The doubled minority also has a slight disadvantage to the single minority case. In the neutral case (type c), the three conditions fares similarly however with the single minority ranking higher overall. In last case of of homophilic majorities and heterophilic minorities, we find no differences between the three conditions. Summarising these figures we see 1) multiple minority situations are never better than single minority cases, and 2) differences are eliminated as longer rankings are considered.

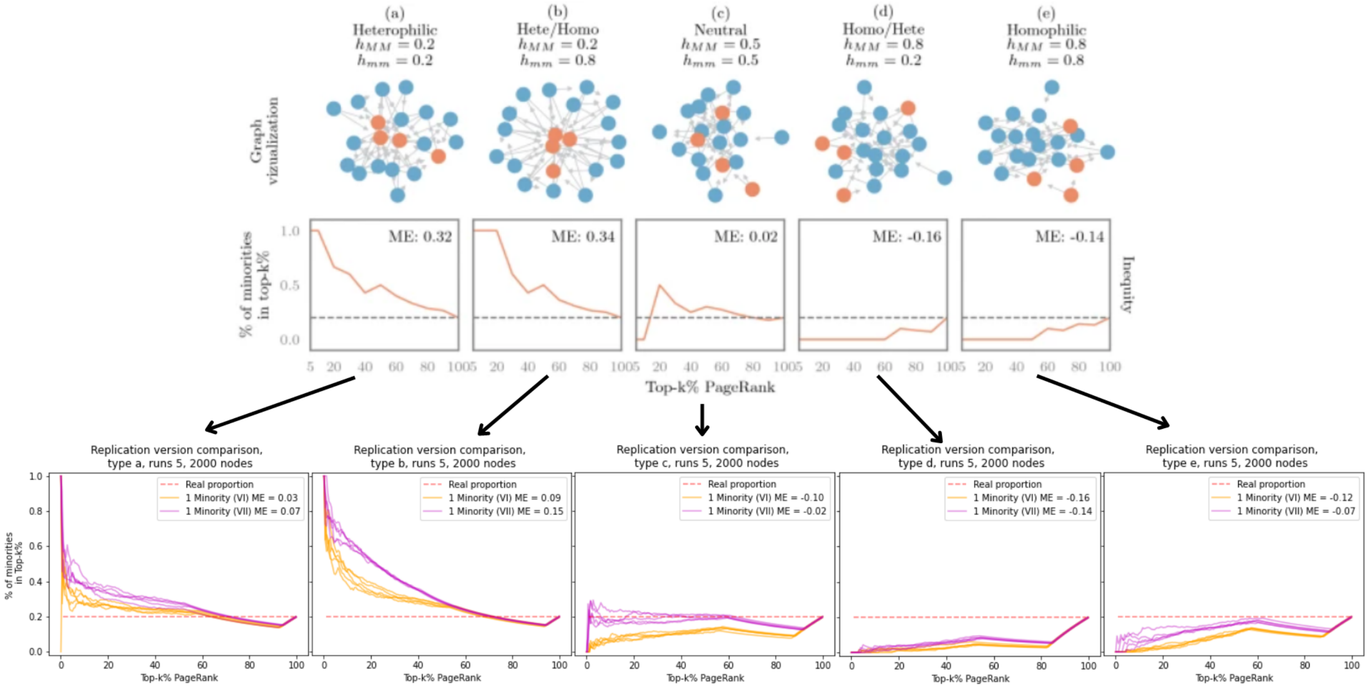


Figure 7: Upper part extracted from Figure 1 of [36]. Comparison of original results on a 20 node network to our simulations on 2000 nodes, applying different versions regarding the expansion of the model. ME is the mean distance of each point regarding the real proportion over all k of the rank. In our case we average them over the $r = 5$ runs in each figure.

	Parameters	(a)	(b)	(c)	(d)	(e)
1 Minority (size: f_m)	hMM	0.2	0.2	0.5	0.8	0.8
	hmm	0.2	0.8	0.5	0.2	0.8
2 Minorities VI ($f_{m1} + f_{m2} = f_m$)	hMM	0.2	0.2	0.5	0.8	0.8
	hMm_i	0.4	0.4	0.25	0.1	0.1
	hm_im_i	0.1	0.4	0.25	0.1	0.4
	hm_im_j	0.1	0.4	0.25	0.1	0.4
	hm_iM	0.8	0.2	0.5	0.8	0.2
2 Minorities VII ($f_{m1} + f_{m2} = f_m$)	hMM	0.2	0.2	0.5	0.8	0.8
	hMm_i	0.8	0.8	0.5	0.2	0.2
	hm_im_i	0.2	0.8	0.5	0.2	0.8
	hm_im_j	0.2	0.8	0.5	0.2	0.8
	hm_iM	0.8	0.2	0.5	0.8	0.2
2 CONFLICTING Minorities VII ($f_{m1} + f_{m2} = f_m$)	hMM	0.2	0.2	0.5	0.8	0.8
	hMm_i	0.8	0.8	0.5	0.2	0.2
	hm_im_i	0.2	0.8	0.5	0.2	0.8
	hm_im_j	0	0	0	0	0
	hm_iM	0.8	0.2	0.5	0.8	0.2

Table 2: Homophily parameters when extending the model of 1 Minority to 2 Minorities and the respective values for each scenario type ((a), (b)...), also includes the fragmentation of the minority group into two conflicting groups.

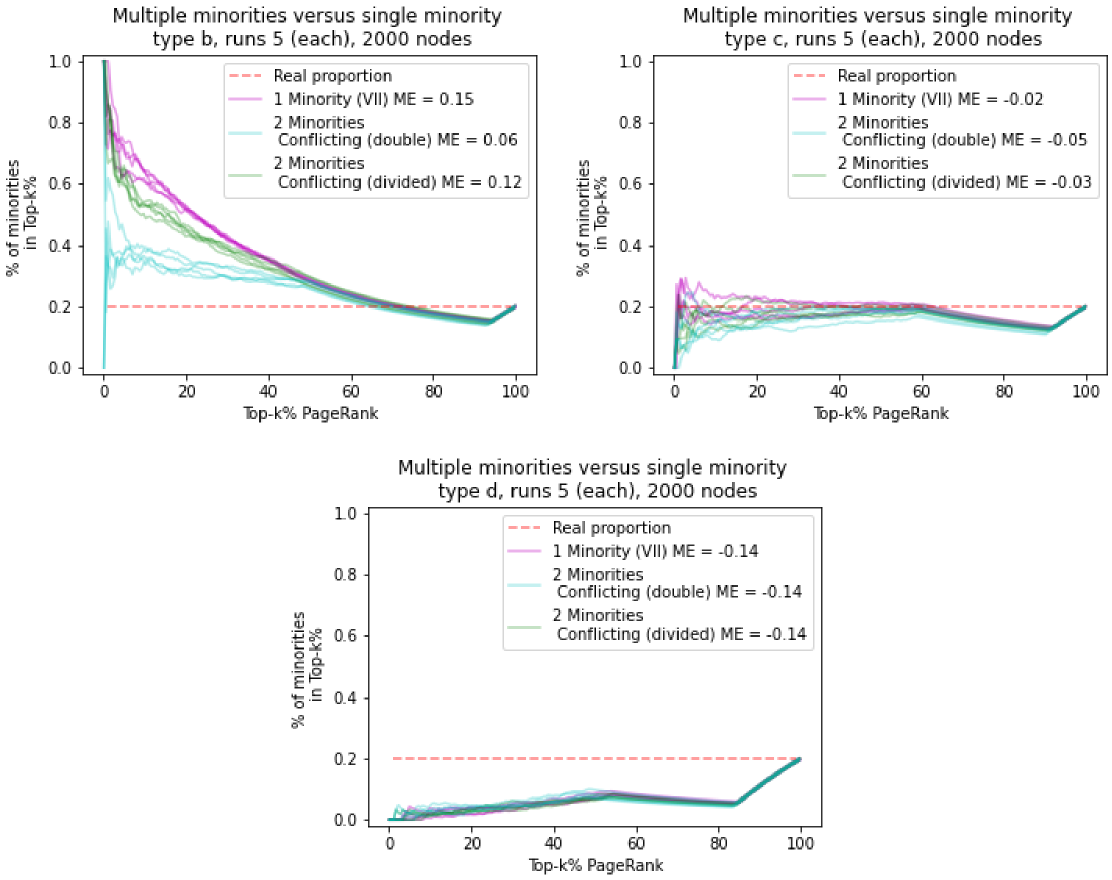


Figure 8: Ranking curves of one 20% minority (following VII discussed previously), a divided 20% minority, and the case of one 20% minority that is in conflict with another minority (in the point of view when this other minority is part of the majority).

Opinion dynamics

We studied opinion spread on two of the different network types: Simulation I is comprised of a neutral majority and homophilic minorities, and Simulation IX comprised of a homophilic majority and heterophilic minorities. For each experiment, a starting network of 2,000 nodes was chosen (e.g. Simulation I or Simulation IX) and initial opinions were determined based on group membership, with 25% of the nodes given a random opinion to allow for some diversity. Opinions were then updated for 1,000 nodes (on average) for each Monte Carlo time step according to the Opinion Dynamics model over 200 time steps total. At each time step, we tracked the total number N_i of each opinion group where i=$\{M$,m_1,$m_2\}$.

In Simulation I, we looked at two sub-cases: first, where the majority is not stubborn and the two minorities are semi-stubborn ($\{\alpha_M$,α_{m1},$\alpha_{m2}\}$ = $\{0.05,0.5,0.5\}$), and second, where all groups are semi-stubborn ($\{\alpha_M$,α_{m1},$\alpha_{m2}\}$ = $\{0.5,0.5,0.5\}$). We considered the same sub-cases for Simulation IX, as well as two additional cases: one in which none of the three groups are stubborn ($\alpha_i = 0$ for all i)[1] and another in which the minorities are very stubborn while the majority is not stubborn ($\{\alpha_M$,α_{m1},$\alpha_{m2}\}$ = $\{0.05,0.8,0.8\}$).

In the results for the neutral majority/homophilic minorities we find an evolution towards near-equal opinion group size when the majority is not stubborn (α_M=0.05). However, in the case where the majority is semi-stubborn ($\alpha_M = 0.5$), the majority opinion fraction N_M/N_{tot} decreases from 0.8 to $\sim$ 0.6 by time step 10 and remains there throughout the simulation. Thus, we find that the adherence parameter of the majority can play a large role in governing opinion fractions in this network type.

These same subcases were investigated on the starkly different network type with a homophilic majority and heterophilic minorities (Simulation IX). Interestingly, the results for these two subcases are indistinguishable from the previous network type (Simulation I). We do note, however, that these analyses consider only the fraction of different opinion groups, and not the means by which the opinions have spread throughout the network; we plan to study the distribution of opinions across the network in the future.

In the case where none of the groups are stubborn ($\alpha_i = 0$ for all i), we also find an evolution towards near-equal opinion share, similar to the $\{\alpha_M$,α_{m1},$\alpha_{m2}\}$=$\{0.05,0.5,0.5\}$ subcase. For the final subcase, where both minorities are stubborn and the majority is not ($\alpha_{m2}\}$=$\{0.05,0.8,0.8\}$), an equal spread of opinions is realized.

B.4 Discussion

Taken together, these simulation results indicate that the differing adherence parameters α_i play a large role in governing the overall opinion fractions in a given network. In future work, we would like to explore the trade-off between α and network homophily by repeating these same sub-cases on the different network types listed in Table B.3. In addition, we need to consider additional metrics for studying opinion spread such as the distribution of opinions across the network. It is most likely in the sub-cases investigated here, that opinion changes are occurring at boundaries between the majority and minority groups, but we would like to verify this hypothesis.

Another interesting result to highlight is the case of two stubborn minorities and a not-stubborn majority ran on Simulation IX (homophilic majority and heterophilic minority). Despite a homophilic majority, this network realizes equal opinion fraction in very few time steps. This suggests that minorities may be able to increase their opinion share throughout a network via a high connectivity with other groups (heterophily) combined with strong adherence to their own group's original opinion. More simulations are needed to verify this claim, in particular, as mentioned above, to study whether this result occurs for the same combination of adherence parameters on different network types.

The main goal in creating such models is to idealize the models to a point that can give us some insights about actual opinion spread. Thus, once we finish our analysis of the the trade-off between the homophily and the adherence parameter, we will look for data to validate our model's empirical validity. A natural place to start would be to look at Twitter data, e.g. such as that in [46]; this study analyzes an individual's 'threshold' to join a collective protest. This might give us a sense of what are reasonable (or actual) adherence parameters shared by a population.

Another important use of network models is to test the plausibility of theories in less formal domains. Our model seems well-suited to test some of the claims made in the growing literature on so-called 'elite-

[1] This case is similar to a majority-vote model

capture', the phenomenon when a minority has disproportional political and epistemic influence due to its powerful status [47].