Voice and silence in public debate:
Modelling and observing collective opinion expression online

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM (Dr. rer. nat.)

im Fachgebiet
Informatik

Vorgelegt
von M.Sc. Felix Gaisbauer
geboren am 28.05.1993 in Passau

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Jürgen Jost, Max-Planck-Institut für Mathematik in den Naturwissenschaften

2. Prof. Dr. Walter Quattrochiocchi, Universität La Sapienza

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 12.09.2022 mit dem Gesamtprädikat magna cum laude

Meiner Großmutter

Danksagung

Diese Dissertation hätte nicht zustandekommen können ohne Unterstützung. Auf fachlicher und institutioneller Ebene gilt hier mein Dank zuvorderst Eckehard Olbrich. Desweiteren: Jürgen Jost, Sven Banisch sowie vielen ehemaligen und aktuellen Mitgliedern der Jost-Gruppe, und definitiv nicht zuletzt: Antje Vandenberg. Auch in persönlichen Belangen bedarf, wer sich in so ein Unterfangen hineinmanövriert, größeren Rückhalts. Armin Pournaki hat es geschafft, sowohl hier als auch dort eine nicht unerhebliche Rolle zu spielen. Dank auch Gregor Bös, Jasper Anger, und weiterer, deren Freundschaft mir trotz des Fehlens namentlicher Erwähnung erhalten bleiben möge. Meinen Eltern natürlich, und meiner Schwester.

Abstract

This thesis investigates how group-level differences in willingness of opinion expression shape the extent to which certain standpoints are visible in public debate online. Against the backdrop of facilitated communication and connection to like-minded others through digital technologies, models and methods are developed and case studies are carried out – by and large from a network perspective.

To this end, we first propose a model of opinion dynamics that examines social-structural conditions for public opinion expression or even predominance of different groups. The model focuses not on opinion change, but on the decision of individuals whether to express their opinion publicly or not. Groups of agents with different, fixed opinions interact with each other, changing the willingness to express their opinion according to the feedback they receive from others. We formulate the model as a multi-group game, and subsequently provide a dynamical systems perspective by introducing reinforcement learning dynamics. We show that a minority can dominate public discourse if its internal connections are sufficiently dense. Moreover, increased costs for opinion expression can drive even internally well-connected groups into silence.

We then focus on how interaction networks can be used to infer political and social positions. For this purpose, we develop a new type of force-directed network layout algorithm. While being widely used, a rigorous interpretation of the outcomes of existing force-directed algorithms has not been provided yet. We argue that interpretability can be delivered by latent space approaches, which have the goal of embedding a network in an underlying social space. On the basis of such a latent space model, we derive a force-directed layout algorithm that can not only be used for the spatialisation of generic network data – exemplified by Twitter follower and retweet networks, as well as Facebook friendship networks – but also for the visualization of surveys. Comparison to existing layout algorithms (which are not grounded in an interpretable model) reveals that node groups are placed in similar configurations, while said algorithms show a stronger intra-cluster separation of nodes, as well as a tendency to separate clusters more strongly in retweet networks.

In two case studies, we observe actual public debate on the social media platform Twitter – topics are the Saxon state elections 2019, and violent riots in the city of Leipzig on New Year's Eve in the same year. We show that through the interplay of retweet and reply networks, it is possible to identify differences in willingness of opinion expression on the platform between opinion groups. We find that for both events, propensities to get involved in debate are asymmetric. Users retweeting far-right parties and politicians are significantly more active, making their positions disproportionately visible. Said users also act significantly more confrontational in the sense that they reply mostly to users from different groups, while the contrary is not the case. The findings underline that naive reliance on what others express online can be collectively dangerous, especially in an era in which social media shapes public discourse to an unprecedented extent.

Deutsche Zusammenfassung

Diese Arbeit befasst sich damit, inwiefern die unterschiedliche Bereitschaft bestimmter Gruppen, ihre Meinung online zu äußern, die Sichtbarkeit ebenjener Meinungen in öffentlichen Debatten beeinflusst. Dies wird sowohl mathematisch modelliert als auch in empirischen Fallstudien (aufbauend auf dafür entwickelten quantitativen Methoden) untersucht. Angesichts der zunehmenden Vereinfachung von Kommunikation und Vernetzung durch digitale Technologien liegt der Fokus hierbei auf der Nutzung und Entwicklung verschiedener Techniken zur Analyse von sozialen Netzwerken.

Zunächst wird ein Modell konzipiert, das die sozio-strukturellen Bedingungen für öffentliche Meinungsäußerung oder sogar Meinungsdominanz verschiedener Gruppen beleuchtet. Dabei wird nicht untersucht, ob Akteure ihre Meinung kollektiv ändern, sondern ob sie sich dazu entscheiden, ihre Meinung öffentlich zu äußern. Meinungsgruppen, deren Meinungen im Modell unverändert bleiben, interagieren miteinander und werden in ihrer Bereitschaft, die Meinung zu äußern, von den Reaktionen anderer beeinflusst. Das Modell wird zunächst spieltheoretisch formuliert. Anschließend wird es durch die Einführung von 'reinforcement learning' in ein dynamisches System überführt, dessen Fixpunkte analysiert werden. Es wird gezeigt, dass eine Minderheit die öffentliche Meinung prägen kann, wenn sie sich intern stark genug vernetzt. Zudem können erhöhte Kosten für Meinungsäußerung auch zahlenmäßig große und gut vernetzte Gruppen von Meinungsäußerung abhalten.

Weiterhin wird ein kräftebasierter Algorithmus zur Visualisierung von Netzwerken entwickelt, der zur Inferenz sozialer oder politischer Positionen verwendet werden kann. Algorithmen, die physikalische Kräfte zur Netzwerkvisualisierung simulieren, werden in verschiedenen wissenschaftlichen Disziplinen häufig benutzt. Ihre Ergebnisse entziehen sich indes rigoroser Interpretation. In dieser Arbeit wird gezeigt, dass die Ableitung eines kraftbasierten Algorithmus aus sogenannten 'latent space models' es ermöglicht, die Einbettung von Akteuren in einem Raum als deren soziale oder politische Positionen zu interpretieren. Der resultierende Algorithmus kann nicht nur zur Visualisierung sozialer Netzwerke, wie z.B. von Follower- oder Retweet-Netzwerken, sondern auch für Umfragedaten verwendet werden. Bereits existierende Algorithmen, die nicht auf einem interpretierbaren Modell beruhen, werden mit dem entwickelten Algorithmus verglichen. Sie resultieren in den untersuchten Fällen in ähnlichen Konfigurationen, unterscheiden sich aber sowohl durch übermäßig starke Trennung dicht miteinander verbundener Knoten, als auch durch die Tendenz, Cluster in Retweet-Netzwerken stärker räumlich zu separieren.

In zwei Fallstudien werden öffentliche Debatten auf Twitter analysiert – einerseits zu den Landtagswahlen 2019 in Sachsen, andererseits zu Unruhen in der Silvesternacht 2019 im Leipziger Stadtteil Connewitz. Durch die komplementäre Betrachtung von Retweet- und Reply-Netzwerken können Unterschiede in der (Gruppen-)Bereitschaft zur Meinungsäußerung erfasst werden. In beiden Fällen ist eine asymmetrische Neigung zur Meinungsäußerung erkennbar: User, die hauptsächlich Inhalte rechtspopulistischer Parteien geteilt hatten, zeigten eine signifikant höhere Aktivität, die ihre Standpunkte überproportional sichtbar machte. Zudem war eine höhere Bereitschaft zur Konfrontation anderer mit der eigenen Meinung sichtbar, d.h. diese Usergruppe reagierte in Debattenbeiträgen häuptsächlich auf Tweets von Usern anderer politischer

Ausrichtung. Umgekehrt waren User, die eher linksliberale Inhalte teilten, sowohl weniger aktiv als auch weniger dazu bereit, auf Tweets von Usern mit anderen Meinungen zu antworten. Die Ergebnisse der Arbeit untermauern, dass jene Standpunkte und Inhalte, die online sichtbar werden, nicht notwendigerweise auch repräsentativ für die Nutzergruppen sind (geschweige denn darüberhinaus) – eine Problematik insbesondere in Zeiten ungekannten Einflusses sozialer Medien auf öffentlichen Diskurs.

Contents

List of Figures

List of Tables

Er stand hinter einem der Fenster, sah durch den zartgrünen Filter der Gartenluft auf die bräunliche Straße und zählte mit der Uhr seit zehn Minuten die Autos, die Wagen, die Trambahnen und die von der Entfernung ausgewaschenen Gesichter der Fußgänger, die das Netz des Blicks mit quirlender Eile füllten; er schätzte die Geschwindigkeiten, die Winkel, die lebendigen Kräfte vorüberbewegter Massen, die das Auge blitzschnell nach sich ziehen, festhalten, loslassen, die während einer Zeit, für die es kein Maß gibt, die Aufmerksamkeit zwingen, sich gegen sie zu stemmen, abzureißen, zum nächsten zu springen und sich diesem nachzuwerfen; kurz, er steckte, nachdem er eine Weile im Kopf gerechnet hatte, lachend die Uhr in die Tasche und stellte fest, daß er Unsinn getrieben habe. – Könnte man die Sprünge der Aufmerksamkeit messen, die Leistungen der Augenmuskeln, die Pendelbewegungen der Seele und alle die Anstrengungen, die ein Mensch vollbringen muß, um sich im Fluß einer Straße aufrecht zu halten, es käme vermutlich – so hatte er gedacht und spielend das Unmögliche zu berechnen versucht – eine Größe heraus, mit der verglichen die Kraft, die Atlas braucht, um die Welt zu stemmen, gering ist, und man könnte ermessen, welche ungeheure Leistung heute schon ein Mensch vollbringt, der gar nichts tut. *R. Musil*

Chapter 1

Introduction

Through digitalization and facilitated communication, societies produce and are confronted with an increasing amount of information. This development has made human behaviour observable to an uprecedented extent. Continuous data collection, storage of online (inter-)actions, and the development of novel techniques for their analysis have enabled researchers to study social phenomena in new scale and resolution. Especially the scientific potential of digital trace data – data documenting the interactions of users with digital devices or services, which constitute a seemingly unobstructed view of actual behaviour of individuals – has been highlighted (see e.g. [95, 6, 75]). In light of this apparent "measurement revolution" [75] for the social sciences, some diagnosed an "end of theory" [6], indicating that the sheer mass of available data made it possible to do without social theory and models altogether. Researchers should let data "speak for itself" [95], which implied that the focus of the social sciences ought to shift away from seeking explanation for human behaviour towards its prediction on the basis of available data. Approaches were developed which took digital traces of humans online at face value: This included forecasting election results on the basis of simple tweet counts [150] or near-term values of economic indicators via Google Trends data [35] – seemingly, with surprising success. Raw data alone appeared to be sufficient for future social research. These contributions had significant impact on the scientific community.[1]

Often, however, they were not met by agreement: That Twitter can be used to predict election results was rebutted quickly [74, 57], and the spectacular claims about the future of social science were contested by many [25, 67], not least due to the naturalization of social patterns found in (digital) trace data [147]. Computational social science, the field in which such research is conducted, studies social phenomena with computational approaches – computational methods are both developed and applied to typically large-scale, and sometimes simulated, human behavioural data (see e.g. [85]). It is necessarily inter- or multi-disciplinary and involves, among others, social and computer scientists, physicists, and mathematicians. Problematically, "collaboration

[1] [6] (notably, published in a non-scientific magazine) has been cited more than 2,000 times in the scientific literature according to Google Scholar (accessed on 01.08.2021), [150] more than 3,000 times, and [95] more than 7,000 times – to which this thesis will contribute one more for each.

[between faculties] is often not encouraged, and too often is discouraged [...] Computational researchers and social scientists tend to be in different units in distinct corners of the university, and there are few mechanisms to bring them together." [85, p. 1060] Therefore, the field can be prone to the reproduction of divides between (and within) disciplines – for which the purely data-focused approaches sketched above seem emblematic. Of course, the referenced works stand out in their optimism with respect to the potential of raw digital trace data in the study of human behaviour. Programmatic publications in computational social science have stressed value of pluralism and co-existence of methods [44, 89, 67, 160]. Segregative, data-centered approaches and the (re-)creation of dichotomies, stemming from different epistemic values in different disciplines, such as explanation versus prediction [67], quantitative versus qualitative [89], or solution-oriented versus theoretical [44], have been marked as unproductive for the field. But the general question of what role and importance solution-oriented [160], prediction-centered [67] approaches should play in this endeavour remains – or, phrased the other way round: Where social theory, and explanation, are needed, and where one might be able to do without them.[2]

Explanatory approaches are necessary, as will become visible in the course of this work, if one asks how the data that is supposed to be analysed has come about. Virtually always if one wants to make predictions about phenomena which are more general than the data under investigation, understanding data-generating processes is vital: If one only explores data that is available in digital traces and makes those the basis of one's prediction, there lies a potential cause of error [147, 89]. The question here is: What phenomena of interest are reflected by what becomes visible and therefore *accessible* as digital traces on, say, social media platforms [75, 133]? What are the limits and pitfalls of digital trace data taken at face value? Users of online platforms might not be representative of the general population [99], platform design might favour certain modes of interaction and discourage others [147], or certain ranking algorithms might be biased with respect to minority groups, e.g. through network-immanent homophily effects [79].

It might also be the case that behavioural differences between users or user groups on the platform skew purely quantitative analyses.[3] Some users, even though part of a platform, might even *refrain from expressing themselves altogether.* The phenomenon of silence of certain groups, especially in political contexts, has received a great deal of attention in the analogue realm [112, 60]. A closer look on its effects on online environments is in order. Considerations that were valid in the offline world might - at least partly - not applicable any longer [123]. Moreover, online platforms promised from their inception on to give voice to the ones who had previously been unheard. To be able to evaluate this, a thourough investigation of which content, users and social

[2]Explanation can be interpreted very differently, of course: Some might be interested in plausible mechanisms on the micro level which then shape macro outcomes [44], others in a more qualitative *understanding* [89].

[3]This, in fact, turned out to be the case for the election prediction method of Tumasjan et al. [150]: If one includes the Pirate Party, a small political party concerned with digitalization and digital self-determination which received only a negligible number of votes, in their analysis, the whole prediction collapses [74]. Their voters and party members were mostly tech-savvy and had a strong interest in digital technologies, and were therefore extremely impactful on Twitter.

groups become audible (or, rather, visible[4]) in online environments, and which do not, is needed.

This thesis attempts such an account: It inquires how the interplay of expressive and silent users shapes what becomes visible on social media platforms against the backdrop of facilitated communication and connection to (especially like-minded) others in these spaces. This is done via modelling, the development of methods, and case studies – which requires methods and concepts from very different disciplines. In 'Data Theory,' Simon Lindgren approaches the question of the relation of data analysis and social theory in the digital age. He argues that "theory needs data [...] we shall look at how knowledge about particular data can be advanced through some particular social theory [...] The overarching goal is the productive meeting of the two." [89, p. 21] He calls for a "patchwork of solutions" [89, p. 28] – a certain eclecticism with respect to both theory and method in data-driven research – in order to bring quantitative data analysis and social theory closer together. In that spirit, the present work is somewhat eclectic in the methods and models developed. Its focus is quantitative, and models and methods are developed with techniques from network science, dynamical systems, game theory and statistics. But on the other hand, and for the reasons sketched above, we[5] will seek to ground the approaches in plausible accounts of human interaction. A game-theoretic model will subsume a theory from communication science, the spiral of silence theory [111, 113], a force-directed layout algorithm will be derived from a latent space model of homophilious interactions, and we will interpret results from our case studies with the concept of counterpublic spheres [48], which originated from a critique of Habermas' concept of the public sphere [62, 61].

This work comprises three parts, one theoretical, one methodical, and one observational:

(i) A game-theoretically substantiated model of public opinion expression, where we identify social-structural conditions for which certain opinion groups might be disproportionately or even exclusively audible.

(ii) A statistical method to infer political positions of actors via a force-directed layout of their interaction networks.

(iii) A method to assess ideological differences in engagement in public debate on Twitter, applied in two case studies.

Their common focal point is the question of which social groups are highly willing to express their opinions online, and which ones are not, explored by and large from a social-structural perspective.

The model in Ch. 2 will shed light on an effect of social interaction that has long been ignored in opinion dynamics: That opinion exchange between individuals can also have an effect on how willing they are to express their opinion publicly. Groups

[4] In the following, it will not be avoidable to mix sensory perceptions in the interchangeable usage of these two terms.

[5] While I (the author of this thesis) am leading author of all publications this thesis comprises and sole contributor to all of its other parts, the publications were produced in collaboration with other scientists. Using 'we' as a pronoun in parts that subsume a publication which was done in collaboration with others, and 'I' in all others appears to be rather confusing for the reader, which is why I will use 'we' throughout (except, of course, in this footnote).

of agents with different opinion on an issue interact with each other, changing the willingness to express their opinion according to the feedback they receive from others on their opinion. We formulate the model as a multi-group game and investigate its Nash equilibria. We also provide a dynamical systems perspective: Using the reinforcement learning algorithm of Q-learning, we reduce the N-agent system in a mean-field approach to two dimensions which represent the two opinion groups. This two-dimensional system is analyzed in a comprehensive bifurcation analysis of its parameters. The model identifies social-structural conditions for public opinion predominance of different groups. Among other findings, we show that a minority can dominate public discourse if its internal connections are sufficiently dense. Moreover, increased costs for opinion expression can drive even internally well-connected groups into silence.

In order to be able to compare activity differences between different social groups empirically, one must, possess a method to identify these groups. In Ch. 3, a force-directed network layout algorithm is developed. The algorithm places individuals in a latent social space in which clusters of users then may be interpreted as social groups. This method brings together two strands of research: Latent space approaches to network analysis and force-directed layout algorithms. The latter are used ubiquitously for data exploration, illustration, and analysis. Nevertheless, an interpretation of the outcomes of graph drawings with force-directed algorithms is not straightforward. We argue and show that explicit interpretability can be provided by latent space approaches, which have the goal of embedding a network in an underlying social space. There, the distance between nodes influences the probability of a tie between them. We derive force equations in order to infer the latent space for different interaction types, with which e.g. follower and friendship networks, sharing networks, or surveys can be spatialised. The force-directed layout can not only be used for the spatialisation of generic network data – exemplified by Twitter follower and retweet networks, as well as Facebook friendship networks – but also for the visualization of surveys. Comparison to existing layout algorithms (not grounded in an interpretable model) reveals that node groups are placed in similar configurations. Nevertheless, the other algorithms show a stronger intra-cluster separation of nodes, as well as a tendency to separate clusters more strongly in retweet networks.

Ch. 4 investigates public debate on Twitter via network representations of retweets and replies. We argue that tweets observable on Twitter have both a direct and mediated effect on the perception of public opinion. Through the interplay of the two networks, it is possible to identify group-level differences in willingness of opinion expression in reply sections on the platform. The method is employed to observe public debate about two events: The Saxon state elections and violent riots in the city of Leipzig in 2019. We show that in both cases, different opinion groups exhibit different propensities to get involved in debate, and therefore have unequal impact on public opinion. Users retweeting far-right parties and politicians are significantly more active, hence their positions are disproportionately visible. Said users also act significantly more confrontational in the sense that they reply mostly to users from different groups, while the contrary is not the case.[6]

[6]The chapter (as well as the preceding one), of course, *do* base their analyses on digital trace data. This might appear to contradict previous remarks about the fact that users that refrain from expressing themselves

Ch. 5 concludes with a summary.
Parts of this thesis contributed considerably to the following publications:

- Ch. 2: Dynamics of opinion expression (2020). Felix Gaisbauer, Eckehard Olbrich and Sven Banisch. Phys. Rev. E 102, 042303.
- Ch. 3: Grounding force-directed network layouts with latent space models. Felix Gaisbauer, Armin Pournaki, Sven Banisch and Eckehard Olbrich (2021). arXiv:2110.11772 [cs.SI]
- Ch. 4: Ideological differences in engagement in public debate on Twitter. Felix Gaisbauer, Armin Pournaki, Sven Banisch and Eckehard Olbrich (2021). Plos one, 16(3), e0249241.

altogether, e.g. on a social media platform, do not show up in digital traces. But through developing methods for which the opinion/political position on an issue can be discerned through (network) data independent of the actual opinion expression, one can go beyond isolated inspection of opinions articulated online. Users who did not express their opinion through comments at all can be included in the analysis, and opinion groups can be compared in their willingness of opinion expression. Certain limitations remain, of course: If users do not interact with others in any way, they will also not show up in the collected data.

Chapter 2

Modelling dynamics of opinion expression

2.1 Silence in models of opinion exchange

Fundamental to models of opinion dynamics is the assumption that people's opinions are, in some way or another, influenced by the opinion of their peers. There is an extensive amount of models of opinion change in social systems (see [90, 32, 46] for reviews). But while it is a plausible assumption that people who express their opinion about an issue are sensitive to approval and disapproval, feedback on the opinion need not necessarily lead to its reconsideration. It might also affect one's willingness of opinion *expression*: The more positive (negative) the feedback, the more (less) motivated one feels to publicly express one's opinion.

In the opinion dynamics literature, this approach has remained rather unexplored. However, it is worth to be considered: In general, people are not always willing to reveal their opinion on certain issues to others [94]. A recent study shows that only a minority of users who consume news online are also involved in sharing and discussing them [77]. Thorough research on opinion dynamics must take into account that some individuals might choose to not express their opinion publicly, which has profound effect on how others perceive the opinion climate in a social system. We will hence, in this chapter, focus on a model of the *expression of*, and not the change in, opinions. Models have been developed which distinguish between internal and publicly revealed opinion of agents [83, 164, 53, 54, 70, 42, 135, 34, 33], often building on the seminal experiments of Asch [8] (see also [84]). As a reaction to peer pressure, agents might publicly display conformity, even though their internal opinion remains unchanged. The separation between publicly visible and privately held position is also established in the present work – but in this case, a discrepancy between the own and the (perceived) publicly dominant opinion will result in silence.

A theory of public opinion expression has already been developed half a century ago, with Elisabeth Noelle-Neumann's influential 'spiral of silence' [111, 113]. Roughly speaking, Noelle-Neumann sees the fear of isolation as an essential drive for

how humans behave in public. Especially with respect to morally charged topics, individuals constantly and mostly sub-consciously monitor the 'opinion landscape' around them (they possess a 'quasi-statistical sense' [111, 113]) and might refrain from expressing their opinion if they believe to be part of a minority. On the other hand, a belief to hold the majority position might encourage them to express their view. Since each individual's decision whether to express her[1] opinion or not influences how others perceive the opinion landscape, whose evaluation might then change accordingly, a dynamical development (for which Noelle-Neumann used the metaphor of a spiral) follows in which the seemingly dominant opinion fraction becomes more and more vocal and the perceived minority fraction becomes more and more silent.

Noelle-Neumann's spiral of silence is particularly interesting for mathematical modelling since it links a micro mechanism with a dynamical development at the macro level. We will follow this route in the proposed model, as well. Nevertheless, the deeply rooted fear of isolation that Noelle-Neumann postulates will not be necessary for the model to gain traction.[2] It suffices that the behavioral adjustment of agents depends solely on the social feedback they receive when they express their opinion – and that agreement is seen as positive feedback, while disagreement is perceived negatively. This approach is embedded in an account of social influence termed social feedback theory [13], an affective experience-based interaction mechanism that has already been shown to lead to opinion polarization in connected networks of sufficiently strong community structure [12]. In the present approach, the effect of social interaction is directed towards the *willingness* of, or incentives for, individuals to publicly express their opinion. We investigate the structural conditions that promote or hinder opinion expression of different opinion groups. While there have been efforts to model opinion expression and specifically the spiral of silence, they are either in large parts simulative [137, 129, 144, 156, 56] or directed towards the effect of specific circumstances on the spiral of silence (mass media [136], social bots [129], or the long-time effect of charismatic agents [56]). Granovetter and Soong [60], and subsequently Krassa [82], employ a threshold model of opinion expression which only applies to cases in which a certain opinion is already suppressed. We aim here towards a more general, structural understanding of the dynamics of opinion expression.

In the following, we will first describe the baseline social structure and the two central structural parameters of the model. We then represent the model as a multi-group silence game on the agent network, and investigate its Nash equilibria with respect to its structural parameters. A dynamical systems perspective is provided by the introduction of Q-learning and a subsequent two-dimensional approximation of the dynamical system. This makes it possible to perform a bifurcation analysis for the different parameters involved. We conclude with a discussion of the results and an outlook. The present chapter is concerned with a model which does not exclusively apply to online settings, but is of increased relevance there due to the facilitation of communicative ties, especially to like-minded others. It explores social-structural conditions for public

[1] In this work, we will use 'she', 'her', etc. as the generic third-person singular pronouns.

[2] This fear of isolation implies a need for being accepted in a social environment. Especially in online contexts, which are our focus here, this assumption might not be fulfilled: Whether some users, possibly all anonymous, in some corner of the internet accept an individual as part of their group or not might not always be a strong motivation for the individual's actions.

opinion co-existence between or public predominance of certain groups.

2.2 Social-structural setting

For simplicity, we assume that there are two groups of individuals holding two different opinions on an issue. The opinion of an agent i, o_i, is given by either 1 or 2, depending on the group she belongs to. G_1 is the group of agents holding opinion 1, G_2 the one of agents holding opinion 2. Agents are connected to each other according to probabilities of the stochastic block matrix M (the entries q_{11}, q_{22} and q_{12} in the different blocks represent the probability of every connection within that block, self-connections excluded),

$$M = \left(\begin{array}{c|c} q_{11} & q_{12} \\ \hline q_{12} & q_{22} \end{array} \right). \tag{2.1}$$

In each interaction step, an undirected, unweighted network is generated from M, for which the probability of there being an edge between any two agents belonging to opinion group G_1 is given by q_{11}, and analogously q_{22} for G_2. Cross-group connection probabilities are given by q_{12}. Since they are probabilities, $q_{11}, q_{22}, q_{12} \in [0,1]$.

We can express the expected fraction of neighbors that hold the same opinion as an agent by[3]

$$f_{11} = \frac{(N_1 - 1)q_{11}}{(N_1 - 1)q_{11} + N_2 q_{12}} \tag{2.2}$$

for agents belonging to opinion group G_1 and

$$f_{22} = \frac{(N_2 - 1)q_{22}}{(N_2 - 1)q_{22} + N_1 q_{12}} \tag{2.3}$$

for agents that are part of opinion group G_2. The expected fractions of neighbors belonging to the other opinion group are consequently

$$f_{12} = \frac{N_2 q_{12}}{(N_1 - 1)q_{11} + N_2 q_{12}} \tag{2.4}$$

for agents of G_1 and

$$f_{21} = \frac{N_1 q_{12}}{(N_2 - 1)q_{22} + N_1 q_{12}} \tag{2.5}$$

[3]Note here that these are the fractions of neighbors with a certain *internal* opinion. Whether these opinions are also visible to others will be subject of the next section.

for agents of G_2. We now introduce two central structural parameters, γ and δ. They are the ratios of the expected in-group to the out-group connections for each opinion group and given by

$$\gamma = \frac{N_1 - 1}{N_2} \cdot \frac{q_{11}}{q_{12}} \tag{2.6}$$

and

$$\delta = \underbrace{\frac{N_2 - 1}{N_1}}_{\text{group sizes}} \cdot \underbrace{\frac{q_{22}}{q_{12}}}_{\text{probabilities}}, \tag{2.7}$$

$\gamma > 1$ or $\delta > 1$ means that the agents of one opinion group on average have more connections to others that hold the same opinion, while $\gamma < 1$ or $\delta < 1$ indicates that agents of the opinion group are more strongly connected to agents holding a different opinion. In the following, if we say that an opinion group is internally well-connected, we mean that the structural parameter of the group is bigger than 1. With γ and δ, the above fractions (2.2), (2.3), (2.4) and (2.5) can be simplified to

$$f_{11} = \frac{\gamma}{\gamma + 1}, \qquad f_{12} = \frac{1}{\gamma + 1}, \tag{2.8}$$

$$f_{22} = \frac{\delta}{\delta + 1}, \qquad f_{21} = \frac{1}{\delta + 1}. \tag{2.9}$$

2.3 A silence game

We use the social structure described in section 2.2 as the setting of a 'silence game,' in which the opinions of the agents are fixed according to their group affiliation and do not change. Game-theoretic approaches to opinion dynamics provide an interesting avenue for research since certain cognitive mechanisms, such as the minimization of cognitive dissonance, can be modelled by utility maximization (see also [12, 13]). Therefore, in some scenarios, it might be insightful to formulate opinion dynamics models as games. This is even more evident with respect to opinion *expression* dynamics, since expression constitutes a conscious action rather than an (often partly unconscious) process like opinion formation. Especially in online environments, where opinion expression is facilitated and can be rewarded almost instantly, the effect of different collective incentive structures can be examinded with game theory.[4]

Game theory is a methodology to model and analayse situations of interactive decision making that involves several participants (called players), in which the decision of

[4]The argument for incentive-based opinion dynamics sketched here parallels (at least to some extent) Robert Lucas' critique of econometric models that assume invariant decision rules under changing government policy [91]. Lucas argues that "given that the structure of an econometric model consists of optimal decision rules of economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models" [91, p. 41]. Effects of regulations (in the case at hand here, this might include e.g. algorithmic changes of social media platforms) can barely be evaluated if a behavioral mechanism is already assumed beforehand. But if decisions of individuals are guided by incentives that are influenced by such regulations, such an evaluation is (in principle) possible.

each player affects the outcome of all others. It is used to predict the outcomes of collective interactions based on the assumption that players seek to maximize their utility. A normal-form game is an ordered triple $G = (N, (A_i)_{i\in N}, (u_i)_{i\in N})$. N is a finite set of players. For each player i, A_i describes the set of all possible actions from which the player can choose. The set of all vectors of actions is denoted by $\mathbf{A} = A_1 \times A_2 \times ... \times A_n$, with typical element $\mathbf{a} = (a_i)_{i\in N}$. $u_i : \mathbf{A} \to \mathbb{R}$ is a function associating each vector of actions with the payoff (or utility) $u_i(\mathbf{a})$ to player i. Generally, strategies s_i are probablity distributions over the actions, taken from $S(A_i)$, the set of probability distributions over the action set A_i. A strategy profile $\mathbf{s} = (s_1, s_2, ..., s_n)$ is a vector of strategies with one strategy for each player. A strategy is called pure if $s_i(a_i) = 1$ for some action $a_i \in A_i$ and 0 for all other actions, otherwise it is called a mixed strategy. The expected utility[5] of a strategy profile for player i, $u_i(\mathbf{s})$, is given by

$$u_i(\mathbf{s}) = \sum_{\mathbf{a}\in\mathbf{A}} u_i(\mathbf{a}) \prod_{j=1}^{n} s_j(a_j). \tag{2.10}$$

A solution concept of a game is a formal rule which governs what strategies are adopted by players. Different solution concepts exist, out of which the most prominent one is constituted by the *Nash equilibrium*. A Nash equilibrium (NE) is definded as follows: A strategy profile $\mathbf{s}^* = (s_1^*, ..., s_n^*)$ is a NE if for each player $i \in N$ and each strategy $s_i \in S(a_i)$,

$$u_i(\mathbf{s}^*) \geq u_i(s_i, \mathbf{s}^*_{-i}). \tag{2.11}$$

The payoff vector $u(\mathbf{s}^*)$ is the equilibrium payoff corresponding to the NE $\mathbf{s}^*$, and $\mathbf{s}^*_{-i}$ denotes a strategy selection for all players but i.

The silence game consists of N players, for which each player can choose an action from $A_i = \{e, s\}$. e refers to public expression of opinion, and s to silence. Expected utilities of a player are the sum of (expected) payoffs earned from pairwise interaction with each neighbor on the network drawn according to Sec. 2.2. These local interactions are governed by the payoff matrices given in Table 2.1 and depend on the opinion groups the two agents belong to.[6] If two expressive agents are paired that share an opinion, they are rewarded with a payoff of $1 - c$. The constant c accounts for the costs of opinion expression – we may think of the effort of writing a reply to someone in social media, or the effort of joining a demonstration for or against some issue. On the other hand, if two disagreeing players who express their opinion meet, payoff is $-1 - c$. If one of the two is silent, and the other one expressive, the expressive player is left with the costs of expression $-c$; and if both are silent, payoff for both amounts to 0.

In the following paragraphs, we will first restrict ourselves to pure strategies and their NEs.[7] The expected utility of an individual i choosing to express herself is given

[5] We are overloading notation here by calling both the utility and the expected utility u_i.

[6] Games of this general type have been referred to as 'multiple-group games,' e.g. in [105] and [104]. [104] contains a more formal definition of such a setting in Def. 7. In [105], it is noted that also the contributions of [49, 38, 126] describe multiple populations interacting together, although not in similar detail.

[7] Mixed strategies are only metastable in the sense that it only takes strategy change by one agent to make it favorable for all players to adopt a pure strategy, as will be explained in a paragraph below.

Table 2.1: Payoffs depending on group affiliation of the players. For G^{11} and G_{22}, the players A and B are of the same opinion group, while for G^{12}, they are from different groups.

G^{11}, G^{22}

		A: e	A: s
B	e	$1-c, 1-c$	$-c, 0$
	s	$0, -c$	$0, 0$

G^{12}

		A: e	A: s
B	e	$-1-c, -1-c$	$-c, 0$
	s	$0, -c$	$0, 0$

as follows:

$$u_i(e, \mathbf{a}_{-i}) = \sum_{\substack{j \in G_1 \\ j \neq i \\ a_j = e}} q_{11} - \sum_{\substack{j \in G_2 \\ a_j = e}} q_{12} - c' \tag{2.12}$$

if $i \in G_1$, with $c' = c((N_1 - 1)q_{11} + N_2 q_{12})$, and

$$u_i(e, \mathbf{a}_{-i}) = \sum_{\substack{j \in G_2 \\ j \neq i \\ a_j = e}} q_{22} - \sum_{\substack{j \in G_1 \\ a_j = e}} q_{12} - c'' \tag{2.13}$$

if $i \in G_2$, with $c'' = c((N_2 - 1)q_{22} + N_1 q_{12})$. In the above equations, we compare the expected number of neighbors of an agent i who *publicly* (i.e. they choose action $a_j = e$ with probability 1) agree with i with the expected number of publicly disagreeing neighbors (and costs of opinion expression). The utility for silence is always 0, no matter the action of the other players. Put differently, the payoffs for opinion expression depends on the players' social environments. If they are surrounded by predominantly disagreeing others, they prefer to be silent. If not, they will have an incentive to express their opinion. But only the expressive agents shape the subjective impression of the opinion landscape of each individual. Silent ones do not contribute – after all, silence implies that the individual's opinion is not public.

An agent i prefers e over s if on average, in i's neighborhood, more agents who share i's opinion speak out (if costs of expression are set to 0 – otherwise, they need to be compensated for, as well). The condition for e being strictly preferred to s for a player i can then be written as

$$\frac{\sum_{\substack{j \in G_1 \\ j \neq i \\ a_j = e}} q_{11}}{(N_1 - 1)q_{11} + N_2 q_{12}} > \frac{\sum_{\substack{j \in G_2 \\ a_j = e}} q_{12}}{(N_1 - 1)q_{11} + N_2 q_{12}} + c \tag{2.14}$$

if i is part of opinion group G_1 and

$$\frac{\sum_{\substack{j \in G_2 \\ j \neq i \\ a_j = e}} q_{22}}{(N_2 - 1)q_{22} + N_1 q_{12}} > \frac{\sum_{\substack{j \in G_1 \\ a_j = e}} q_{12}}{(N_2 - 1)q_{22} + N_1 q_{12}} + c \tag{2.15}$$

if $i \in G_2$. Here, the terms are normalized by the expected overall number of neighbors of the agent.[8]

If the respective inequality is fulfilled for a player, she prefers to speak out. If the two sides of (2.14) or (2.15) are equal, the individual is indifferent in her preference over the actions.

In the system, the equilibrium condition is met if there is a strategy profile for which each individual that expresses herself has Eq. (2.14) or Eq. (2.15) (depending on the opinion group of the agent) satisfied, and if for each individual that is silent, the corresponding inequality is not fulfilled.

It is already visible in Eqs. (2.14) and (2.15) that apart from the fact that an individual does not account for her own expressed opinion in the inequality ($i \neq j$ in the sum on the left-hand side), the rest of the contributions in the inequalities are the same for all agents of one opinion group. It is also visible that if (2.14) or (2.15) is satisfied for an agent i that expresses herself, it must be satisfied for all silent individuals of her group as well: They 'see' one more agent expressing their opinion than i, since i does not account for herself in her evaluation of her environment. Hence, there is an additional positive term on their left-hand side. On the other hand, if the inequality is not fulfilled for a silent agent of one group, it can neither be fulfilled for an expressive one. Therefore, in a pure-strategy equilibrium, all agents of one opinion group must choose the same action.

This simplifies the inequalities above. If all agents of an opinion group act the same, Eqs. (2.14) and (2.15) can be expressed in terms of the structural parameters γ and δ. Four pure-strategy NEs might be possible, depending on γ and δ. Both groups can be silent, or only one of them, but not the other, or none:

- If both groups express their opinion (we call this state (e,e); the first entry stands for the collective action of G_1, the second for the action of G_2), the following conditions must be satisfied to make this state a NE:

$$\frac{(N_1-1)q_{11}-N_2q_{12}}{(N_1-1)q_{11}+N_2q_{12}}-c=\frac{\gamma-1}{\gamma+1}-c>0,^{9} \tag{2.16}$$

$$\frac{(N_2-1)q_{22}-N_1q_{12}}{(N_2-1)q_{22}+N_1q_{12}}-c=\frac{\delta-1}{\delta+1}-c>0. \tag{2.17}$$

- (e,s) is a NE if

$$\frac{(N_1-1)q_{11}}{(N_1-1)q_{11}+N_2q_{12}}-c=\frac{\gamma}{\gamma+1}-c>0, \tag{2.18}$$

$$-\frac{N_1q_{12}}{(N_2-1)q_{22}+N_1q_{12}}-c=-\frac{1}{\delta+1}-c<0. \tag{2.19}$$

[8] The reason for this normalization will become apparent in equations (2.16)-(2.22): We can then express the conditions for the pure-strategy Nash equilibria in terms of γ and δ.

[9] We use equation (2.6) in the equivalence.

- (s,e) is a NE if

$$-\frac{N_2 q_{12}}{(N_1-1)q_{11}+N_2 q_{12}} - c = -\frac{1}{\gamma+1} - c < 0, \tag{2.20}$$

$$\frac{(N_2-1)q_{22}}{(N_2-1)q_{22}+N_1 q_{12}} - c = \frac{\delta}{\delta+1} - c > 0. \tag{2.21}$$

- (s,s) is a NE if

$$c > 0. \tag{2.22}$$

The different existence regimes of the pure-strategy NEs are given in Fig. 2.1. Silence is always a Nash equilibrium for non-negative expression costs. If γ and δ are both smaller than $\frac{c}{1-c}$, then even if all group members express their opinion and the other opinion group is silent, it is too costly (compared to the amount of connections to agents of the own opinion group) to express one's opinion and the only NE is the one in which all individuals are silent. If γ or δ or both are bigger than $\frac{c}{1-c}$, but smaller than $\frac{c+1}{1-c}$, either both opinion groups are silent or one of the groups expresses themselves, but not both: The strength of internal connections of each group are not sufficient to account for the negative influence of the other, expressive group. Not both Eq. (2.16) and Eq. (2.17) can be satisfied. Hence, this structural regime only allows public opinion predominance of one group (or complete silence).[10] If γ and δ are both bigger than $\frac{c+1}{1-c}$, it is possible that both opinion groups express their opinon publicly at the same time. Then, the positive influence of the in-group members still dominates, even if all out-group members are expressive as well. Hence, also Eqs. (2.16) and (2.17) are satisfied. Note that even though all players of one opinion group need to adopt the same action in a Nash equilibrium, the situation is different than in a simple two-player game (even if we allow a player to interact with herself). Since one player represents a whole group then, if she changes strategy, the whole group would do so. Then, (e,e) is the only possible NE for γ and δ larger than $\frac{1+c}{1-c}$. For γ and δ larger than $\frac{c}{1-c}$, (e,s) or (s,e) are NE, and if only one of the two structural parameters fulfills that condition, the corresponding group is expressive. If γ and δ are smaller than $\frac{c}{1-c}$, only (s,s) is a NE.

Obviously, there are also mixed-strategy NEs. Suppose the situation is as follows: The agents of each group mix their actions uniformly such that each agent is exactly indifferent between expressing herself or staying silent. Then, no one has an incentive for action change, and we therefore have a NE. This equilibrium is, nevertheless, just metastable in the sense that it takes only one agent to increase (or decrease) her expression probability in order to make it favourable for all other agents of one opinion group to express themselves (or become silent).

γ and δ do not only depend on the number of agents holding one or another opinion. They are also influenced by the internal connection probabilities of agents of one opinion group. Hence, a well-connected minority group can dominate public discourse

[10] If either only $\gamma > \frac{c}{1-c}$ or $\delta > \frac{c}{1-c}$ is satisfied, it is clear which opinion will dominate publicly (if any). If both are satisfied, the situation becomes more interesting in the sense that it depends on the initial conditions and the dynamical development of the system which opinion will predominate. We will approach these issues in sections 2.4 and 2.5.

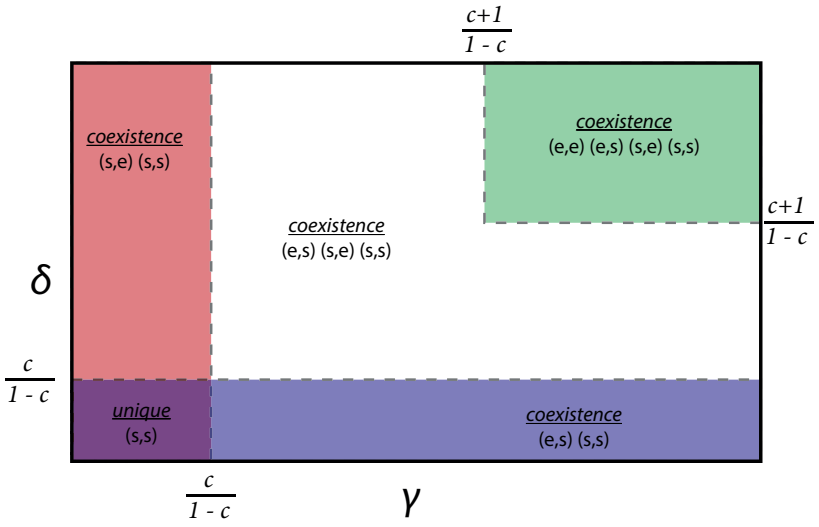


Figure 2.1: The available pure-strategy Nash equilibria in different regimes of γ and δ. The equilibria are abbreviated by either *e* for expression or *s* for silence for each opinion group (the first entry is for the collective action of G_1, the second for the one of G_2). For costs $c > 0$, γ and δ below $\frac{c}{1-c}$ will lead to a situation in which the only available Nash equilibrium is one in which no one expresses her opinion publicly. An increase in the structural parameters above this threshold leads to additional Nash equilibria in which at least one of the two opinion groups speaks out. If both γ and δ are bigger than $\frac{c+1}{1-c}$, an additional Nash equilibrium arises in which all agents express their opinion.

if the corresponding structural parameter is above the threshold of $\frac{c}{1-c}$. But while the regimes of different NEs in Fig. 2.1 are displayed correctly, it might give the impression that γ or δ are parameters that can be tuned by simply increasing the probability of a connection between two agents of the same group, that is, q_{11} or q_{22} (all other parameters, including q_{12}, fixed). This is not the case. Some numerical minorities cannot be balanced by increasing internal connections since q_{11} and q_{22} are bounded by 1. If there are too few agents in one opinion group, even setting q_{11} or q_{22} to 1 will not be elevate γ or δ above a certain threshold. This is made visible in Fig. 2.2. The figure shows the different existence regimes of the NEs for different combinations of internal connection weights q_{11} and q_{22} and partitions of a total of $N = 100$ agents between groups G_1 and G_2. q_{12} and c are fixed. Each point in the plot stands for a combination of the number of agents in opinion group G_1, N_1, and the in-group connection probability q_{11}, out of which one can compute the value of γ. The lines of constant γ are plotted in red. Since the overall number of agents $N = 100$ is fixed, N_2 is not independent and determined by the choice of N_1 by $N - N_1$. If we just assume that $q_{22} = q_{11}$, each point in the plot at the same time represents also a combination of the relevant parameters of opinion group G_2 out of which one can compute δ. Curves of constant δ are the blue lines and symmetrical to the γ-curves with respect to $N_1 = 50$.

A vertical line in the plot, e.g. at $N_1 = 20$, can be interpreted as follows: Each constant γ or δ value that it intersects on its way to $q_{11} = q_{22} = 1$ is reachable for this partition of agents in the two groups if q_{11} and q_{22} are tuned accordingly. But if there is no intersection for a specific γ or δ, then even if the internal connection probabilities are maximized (i.e. one opinion group is completely connected internally), the structural strength of the respective group cannot reach that value due to their limited group size. For $N_1 = 20$, a state in which both opinion groups are expressing themselves (the upper right, green area in Fig. 2.1) cannot be reached since opinion group G_1 has too few agents to produce a γ high enough to satisfy Eq. (2.16). In general, there are numerical thresholds (dependent on the costs c, the cross-group connection probability q_{12} and the overall number of agents N) below which reaching a state in which both group express themselves or in which the own group becomes dominant becomes impossible from a game-theoretic perspective. The game-theoretic approach hence comes with (all other parameters fixed) limits for the effect of group-internal coordination in the form of internal cohesion on public discourse.

2.4 Q-learning and a dynamical systems perspective

The analysis conducted so far does not answer questions of equilibrium selection or stability of equilibria. In this section, we will introduce a dynamical systems perspective to approach these issues. To this end, we describe the development of the system as reinforcement learning dynamics. Reinforcement learning is based on the idea that an action that is followed by a satisfactory outcome strengthens the inclination to repeat that action [142]. In contrast to the game-theoretical setting, agents do not have full access to the payoff matrix. Agents are taken to be players in a normal form game, which is played repeatedly, in order to improve their strategy over time [116]. Reinforcement learning can be utilized to solve Markov decision processes, which model a sequential

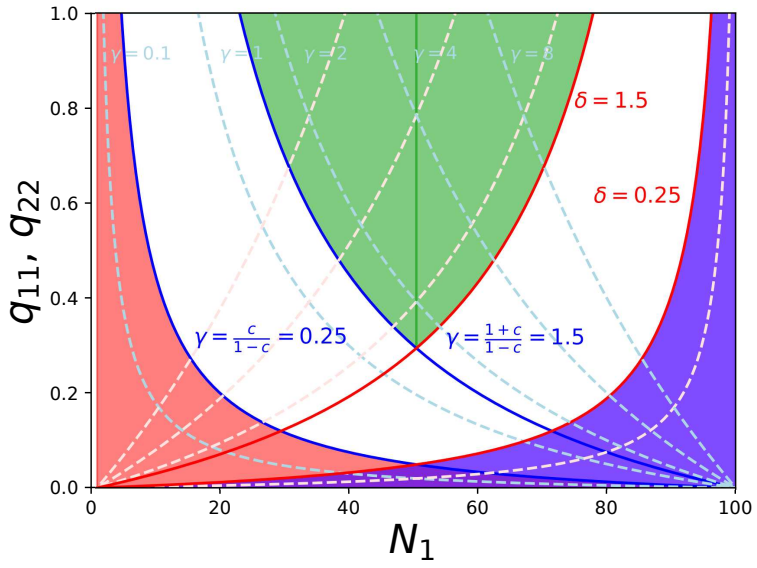


Figure 2.2: The constant γ- and δ-curves for $N = 100$ agents, $q_{12} = 0.2$ and $c = 0.2$. They are plotted with respect to N_1, $N_2 = N - N_1$, and $q_{11} = q_{22}$. Each blue curve (starting at $N_1 = 100$, $q_{11} = 0$) stands for a combination of the number of opinion group members N_1 and internal connection weights q_{11} that yields a constant value of the structural parameters γ, each red one (starting at $N_1 = 0$, $q_{11} = 0$) for a combination of N_2 and q_{22} that produces constant δ. The color-coding for the different Nash equilibrium regions is analogous to Fig. 2.1. It is visible that the numerical minority of an opinion group cannot always be compensated by increasing q_{11} (or q_{22}), the probability of a connection between two agents of the same opinion group. Moreover, the fixed γ- and δ-curves are symmetric with respect to $N_1 = N_2 = 50$, where they intersect. (For better readability, the dashed δ-curves have not been labelled. They correspond to their γ-counterparts.)

decision-making problem where the state transition and reward function depend only on the current state of affairs and the applied action [159].

More specifically, in this section we will investigate dynamics induced by Q-learning, a form of model-free reinforcement learning [159]. There, the players' strategies are represented by Q-functions that characterize relative utility of a particular action.[11] Players then make use of these estimates to select strategies. Reinforcement learning shares many objectives and assumptions with evolutionary game theory: There, a game is played among a population not only once, but repeatedly. Players are uninformed of the preferences of the opponent players. An evolutionary stable strategy is a strategy which, upon being adopted by the population, cannot be replaced by another, more successful strategy initially played only by a small fraction of agents. One way of modelling a dynamical process in which the (expected) proportions of various strategies in a population evolve is given by replicator dynamics. A variety of publications deal with the connection of replicator dynamics to different types of reinforcement learning. For an overview, see e.g. [20]. A specific connection to Q-learning has been made in [151], where continuous-time Q-learning is decomposed into two terms: One that gives an selection mechanism in the form of replicator dynamics (i.e. exploitation), and an additional term that introduces exploration of new strategies.

In Q-learning, the reinforcement mechanism that updates the agent's Q-value of an action (in this case: willingness to express her opinion) is given by [141, 142]

$$Q_i^{t+1} = (1-\alpha)Q_i^t + \alpha r_i^t, \tag{2.23}$$

where r_i^t is the reward for agent i at time step t upon expression. The reward depends on who the agent interacts with. Interaction is pairwise and takes place on the network (drawn again from M in each time step) of section 2.2. Payoffs for opinion expression can be found in the local interactions sketched in Table 2.1, and are given by

$$r_i^t = \begin{cases} -c & \text{for random neighbor being silent,} \\ -1-c & \text{for disagreeing random neighbor,} \\ 1-c & \text{for agreeing random neighbor.} \end{cases} \tag{2.24}$$

The Q-function is expected to converge to the expected reward for public opinion expression over time. (2.23) describes Q-learning for myopic agents, i.e. with discount factor 0. Moreover, we can restrict ourselves to one Q-function per agent, which estimates the reward for opinion expression, since the reward for silence is always 0. The probability of expression is a function of the value of Q_i. Simply choosing the action with the highest Q-value might generally lead to globally suboptimal solutions. Thus, one needs to incorporate some way of exploring less-optimal strategies [80]. Here, we assume a Boltzmann action selection mechanism, i.e. the probability of expression of agent i is given by

$$p_i^t = \frac{1}{1+e^{-\beta Q_i^t}}, \tag{2.25}$$

and the probability of staying silent by $1-p_i^t$. If $\beta = 0$, the action choice of the agent is completely independent of the Q-values and randomized. For increasing β, the agent

[11] [13] provides a more detailed justification for this choice including evidence from neuroscience.

becomes more sensitive in her action selection towards her current evaluation of her local opinion environment. Then, a positive Q-value indicates that it is more likely for her to express herself than not, while a negative one indicates the opposite. If $\beta \to \infty$, the probabilities of the actions become deterministic.

In the case under investigation here, the expected reward for agent i upon opinion expression is given by either (if i belongs to opinion group G_1)

$$\begin{aligned}\mathbb{E}_p[r_i^t] = \quad & -c + f_{11}\frac{1}{N_1 - 1}\sum_{\substack{j\in G_1\\ j\neq i}}\frac{1}{1+e^{-\beta Q_j^t}} - \\ & f_{12}\frac{1}{N_2}\sum_{j\in G_2}\frac{1}{1+e^{-\beta Q_j^t}},\end{aligned} \tag{2.26}$$

or (if i belongs to opinion group G_2)

$$\begin{aligned}\mathbb{E}_p[r_i^t] = \quad & -c + f_{22}\frac{1}{N_2 - 1}\sum_{\substack{j\in G_2\\ j\neq i}}\frac{1}{1+e^{-\beta Q_j^t}} - \\ & f_{21}\frac{1}{N_1}\sum_{j\in G_1}\frac{1}{1+e^{-\beta Q_j^t}}.\end{aligned} \tag{2.27}$$

We follow [80], where Q-learning in two-player two-action games is investigated, and take the continuous-time limit of the Q-learning equation (2.23). In this limit, we divide time into intervals of δt. We replace $t+1$ with $t+\delta t$ and α with $\alpha'\delta t$. This yields

$$Q_i(t+\delta t) - Q_i(t) = \alpha'\delta t(r_i(t) - Q_i(t))$$

and hence

$$\dot{Q}_i = \alpha'(r_i(t) - Q_i(t)). \tag{2.28}$$

Over time, the difference of the largest and the lowest Q-value of an opinion group decays at least exponentially in expectation (see Appendix A for the estimation):

$$\frac{d}{dt}(Q_{i\in G_1}^{\max} - Q_{i\in G_1}^{\min}) \leq -\alpha'(Q_{i\in G_1}^{\max} - Q_{i\in G_1}^{\min}),$$

$$\frac{d}{dt}(Q_{i\in G_2}^{\max} - Q_{i\in G_2}^{\min}) \leq -\alpha'(Q_{i\in G_2}^{\max} - Q_{i\in G_2}^{\min}).$$

That is, the Q-values of the agents of one group are expected to converge over time. This allows us to employ a mean-field approximation for the expected reward of the two opinion groups: We introduce the average Q-values for each opinion group[12]

$$Q_1(t) = \frac{1}{N_1}\sum_{i\in G_1} Q_i(t), \qquad Q_2(t) = \frac{1}{N_2}\sum_{i\in G_2} Q_i(t). \tag{2.29}$$

[12]Note the slight abuse of notation here: From now on, the index of Q and p will not indicate single individuals any more, but the average Q-value and the corresponding expression probability of the different opinion *groups*.

This means that we do not distinguish any more between the agents of the respective opinion groups. We assign them the average of their group's Q-value. This simplification will have an effect on the probability of opinion expression for the individuals. Instead of averaging over each group's probability of expression, we simply insert the averaged Q-values into the equation:

$$\frac{1}{N_1}\sum_{j\in G_1}\frac{1}{1+e^{-\beta Q_j(t)}} \longrightarrow \frac{1}{1+e^{-\beta Q_1(t)}} = p_1(t), \tag{2.30}$$

$$\frac{1}{N_2}\sum_{j\in G_2}\frac{1}{1+e^{-\beta Q_j(t)}} \longrightarrow \frac{1}{1+e^{-\beta Q_2(t)}} = p_2(t). \tag{2.31}$$

The expected rewards for the different opinion groups are given by the equations[13]

$$\mathbb{E}_p[r_1(t)] = -c + \frac{\gamma}{\gamma+1}p_1(t) - \frac{1}{\gamma+1}p_2(t), \tag{2.32}$$

$$\mathbb{E}_p[r_2(t)] = -c + \frac{\delta}{\delta+1}p_2(t) - \frac{1}{\delta+1}p_1(t), \tag{2.33}$$

where the probabilities of expression for each group are $p_1(t)$ and $p_2(t)$, and it is not distinguished any more between the individuals.

We can therefore write a two-dimensional approximation of the system as follows:

$$\dot{Q}_1(t) = \alpha'(-c + \frac{\gamma}{\gamma+1}p_1(t) - \frac{1}{\gamma+1}p_2(t) - Q_1(t)), \tag{2.34}$$

$$\dot{Q}_2(t) = \alpha'(-c + \frac{\delta}{\delta+1}p_2(t) - \frac{1}{\delta+1}p_1(t) - Q_2(t)). \tag{2.35}$$

According to equations (2.34) and (2.35), we can produce a phase portrait of the system including its trajectories and fixed points for given exploration rate β, structural parameters γ and δ, and costs of expression c. An example of how the phase portraits change with γ and δ is given in Fig. 2.3. There, it is visible that the stable fixed points of the system include basins of attraction, that is, regimes of values of Q_1 and Q_2 for which the system is expected to end up in those fixed points.[14] The basins of attraction in the two-dimensional approximation correspond exactly to those of the stochastic N-agent system in the limit $\alpha \to 0$. For larger α, both fixed points and basins of attraction do not necessarily correspond to the two-dimensional approximation. We show averages over simulation runs for different values of α in Fig. 2.4.

[13] f_{11}, f_{12}, f_{21}, and f_{22} have been replaced according to equations (2.8) and (2.9) with $\frac{\gamma}{\gamma+1}$, $\frac{1}{\gamma+1}$, $\frac{\delta}{\delta+1}$, and $\frac{1}{\delta+1}$.

[14] Hence, when we speak of a stable fixed point here we mean that the *average* Q-values of the opinion groups can be subjected to small variations and the system is still expected to end up in these fixed points.

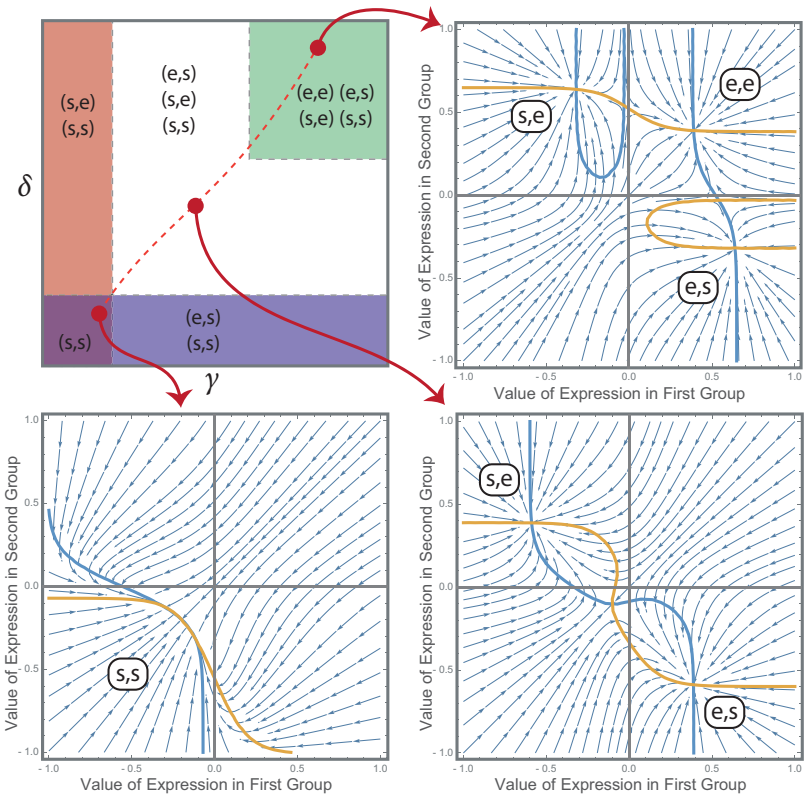


Figure 2.3: Three phase portraits of the Q_1- (x-axis) and Q_2-values (y-axis) of the two-dimensional system for different configurations of γ and δ. We have $c = 0.1$, $\beta = 10$, and structural parameters $\gamma = \delta = 0.1$ (bottom left), $\gamma = \delta = 1$ (bottom right), and $\gamma = \delta = 3$ (top right). The yellow and blue lines in the phase portraits are the isoclines of the equations for Q_1 and Q_2. The fixed points are located at their intersections. As is visible, the (s,s) fixed point disappears in the dynamical system for higher γ- and δ-values. This is due to the finite exploration rate β and the transition from the N-player game of section 2.3 to the two-population game in the mean-field approximation.

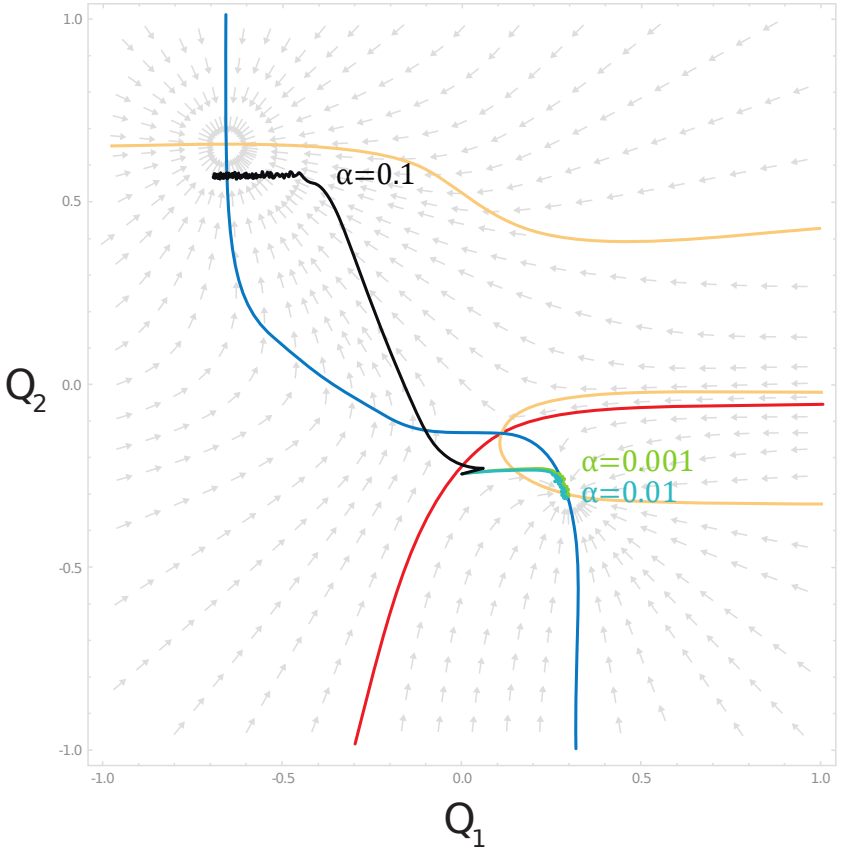


Figure 2.4: The trajectories of the Q-values in simulations, averaged over 50 runs with $N \cdot 10^5$ steps, for different values of α with a starting point close to the border (red line) of the two basins of attraction of the two stable fixed points. The starting Q-values were $Q_{i\in G_1} = 0$, $Q_{i\in G_2} = -0.25$. There were $N = 200$ agents, 100 of each opinion group, and $c = 0.1$, $q_{11} = 0.04$, $q_{12} = 0.05$, and $q_{22} = 0.15$. A relatively big $\alpha = 0.1$ makes the trajectory leave the lower right basin of attraction of the two-dimensional system (black trajectory). Due to the high α, the fixed point of the other basin of attraction is also missed by some margin. The lower α, the closer the trajectories get to the fixed point and the more probable it is that they will stay in the basin predicted by the two-dimensional approximation. For $\alpha = 0.01$ (turquoise) and $\alpha = 0.001$ (light green), the trajectories run towards the predicted fixed point. The yellow (light gray) and blue (dark gray) lines are the isoclines of the equations for Q_1 and Q_2. The fixed points are located at their intersections.

2.5 Bifurcation and stability analysis

In order to find the fixed points of Q_1 and Q_2, we set (2.34) and (2.35) to 0, solve (2.34) for Q_2 and insert it into (2.35), which yields:

$$Q_2 = -\frac{1}{\beta}\ln\left(\frac{1}{\frac{\gamma}{1+e^{-\beta Q_1}} - (\gamma+1)(Q_1+c)} - 1\right) \tag{2.36}$$

$$\begin{aligned}\frac{\delta}{\delta+1} \quad & \left(\frac{\gamma}{1+e^{-\beta Q_1}} - (\gamma+1)(Q_1+c)\right) - \frac{1}{\delta+1}\frac{1}{1+e^{-\beta Q_1}} + \\ & \frac{1}{\beta}\ln\left(\frac{1}{\frac{\gamma}{1+e^{-\beta Q_1}} - (\gamma+1)(Q_1+c)} - 1\right) - c = 0\end{aligned} \tag{2.37}$$

Equation (2.37) now gives us the Q_1-value of the fixed points of the system, with which we can calculate the corresponding Q_2-value by equation (2.36). In essence, the fixed points depend on four parameters: β, γ, δ, and c. We will carry out a bifurcation analysis of the latter three parameters in the following subsections, β-bifurcations can be found in Appendix B.

After having solved equations (2.37) and (2.36) for Q_1 and Q_2, we can assess the stability of the respective fixed points by calculating the eigenvalues of their Jacobian; two negative (real parts of the) eigenvalues indicate a stable attractor. In the following, we analyze the bifurcation structure of the system depending on the different types of parameters in the system.

2.5.1 Structural power

The parameter γ describes the ratio of expected internal versus external connections of G_1. $\gamma > 1$ means that on average each member of G_1 is connected to more agents of the own than of the other opinion group. (Everything stated in this paragraph applies equivalently to δ, which is just the parameter for the ratio of internal versus external connections of the other group.)

As is visible in Fig. 2.5, for small γ (< 0.5), given $\beta = 10$, $\delta = 2.36$ (that is, a quite well-connected opposite opinion group) and $c = 0.1$, there is only one (stable) fixed point with negative Q_1-value and positive Q_2. While γ grows, a saddle-node bifurcation occurs such that one stable and one unstable fixed point appear for positive Q_1 and negative Q_2. Another saddle-node bifurcation occurs at around $\gamma = 2$; and for $\gamma > 4.2$, the low-Q_1 fixed points disappear in another saddle-node.

How can this be interpreted? In essence, an opinion community that is not well-connected internally ($\gamma < 0.5$) will be driven into silence by the opposite opinion group that is internally more cohesive. With increasing γ, that is, increasing internal connectedness, other fixed points appear in which the former group is expressive.[15] With a

[15] To be precise, the Q-values here are only indicative of probabilities of opinion expression according to the Boltzmann action selection which depends on Q. If Q is smaller than 0, the probability of expression is smaller than the probability of staying silent. In the following, if we say that one opinion group is expressive, we mean that they have a Q-value bigger than 0 which makes their probability of expression higher than that of silence.

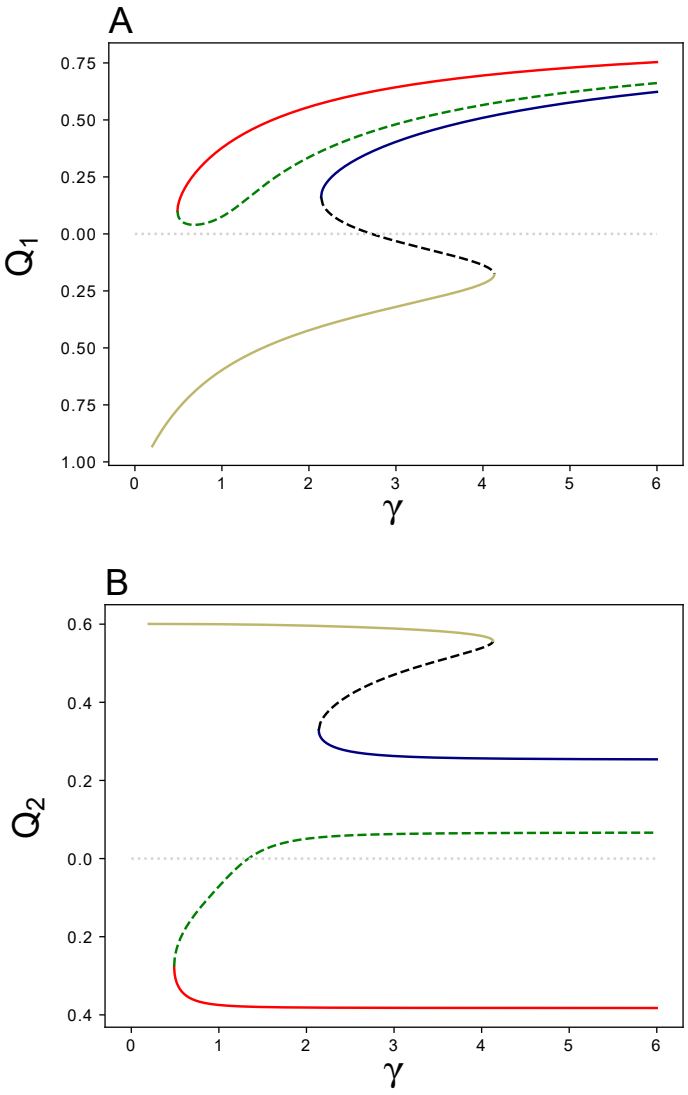


Figure 2.5: The development of the Q_1- (A) and Q_2-value (B) of the fixed points with γ given $\beta = 10$, relatively high $\delta = 2.36$, and $c = 0.1$. The colors of the curves in the two plots indicate the different fixed point pairs of Q_1 and Q_2. A dashed line indicates an unstable fixed point, a continuous one a stable fixed point. It is visible in the plots that a poorly connected opinion group G_1 ($\gamma < 0.5$) will be driven into silence by the other group (beige curve, lowest one in (A), highest one in (B)). With increasing in-group connectivity, fixed points arise for which G_1 expresses their opinion in two saddle-node bifurcations (red for an an (e,s)-equilibrium (highest curve in (A), lowest in (B)) and blue for (e,e) (in-between)). For $\gamma > 4.5$, G_1 is so well-connected that the equilibrium disappears in which the group is silent. The dotted grey line indicates Q-value 0, where the probability of expression passes 0.5.

further increase of γ, G_1 even becomes too cohesive to be driven into silence by the other group: Either the first opinion group is 'loud' alone or both groups express their opinions. Increased internal cohesion of one opinion group can hence have the effect that this group, which is not necessarily a majority, will dominate public discourse.

A lower δ-value (e.g. $\delta = 1.6$) leads to a reduction in available fixed points (Fig. 2.6) such that only two saddle-node bifurcations occur and at high γ only one fixed point remains in which the first opinion group is expressive.

2.5.2 Costs

The costs for opinion expression have a profound impact on the fixed points of the system. If opinion expression is very expensive, (in Fig. 2.7: $c > 0.4$), there is only one fixed point in the system for which both opinion groups stay silent. For decreasing costs, two pairs of fixed points arise in a saddle-node bifurcation. Each of the pairs corresponds to a situation in which one opinion group is expressive, while the other is silent (in Fig. 2.7, we have identical values for γ and δ). The fixed point in which both opinion groups are silent becomes unstable with decreasing c in a pitchfork bifurcation. Below $c = 0.1$, another pitchfork bifurcation arises for which the stable fixed point now corresponds to a state in which both groups are expressing their opinion. Costs can also be negative: Then, the individuals might be intrinsically motivated or externally encouraged to speak out.[16] For sufficiently negative costs (in the case of Fig. 2.7: $c < -0.05$), only one fixed point exists: Everyone has an incentive to speak out, at least for internally well-connected opinion groups. The fixed points for which only one of the groups is expressive disappear in two saddle-nodes.

2.5.3 Asymmetric costs

The model allows us to also assign different costs to each opinion group, such that $c_1 \neq c_2$. Internal motivation for a cause, for example, can be an incentive to speak out and might even be indicated by negative costs (that is, an urge to express one's opinion). Moreover, there might be biases in the infrastructures on which debate takes place such that it takes more effort for one group to speak out than for the other.[17]

The bifurcation in Fig. 2.8 (for the case of two internally well-connected opinion groups) illustrates the effect that different expression costs in the populations exhibit on public discourse. In Fig. 2.8, a bifurcation over c_1 is shown. Negative costs for opinion expression in opinion group G_1 yield two stable equilibria in which opinion group 1 is expressive, either together with opinion group G_2 or alone. With increasing costs, a stable fixed point arises in a saddle node for which G_1 is silent (at $c_1 \approx 0$), while G_2 is expressive. At $c_1 \approx 0.15$ and at $c_1 \approx 0.4$, the two fixed points for which G_1 expresses opinion disappear. For costs that high, opinion group G_1 will not be publicly audible any more. Asymmetric costs can hence drive certain opinion groups into silence.

[16] Ideals of e.g. free speech might have such an effect: People then see it as their duty to voice their opinion, *especially* if it does not conform to the apparent majority.

[17] One may think here about online platforms whose design favours engagement of certain demographic groups or states that encourage certain groups to speak out or try to prevent others from voicing their opinion.

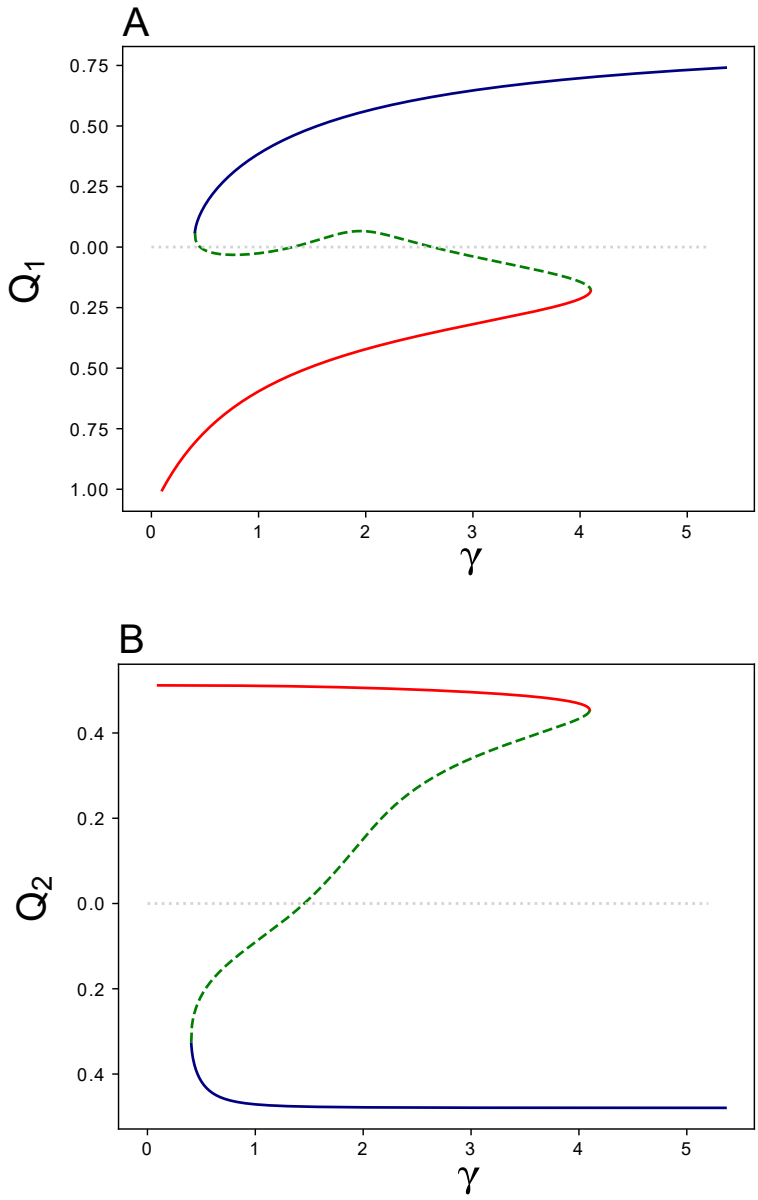


Figure 2.6: The development of Q_1- and Q_2-fixed points with γ given $\beta = 10$, moderate $\delta = 1.6$ and $c = 0.1$. For $\gamma < 0.4$, only group G_2 is expressive. A second fixed point arises for higher γ in which G_1 is predominating public discourse. There is no fixed point in which both groups are expressive.

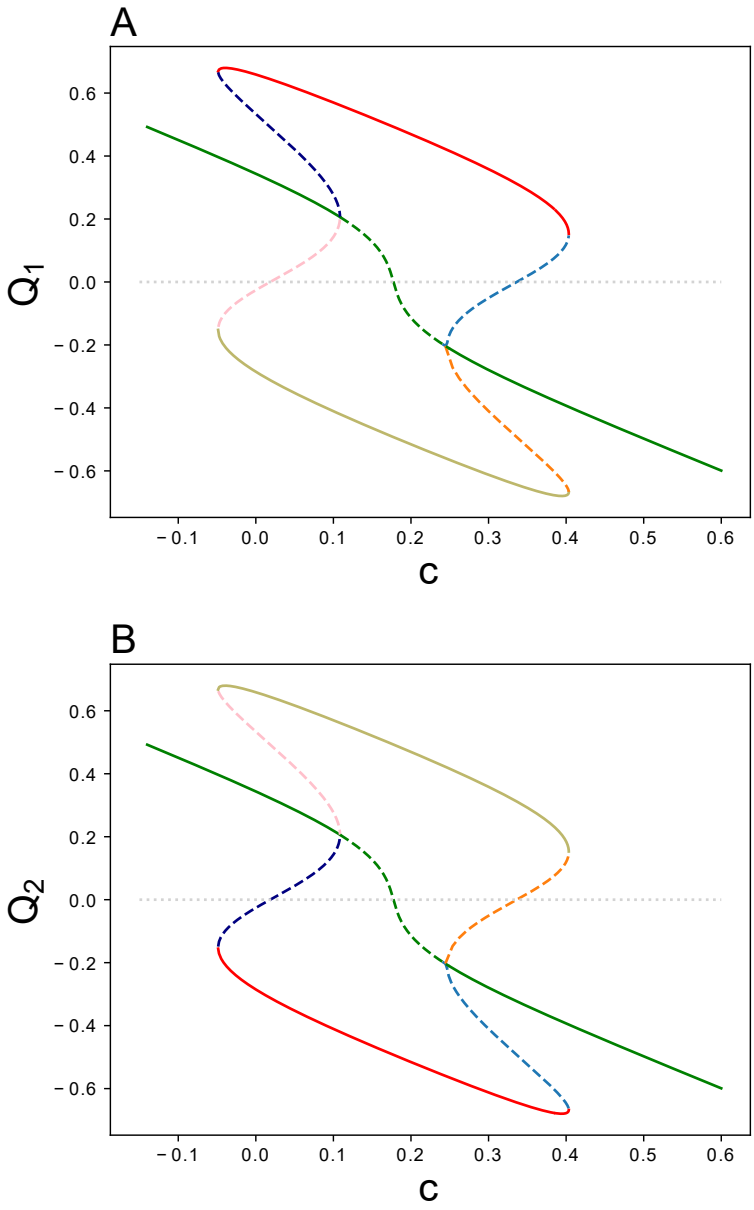


Figure 2.7: The development of the fixed points with c given $\beta = 10$ and $\gamma = \delta = 2.1$. (A) and (B) are symmetric since c is the same for both and has the same impact on both groups if they also have identical structural parameters. If expression is costly, everyone is silent, if it has negative costs, everyone speaks out.

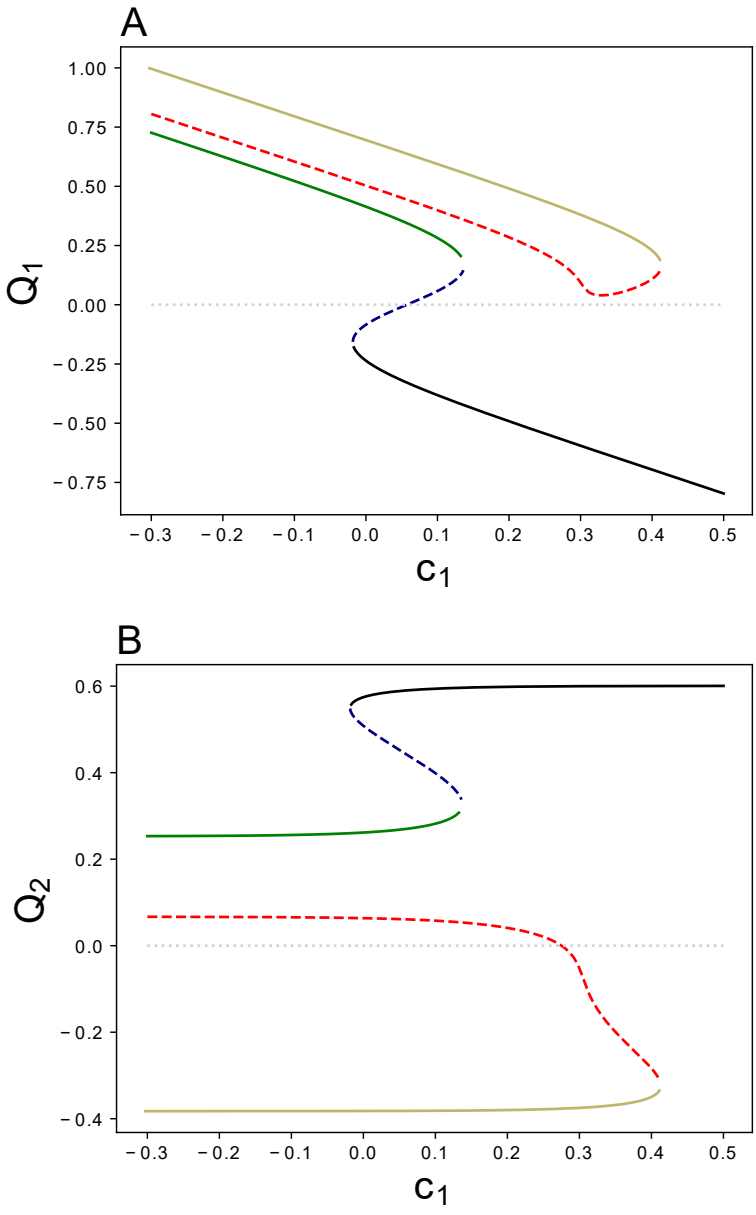


Figure 2.8: Fixed-point development with c_1 independent of c_2, given $\beta = 10$, $\gamma = \delta = 2.36$, $c_2 = 0.1$. Strongly negative c_1 corresponds to a strong motivational disposition (or the facilitation of opinion expression for the group) in the opinion group to express their opinion. There, only fixed points in which this opinion group is expressive exist. For decreasing motivation (or if opinion expression is impeded), fixed points arise in which the second opinion group is the only expressive one.

2.6 Discussion

The present model provides a structural view on collective opinion expression. It reproduces the counterintuitive result postulated by Noelle-Neumann in her theory of the spiral of silence [113, 111], namely the possibility of public predominance of a minority opinion. While the influence of mass media has been stressed in many publications concerning the spiral of silence, we show that no mass media is needed for this effect. Being an internally well-connected community alone can be enough to gain public opinion predominance. Mass media could nevertheless be included in the model as an agent being connected to a large subset of agents across opinion-group borders.[18] Our findings gain additional traction in light of the advent of social media, which facilitated communication among like-minded people and decentralized information distribution. In the model, facilitation of public opinion expression can be accounted for by reduced costs, potentially enabling certain opinion groups to speak out (see Fig.s 2.7 and 2.8). It must be noted here, however, that the essential fear of isolation, which Noelle-Neumann posits as the basis of her theory of public opinion (and the spiral of silence) [111], is not included in this model – and whether it plays a role in online environments is still debated [123]. It suffices for our purposes that disagreement is perceived as a negative experience (and agreement positively), not necessarily motivated by a deeper fear of being excluded from the social group which might appear to be a majority. Silence might follow simply because one does not see any value in discussing with (too many) others of a diametrically opposed opinion. Our research shows that the collective process Noelle-Neumann called the spiral of silence is only one possible outcome of the mircoassumptions on which the theory builds. The present approach also provides conditions for the 'overcoming' of situations in which one group is silent. There, the numerical proportions do not necessarily have to change. An increase in internal cohesion – or reduced expression costs – can be sufficient. On the other hand, we also find that from a game-theoretic perspective, if a minority is too small or costs are too high, even maximum internal cohesion cannot heave the minority opinion into public predominance (see Fig. 2.2).

In [137], the effect of the ego-network size, that is, the (average) number of connections of the agents, on the occurrence of the spiral of silence was investigated. It was concluded that an increase in network density makes it more probable that one opinion group does not speak out publicly. In this chapter, we show that more density might even have the opposite effect. It depends on *where* the additional connections are made: If new connections are guided by homophily, such that the opinion blocks become more cohesive, the spiral of silence might even be overcome (see path (i) in Fig. 2.9). We then arrive at a structure similar to 'echo chambers,' in which mainly the voices affirming one's own view are heard and the others are blocked out (see [19] for a contribution linking opinion dynamics to the emergence of echo chambers). If the additional connections are made between the opinion blocks, both γ and δ decrease, which might make it more probable that the individuals have a more realistic picture of the overall opinion landscape. Then, the spiral of silence is indeed more probable. But

[18] One could then attribute the mass media agent(s) a stronger authority, i.e. impact on the public opinion perception of individuals. For a model of a social system with authoritative leaders and dissenting lower-ranking individuals, see [86].

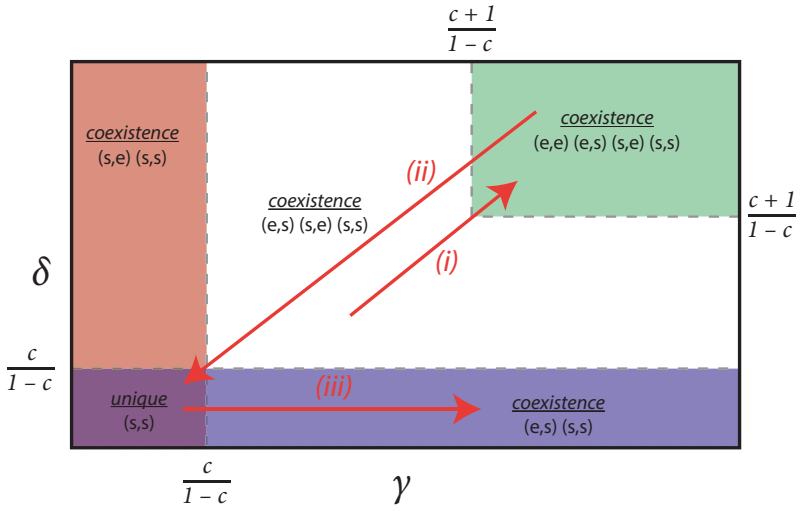


Figure 2.9: Illustration of the transitions between the game-theoretic equilibrium regions for (i) stronger internal cohesion of both opinion groups ('echo chambers'), (ii) less internal cohesion of both (heterophilious connections), and (iii) stronger internal cohesion for only one opinion group ('#metoo').

if the cross-group connections grow even further, both opinion groups misjudge their proportion to their own disadvantage, such that no group speaks out if there are costs associated to opinion expression (path (ii) in Fig. 2.9).

While we have stressed the generality of this model, we want to emphasize its limits as well: The homogeneous network structure of opinion blocks is not particularly realistic. Real social networks are rather heterogeneous, with well-connected and very active hubs and more 'remote' individuals. Nevertheless, stochastic blocks [158] can serve as a baseline for mathematical accessibility. At the moment, the model is restricted to two opinion groups with diametrically opposed opinions. An extension to more opinion groups appears to be a natural next step – one could, for example, include moderate agents that hold opinions that are not as strongly opposed to other groups as in the present model. These might also be structurally moderate in the sense that they sustain more heterogeneous connections. Such an approach could be insightful in the further investigation of, e.g. recent claims about a silencing of moderate voices in social media environments [10].

Moreover, this work is concerned with one way of reacting on social feedback, namely, the change in willingness to express one's opinion. Change in opinion is not included. It is probable that these phenomena take place on different time scales. Also, the social environments prompting opinion formation might be different from the ones in which opinion predominance is fought for. This point will become more visible in Ch. 4, where user interactions with respect to information sharing and discussions are strikingly different. Hence, a combination of models of opinion change and opinion

expression might be in order in a multi-layer network approach, in which opinion formation and the competition for public opinion predominance take place on possibly different but interdependent network structures.

We are also seeking a more systematic larger-scale view on collective phenomena of opinion expression, which are closely related to the parameters γ and δ in the model. A very prominent example of emerging collective opinion expression online, for which this model can provide an explanation, is the Twitter-hashtag '#metoo' and the subsequent movement against sexual harassment and sexual assault. Women found a device (in this case, a hashtag) that allowed them to find and connect to people who had experienced the same, and also to people who supported them – and all of a sudden, many of them decided to speak out (path (iii) in Fig. 2.9). Measurements are an intricate task here: The networks one constructs out of interactions between individuals are only the networks *of interaction*, that is, of only one part of the actions one wants to observe. Silent individuals do usually not show up in such networks since they are not involved in an observable way. While this problem must be circumvented in order to be able to test whether the model is plausible, the model itself is also a warning for researchers dealing with online data of public debate: What becomes visible online might not only not reflect the opinion proportions in society at large, but not even of interested users on the platform itself. We develop a method to investigate ideological differences in engagement in public debate on Twitter in Ch. 4 to approach this from an empirical perspective.

2.7 Summary

We developed a model of opinion dynamics where opinion exchange between individuals has an effect on *the willingness of opinion expression*, and not, as in most previous models of opinion dynamics, the individual's opinion itself. This approach has remained underdeveloped in models of opinion dynamics up to now, while being of relevance especially in online contexts where connecting to like-minded others is facilitated and cost of communication is lowered. The model allows insights into dynamics of collective opinion expression, and investigates how different group structures relate to different regimes of public opinion visibility. We investigated how social structure and costs for opinion expression can prevent or promote collective expression of different opinion groups. We took an incentive-based view on opinion exchange and grounded the model in a plausible micro-mechanism, where agreement is positively and disagreement is negatively received. For mathematical tractability, we used a stochastic blockmodel for two opinion groups as the baseline network structure. We approached the model both from a game-theoretic and from a dynamical systems perspective. The game-theoretic perspective provided an overview over co-existence regimes of different Nash equilibria depending on structural parameters γ and δ, and the cost of expression c. It also provided limits for the number of group members needed in a group in order to reach public audibility (all other parameters constant). To address questions of equilibrium selection and stability, we used the reinforcement learning algorithm of Q-learning. We combined it with a mean-field approach in which the N-agent system was reduced to two dimensions, representing the two opinion groups. In

bifurcation analyses of the structural parameters, we showed how even a strong minority can dominate public debate if it increases its internal connections. Moreover, we investigated how a decrease in expression costs for one group can heave it into public audibility.

Chapter 3

Grounding force-directed network layouts with latent space models

3.1 Using force-directed layouts (not only) to determine opinion groups

This chapter aims to bring together two strands of research: Latent space approaches to network analysis and force-directed layout algorithms (FDLs). FDLs are used ubiquitously for network exploration, illustration, and analysis in a wide variety of disciplines [4, 96, 36, 140, 153, 41, 124, 154, 37]. Nevertheless, it is still unclear how to precisely interpret the outcomes of graph drawings with FDLs, nor what constitutes an appropriate algorithm choice for different network data from the range of FDLs available – points also highlighted recently in [72, 154]. We argue and show that explicit interpretability can be provided by latent space approaches, which have the goal of embedding a network in an underlying social space, and where distances between nodes influence the probability of a tie between them. We derive and implement a new type of FDL, which we call *Leipzig Layout*.

In the context of this thesis, the inferred underlying social space can be used to distinguish between different opinion groups of users – which is a precondition for comparing their willingness of collective opinion expression. Several methods have been developed for this task, such as different community detection algorithms which produce partitions of nodes from network data. The most-used techniques include modularity maximization [108] or fitting a generative model on the data, such as the stochastic blockmodel [68].[1] But, as will become apparent in the following section (and in Ch. 4), FDLs have the potential of revealing more than strict partitions of

[1] Groups of agents holding similar opinions can of course be determined on the basis of other data than networks, e.g. via automated [40] or manual [146] classification of user-generated content. But as the focus of this thesis is structural, we will focus on network-based approaches.

nodes, since e.g. also intermediate zones of "different relational density" [154, p. 9] are made visible, and also distances *between* groups can be determined. The usage of visualization produced by the FDL is, of course, not restricted to distinguishing certain node groups (even though FDLs are very often used for exactly this purpose, as will become apparent in the next section). It can be useful for a variety of tasks: Other applications include identification individual nodes, exploration of arrangement and proximity of survey items (see Sec. 3.5), or even narrative readings of network layouts [41, 154].

In order to arrive at a new type of FDL, we first briefly sketch how FDLs became the predominant tool for graph drawing, their connection to modularity [110], and their shortcomings with respect to interpretability. We also introduce latent space approaches to network analysis, and subsequently show how force terms of a new type of FDL can be derived from said latent space models, where the forces move nodes towards positions and parameters which maximize the likelihood for the network under the given model. We derive force equations for three types of networks: Unweighted networks, cumulative networks (such as the much-studied Twitter retweet networks), and weighted networks. We spatialise a number of real-world networks with the FDL. We also show how existing algorithms, specifically ForceAtlas2 [73], Fruchterman Reingold [51], and Yifan Hu [69], differ from the presented FDLs.

3.2 Latent spaces and FDLs: An intersection

Initially, network visualization algorithms[2] had been conceived to facilitate graph reading – they were supposed to make small networks readable in the sense that paths and nodes in the network were clearly accessible, that the edges had similar lengths and that the network was drawn as symmetric as possible [43]. These readability criteria were also referred to as 'aesthetic' (see e.g. [69, 26, 18, 145]). Progress in network science and the sudden availability of very large network data sets at the end of the millennium – for which a comprehension of individual node positions and paths was illusory – shifted focus: Now, networks needed to be drawn so that community structure and topological features were mediated in the layout. FDLs (partly having been developed already before this complex turn, notably in [51, 43]) turned out to be useful and efficient tools for this task. The algorithms have in common that all nodes repel each other (the repulsive force is usually proportional to a power of of the distance between nodes, i.e. $F_r \propto d^r$), while connected nodes are additionally drawn together by their edges ($F_a \propto d^a$, $a > r$).

Noack, in a seminal work [110], connected FDLs to modularity, one of the most central measures of clustering in networks in use today. Roughly speaking, modularity Q compares the proportion of links connecting nodes within a group of nodes with the proportion expected if the edges in the network were randomly rewired [108]. Community detection algorithms, such as the Louvain algorithm [21], aim to find partitions of a network that maximize this value. Noack showed that, under certain constraints,

[2] This paragraph largely follows [72] in its account of the history of graph drawing and (force-directed) network layout algorithms.

modularity can be transformed into an expression that equals the energy function of force-directed layouts. Constraints for the equivalence are

(i) that nodes can only be placed either at the same position (which then represent nodes in the same cluster) or at distance 1 from each other (if nodes are not in the same cluster).

(ii) that FDLs operate in a space of (at least) $k-1$ dimensions, where k is the number of modularity clusters (usually, FDLs embed networks in a two-dimensional space).

(iii) that the exponents of attractive and repulsive force should be non-negative. (Obviously, if the repulsive force has a negative exponent, placement of nodes at the same position would be impossible.)

For FDLs, this means that if they fulfill (ii) and (iii), energy-minimal states of force-directed layouts are *relaxations* of modularity maximization: They make community structure in networks visible without constraint (i) of having to sort nodes into different, fixed partitions with distance 0 or 1 from each other. They can assign continuous positions in space. Or, phrased the other way round: Modularity is then a special case of the energy function of FDLs. However, Noack's finding is diluted by the fact that network visualizations with FDLs are commonly restricted to two (or at most three) dimensions; and moreover, most FDLs in use today employ a negative exponent for the repulsive force. To balance this, Noack gave qualitative observations of which algorithms, even if they do not *exactly* fulfill the constraints above, tend to produce results that resemble modularity clusterings: Exponents in the forces should be characterized by $a \geq 0$, $r \leq 0$, $a - r \approx 1$, and $a \approx 0$.[3]

The connection to modularity – which, notably, had not been intended in the design of the algorithms – helped give additional credibility to FDLs. All in all, this lead to the widespread adoption of FDLs for network visualization: They were not only used for illustrative purposes [4, 37, 36], but also to explore and analyse network data [124, 153, 41, 96, 154]. It is, however, unclear what information FDLs add to modularity clusterings by placing the nodes in a continuous space. It has been stressed that while they have been widely used, a thorough assessment of what exactly is entailed by the produced layouts has not been provided yet: "practices lead [...] the way, as the discussion in the academic literature has lagged behind" [72, p. 29].[4] And on the question of what it means that two nodes are placed close to each other by a certain FDL, answers have remained somewhat tentative: "While in spatialized networks closer nodes *tend to be* more directly or indirectly associated, no strict correlation should be assumed between the geometric distance and the mathematical distance" [154, p. 4].[5]

[3] ForceAtlas2 ($a = 1$, $r = -1$) is in that sense more closely related to modularity than FruchtermanReingold ($a = 2$, $r = -1$) or Yifan Hu (which uses similar forces to FruchtermanReingold, but with a multilevel algorithm).

[4] [72] departs into a somewhat different direction than the work presented here by proposing certain interventions which help interpret what is visible in FDLs that are already in use today; but both approaches try to tackle the shortcomings of FDLs that are sketched above.

[5] Technically, a geometric distance is of course also a mathematical distance. The authors apparently refer to a kind of graph-theoretic distance, e.g. the shortest path between two nodes, with the expression 'mathematical distance'.

Moreover, many different types of FDLs have been developed, several of which at least approximately subsume modularity clustering.[6] But which one of them constitutes an appropriate choice for a certain data set at hand?[7]

FDLs are often implemented in easily accessible tools such as Gephi [17], while not all researchers using the tool might possess the methodological training to assess the mechanics behind them. But the problems sketched above give an additional explanation for the fact that the limits or benefits of a chosen FDL are usually not discussed and "tools such as Gephi [are often treated as] as black boxes" [28]. It hence appears that FDLs lack *grounding*, for example through an underlying model generating the forces, with which the distances between nodes could be given a rigorous interpretation, and forces could be chosen that are suitable to the network data one wants to analyse. We propose to base a new type of FDL on latent space models of network analysis, with which network layouts can be interpreted explicitly.

Latent space approaches to (social) network analysis have been developed to infer social or political positions of actors in an underlying latent space from their interactions. They represent a class of models based on the assumption that the probability that two actors establish a relation depends on their positions in an unobserved social space [66]. The social space can be constituted by a continuous space, such as an Euclidean space, or a discrete latent space, where each node is in one of several latent classes [93].[8] In the physics literature, models of this type were introduced under the name of spatially embedded random networks [16]. There, the Waxman model [161] and random geometric graphs [121, 39] were recognized as specific examples.[9] Recently, latent space models have been employed successfully in the estimation of continuous (one-dimensional) ideological positions from social media data [15, 14, 71], specifically from Twitter follower networks. The works covered large quantities of users and showed good agreement with e.g. party registration records in the United States [15]. The estimation of positions in the latent space was achieved with correspondence analysis in [15], while in [14], a Bayesian method was used where the posterior density of the parameters was explored via Markov Chain Monte Carlo methods.

This is where the present work intersects: We attempt to take an alternative route in order to arrive at a specific form of force equations for FDLs. We obtain the forces on the basis of latent space models of node interactions. The positions of the nodes in an assumed latent space influence the probability of ties between them – the closer their positions, the more probable it is that they form a tie. We derive an FDL as a maximum likelihood estimator of such a model. This approach clarifies the underlying assumptions of our layout algorithm and makes the resulting layout *interpretable*. We derive three different forces for three different types of networks, specifically adapted to the task of embedding them in a political space: unweighted, cumulative, and weighted

[6] We note here that modularity clustering is not without significant weaknesses, such as its resolution limit [47] or strong degeneracies of high-scoring solutions [59].

[7] Certain quality measures to *compare* network layouts have been proposed, such as the normalized atedge length [109] corresponding to the total geometric length of the edges of a network divided by the graph density and the total geometric distance between nodes. But these do not give meaning to the produced layout beyond network-immanent topological features.

[8] The latter case includes the well-studied stochastic block model [68].

[9] No links to the latent space approaches of [66, 63] were made.

networks. Moreover, alternative interaction models can in principle be used to develop force-directed layouts in a completely analogous way. For this, the present work can serve as a blueprint.

If one wants to take network layouts seriously, an approach highlighting the underlying assumptions of a layout and guiding its interpretation is necessary. While some might claim that the visualization of a network only serves illustrative purposes, their wide-spread use, not only for exploration and illustration, but also visual analysis of networks [37, 153, 41, 96, 154] underscores the necessity of this enterprise: Exploration and interpretation are, in practice, guided by force-directed layouts for many researchers from a variety of disciplines.

3.3 From latent space models to force equations

In this section, we will show how force terms in a force-directed layout algorithm can be derived from latent space models of node interactions. Central to this procedure is the assumption that nodes tend to form ties to others that are close to them in a latent social space. The closer two nodes, the higher the probability that one forms a tie to the other. Since none of the positions (as well as none of the additional parameters of the statistical model which will be introduced in the corresponding subsections) are directly observed, the statistical problem posed here is their inference. Given the underlying model, one can determine the likelihood function $L(G)$ for any observed network. The positions and parameters are then inferred via maximum likelihood estimation. In our approach, this is done by treating the negative log-likelihood as a potential energy. The minima of this potential energy are the local maximizers of the likelihood. Its derivatives with respect to the positions and parameters of the nodes can be considered as forces that move the nodes towards positions that maximize the likelihood.

We will cover three different types of directed[10] networks: Unweighted networks, such as the follower networks covered by Barberá and colleagues [15, 14], cumulative networks (which include Twitter retweet networks), and weighted networks. We will present the derivation of the forces for the unweighted case in detail. The complete derivations for the other two cases are given in Appendices C and D.

3.3.1 Unweighted networks

Consider an unweighted graph $G = (V, E)$ with nodes $i \in V$ and edges $(i, j) \in E$. The graph can be described by an adjacency matrix $A = \{a_{ij} | a_{ij} = 1 \text{ if } (i, j) \in E\}$. Now let us assume that the nodes are represented by vectors $\mathbf{x}_i \in \mathbb{R}^n$ and $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ denotes the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. We formulate the probability of a tie as

$$p(a_{ij} = 1) = \text{logit}^{-1}(\alpha_i + \beta_j - d_{ij}^2) = \frac{\exp(\alpha_i + \beta_j - d_{ij}^2)}{1 + \exp(\alpha_i + \beta_j - d_{ij}^2)}. \tag{3.1}$$

[10] Undirected networks are implicitly included as a special case where $a_{ij} = a_{ji}$ and each node only has one additional parameter α.

We choose the probability to be dependent on the squared Euclidean distance between the two node positions.[11] α and β are additional parameters that also influence the probability of a tie. α_i can be interpreted as an activity parameter related to the out degree of node i: The higher α_i, the higher the probability of a tie from i to others. β_j influences the probability of ties *to* j (see [15, 14, 66]) and is related to the in degree of node j. The parameters allow nodes that occupy the same position in space to have different degrees – as an example, there might be people with roughly the same political position as, say, a state leader, but it is generally unreasonable to expect that these users have the same amount of followers on social media.[12]

For a given graph G, the likelihood function $L(G)$ can be written as the product of the probability of an edge *if* there exists an edge between two nodes, and the probability of there not being an edge if not:

$$\begin{aligned} L(G) &= \prod_{(i,j)\in E} p(a_{ij}=1) \prod_{(i,j)\notin E} (1-p(a_{ij}=1)) \\ &= \frac{\prod_{(i,j)\in E} \exp\left(\alpha_i+\beta_j-d_{ij}^2\right)}{\prod_{\substack{i,j \\ i\neq j}} \left(1+\exp\left(\alpha_i+\beta_j-d_{ij}^2\right)\right)}. \end{aligned} \tag{3.2}$$

The logarithm is given by

$$\begin{aligned} LL(G) &:= \log L(G) \\ &= \sum_{(i,j)\in E} (\alpha_i+\beta_j-d_{ij}^2) - \sum_{\substack{i,j \\ i\neq j}} \log(1+\exp(\alpha_i+\beta_j-d_{ij}^2)). \end{aligned} \tag{3.3}$$

If we consider the *negative* log-likelihood as a potential energy, the minima of this potential are the local maximizers of the likelihood. *Its (negative, once again) derivatives with respect to the positions* $\mathbf{x}_i$ *of the nodes can be considered as forces that move the nodes towards positions that maximize the likelihood.* For a concrete node i', an attractive force is generated by node j' if i' establishes a tie to j':

$$F_{\text{att},i'}^{j'} = \frac{\partial}{\partial \mathbf{x}_{i'}} (\alpha_{i'}+\beta_{j'}-d_{i'j'}^2) = -2(\mathbf{x}_{i'}-\mathbf{x}_{j'}). \tag{3.4}$$

If j' also establishes a tie to i', the same attractive force is applied again. On the other hand, a rejecting force is always present for each possible tie:[13]

$$F_{\text{rej},i'}^{j'} = -\frac{\partial}{\partial \mathbf{x}_{i'}} \log(1+\exp(\alpha_{i'}+\beta_{j'}-d_{i'j'}^2)) = \frac{1}{1+\exp(-\alpha_{i'}-\beta_{j'}+d_{i'j'}^2)} 2(\mathbf{x}_{i'}-\mathbf{x}_{j'}). \tag{3.5}$$

[11] That the probability is dependent on the squared distance is also assumed in [14], while in [15, 66], the linear distance is used.

[12] On the other hand, some users might simply be more active than others, hence forming more ties, while sharing a political position.

[13] Note the sign reversal in the exponent of the exponential function in the denominator in the last equivalence, which stems from $\exp(\alpha_{i'}+\beta_{j'}-d_{i'j'}^2)/(1+\exp(\alpha_{i'}+\beta_{j'}-d_{i'j'}^2)) = 1/(1+\exp(-\alpha_{i'}-\beta_{j'}+d_{i'j'}^2))$.

Another repulsive force on i' appears for this node pair for the potential tie from j' to i'.

The derivative of Eq. (3.3) with respect to $\alpha_{i'}$ and $\beta_{i'}$ gives us the forces on the parameters of node i', such that

$$F^{j'}_{\alpha_{i'}} = a_{i'j'} - \frac{1}{1+\exp(-\alpha_{i'} - \beta_{j'} + d^2_{i'j'})} = a_{i'j'} - p(a_{i'j'} = 1) \tag{3.6}$$

and

$$F^{j'}_{\beta_{i'}} = a_{j'i'} - \frac{1}{1+\exp(-\alpha_{j'} - \beta_{i'} + d^2_{i'j'})} = a_{j'i'} - p(a_{j'i'} = 1). \tag{3.7}$$

The sum over all forces on the single parameters yields the difference between the actual and the expected in/out degree. In the equilibrium state, where this sum yields 0, the real in/out degree of nodes equals the one expected under the model of Eq. (3.1):

$$\sum_{j'} F^{j'}_{\alpha_{i'}} = d^{\text{out}}_i - \langle d^{\text{out}}_i \rangle, \tag{3.8}$$

$$\sum_{j'} F^{j'}_{\beta_{i'}} = d^{\text{in}}_i - \langle d^{\text{in}}_i \rangle. \tag{3.9}$$

3.3.2 Cumulative networks

Force equations can also be derived for networks which are constituted by a number of binary signals between nodes – for example, when users of an online platform create several posts, each of which can be taken up by others (e.g. through liking or sharing the content). A much-studied case are Twitter retweet networks, which are frequently employed to investigate opinion factions on the platform [36, 37] (see also Ch. 4).

We consider, for each type of action k initiated by a node j (e.g. a tweet), an unweighted graph $G^k_j = (V, E^k_j)$ with nodes $i \in V$ and edges $(i,j) \in E^k_j$, where an edge means that i has formed a tie with j upon action k. The graph for each k can be described by an adjacency matrix $A^k_j = \{a_{ij} | a_{ij} = 1 \text{ if } (i,j) \in E^k_j\}$.[14]

Analogously to the unweighted case, we assume the probability of establishing *a single tie upon action k from user i to user j* with

$$p(a^k_{ij} = 1) = \frac{1}{1+\exp(-\alpha_i - \beta_{jk} + d^2_{ij})}, \tag{3.10}$$

where where each action k of j has its own parameter β_{jk} which affects the in degree of j.

The log-likelihood for the cumulative network can be written as

$$LL(G) = \sum_j \sum_{k=1}^{m_j} \Big(\sum_{(i,j)\in E_{jk}} (\alpha_i + \beta_{jk} - d^2_{ij}) - \sum_{\substack{i \\ i\neq j}} \log(1 + e^{-\alpha_i - \beta_{jk} + d^2_{ij}}) \Big) \tag{3.11}$$

[14] Which constitutes an m-star graph with $m = |E^k_j|$.

Here, in addition to user pairs, we sum over all actions k. The derivation of forces for this case is largely analogous to the unweighted case and can be found in Appendix C, along with the concrete force equations.

3.3.3 Weighted networks

So far, we have assumed a binary signal between node pairs – e.g. whether an individual follows another or not, or whether someone shares certain content of another individual or not. A natural extension to non-binary cases is the ordered logit or proportional odds model. There, a response variable has levels $0, 1, ..., n$ (e.g.: people rate their relationships to others on a scale from 0 to 6, or similar). To this end, we consider the general case of weighted networks with finite weights that can be transformed into natural numbers (with 0), i.e. an adjacency matrix $A = \{a_{ij} = k | k \in \mathbb{N}_0\}$. The probability of the variable being greater than or equal to a certain level k is given by [64, 81]:

$$p(a_{ij} \geq k) = \frac{1}{1 + \exp(-c_k - \alpha_i - \beta_j + d_{ij}^2)}, \tag{3.12}$$

where $k = 0, 1, ..., n$. ($c_0 = \infty$, $c_{n+1} = -\infty$.) The probability of a_{ij} equal to a certain k is given by

$$p(a_{ij} = k) = P(a_{ij} \geq k) - P(a_{ij} \geq k + 1). \tag{3.13}$$

The likelihood $L(G)$ is given by

$$L(G) = \prod_{i,j} \left(\frac{1}{1 + \exp(-c_{a_{ij}} - \alpha_i - \beta_j + d_{ij}^2)} - \frac{1}{1 + \exp(-c_{a_{ij}+1} - \alpha_i - \beta_j + d_{ij}^2)} \right), \tag{3.14}$$

and the log-likelihood by

$$LL(G) = \sum_{\substack{i,j \\ i \neq j}} \log \left(\frac{1}{1 + \exp(-c_{a_{ij}} - \alpha_i - \beta_j + d_{ij}^2)} - \frac{1}{1 + \exp(-c_{a_{ij}+1} - \alpha_i - \beta_j + d_{ij}^2)} \right). \tag{3.15}$$

The force equations derived from the log-likelihood (for the case of a three-point scale) can be found in Appendix D. Potential applications of a visualization with these forces are manifold. In smaller data sets, a non-binary signal between nodes, e.g. rating of the relationships between individuals of a social group, might be given between all node pairs. But often, there might be cases for which a subset of individuals (say, politicians, public figures, etc.) or items (e.g. the importance of political goals, technologies, etc.) are rated by others. Then, only the rating individuals receive an α, while only the rated have a β-parameter. The interpretation of the parameters α and β might need adjustment: They now rather refers to the tendency of an individuals to give/receive rather high/low ratings.

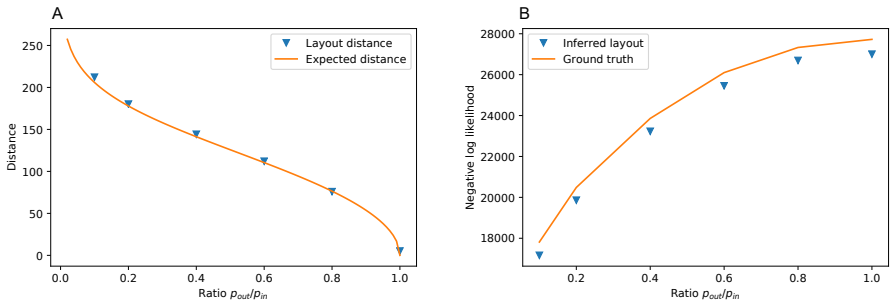


Figure 3.1: Expected distance of an SBM (two blocks, 100 nodes each) with varying p_{out} compared to the distance between the center of mass of the clusters in the proposed layout algorithm, averaged over five runs (A). Not only is the inferred distance by the force-directed layout algorithm nearly identical to the expected one, but the log-likelihood of the inferred latent space surpasses the ground truth in all cases (B).

3.4 Implementation and validation

The force-directed layout algorithm is implemented in JavaScript, building upon the d3-force library [24]. There, force equations are simulated using a velocity Verlet integrator [155, 143]. A ready-to-use implementation, which we call *Leipzig Layout*, is available: `https://github.com/pournaki/leipzig-layout`.[15] It builds upon the force-graph library [9] to interactively display the graph and the evolution of node positions in the simulation of forces. Note that, at the current state, this layout tool works reasonably fast for networks below 10,000 links.[16]

Validation for the unweighted case was performed by testing the agreement with the expected distance of a stochastic block model (SBM) of two blocks with varying p_{out} and $p_{\text{in}} = 0.5$. In the above model, the expected distance can be computed by placing the nodes of each block on the same point in space, and then choosing the distance d between the two blocks so that $p_{\text{out}} = \frac{1}{1+\exp(-d^2)}$ (α and β are set to 0). With this underlying latent space, one can then draw a network according to the given probabilities and let the layout algorithm infer the latent space again. Averaged over 5 runs and for 100 nodes per block, we observe that the inferred distance of the centers of mass of the blocks are nearly identical to the expected one (Fig. 3.1 A), and the log-likelihood of the inferreds latent space surpasses the one of the actually drawn one in all cases (Fig. 3.1 B).

Moreover, we compared the inferred negative log-likelihood with the one for the actually drawn network from a Gaussian distribution of two groups of nodes with a σ of 1/12 and a distance of 5/6 between the groups with varying node number, averaged

[15] For the moment, this implementation is restricted to unweighted graphs. An extension for weighted and cumulative graphs will be published on the same repository.

[16] For larger networks, approximations such as Barnes-Hut are usually employed to ensure fast convergence. In the case of the presented forces and parameters, Barnes-Hut is not directly applicable. We therefore leave this as a future task for the development of our layout algorithm.

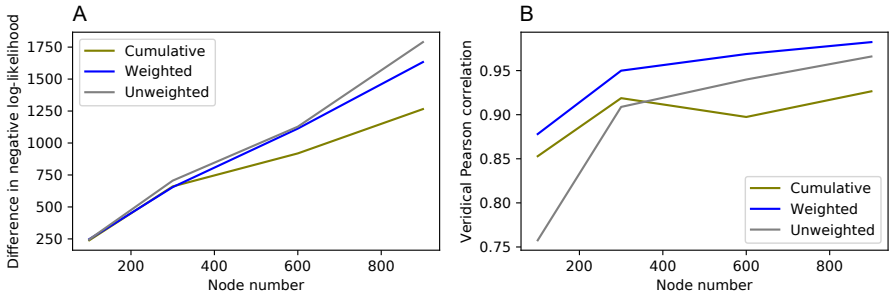


Figure 3.2: Difference between negative log-likelihood of ground truth and inferred negative log-likelihood (A), for a Gaussian distribution of two groups of nodes with a sigma of 1/12 and a distance of 5/6 between the groups with varying node number (averaged over three runs). In all cases, the log-likelihood of the inferred latent space surpasses the ground truth (i.e. difference in negative log-likelihood is positive). Still, similarity between ground truth and inferred distances between nodes is high (increasing with node number), as is visible in the average veridical Pearson correlation between distance matrices (B).

over three runs. In all cases, the log-likelihood of the inferred latent space is higher (i.e. the negative log-likelihood lower) than the ground truth. Still, similarity to the ground truth distances between nodes was high throughout, which we assessed with a Mantel test [92]. Veridical correlation between distance matrices can be inspected in Fig. 3.2 B, the average z-score is reported in Appendix F.

3.5 Real-world networks

Next, we use Leipzig Layout to spatialise several real-world networks: Undirected Facebook friendship networks, the directed Twitter follower network of the German parliament, the retweet network of Twitter debate surrounding the publication of a letter on free speech by Harper's magazine, and a survey on different types of energy-generating technologies.

Facebook100: Haverford & Caltech The Facebook100 data set consists of online social networks collected from the Facebook social media platform when the platform was only open to 100 universities in the US [148]. The data set is of particular interest since it contains social networks with quite rich metadata within a well-defined social environment. We analyse friendship networks – undirected networks where a tie between users represents that both have agreed to connect with each other as 'friends' on the platform.

We spatialise the friendship network of Haverford University in Fig. 3.3 (links have been omitted for better accessibility). On the left, it is visible that students are spatially layered according to their year by the layout algorithm. The first-year students

(light green nodes on top) are visually separated from the others. The layout becomes denser for students who have been at the university for a longer time. The layers are ordered chronologically – lilac nodes are second-year students, blue nodes third-year students, and so on. It seems that if students form cross-year ties, they tend to connect to others from adjacent cohorts. The local assortativity distribution[17] with respect to residence of the students of Haverford has been analysed in detail in [120]. There, it was found that first-year students tend to form ties to other students from their dorms, while students from higher years show less of a tendency to mix only with others they share residence with. This behavioral pattern can also be discerned in the spatialisation: For the first-year students, the students sharing a dorm tend to be placed rather close to each other, while for students from higher years, this is not the case. As a complement, we spatialise the network for the university with the highest overall assortativity with respect to dormitory in the data set: Caltech. There, the students' year does not influence Facebook friendship to a large extent; rather, students' friendships are more strongly guided by their residence [127]. This is reproduced by Leipzig Layout: Fig. 3.4 shows that students that share residence are visibly placed close to each other. On the other hand, students are less strongly grouped according to their university year.

German parliament: Twitter follower network While [14, 15] aim for the estimation of one-dimensional ideological positions of politicians (and their followers), the FDL proposed here embeds nodes in a two-dimensional space. We spatialise the Twitter follower network of all members of the German parliament that have an active Twitter account in Fig. 3.5. The parties (members colored according to their typical party color) are quite visibly separated. They are located along a circle that quite accurately mirrors the political constellation in federal German politics. The center-left to center-right parties (*SPD*, *Bündnis 90/Die Grünen* (Green party), *CDU/CSU*) are positioned between *Die Linke* (Left party) and the market-liberal *FDP*. The *AfD* (blue), a right-wing populist party with which collaboration has been ruled out by all other parties, accordingly occupies a secluded area. Interestingly, within parties, a one-dimensional arrangement is visible (except for the *Greens*). We can show that this mirrors the amount of cross-party ties, as well as the users' activity on Twitter: The further out on an axis between the innermost and outermost party member users are placed (outliers excluded), the smaller their share of ties to other parties (Fig. 3.5 C). The central and densely packed placement of the Greens can be explained by the fact that they tend to use Twitter quite homogeneously (see Appendix G) in the sense that there are no users which are inactive or lack followers: Each of them has an in/out degree of at least 50. Moreover, the party members are followed and follow all other parties (except for the *AfD*) in a well-balanced fashion.

This layout also illustrates the difference between ForceAtlas2 (see (B) in Fig. 3.5) and the layout algorithm at hand here: ForceAtlas2 incorporates a rejecting force between node pairs proportional to d_{ij}^{-1}, which leads to a stronger separation of nodes within clusters. Hence, while the overall arrangement of parties is similar to the Leipzig Layout, parties themselves are more strongly spaced out. A comparison to spatialisa-

[17] For a detailed account of local assortativity, see Ch. 4.

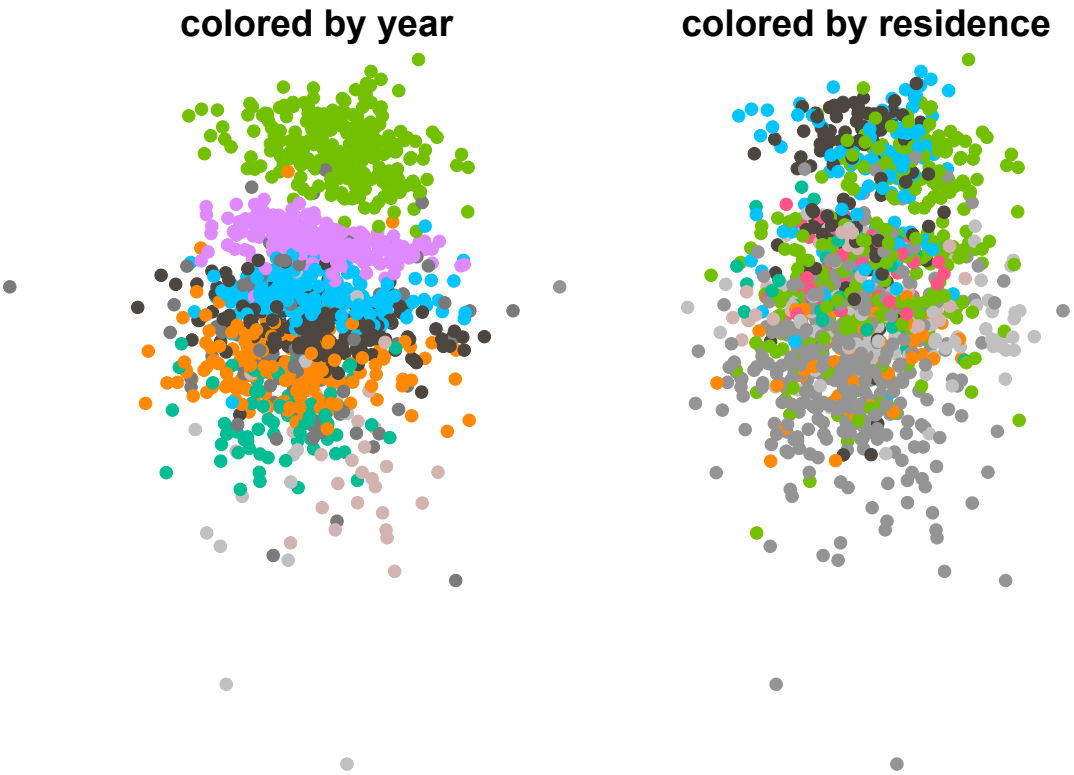


Figure 3.3: Friendship network of students of Haverford University, colored by year (left) and dorm (right). Students are spatially layered by year (chronologically ordered from top to bottom, with first-year students colored green, second-year students colored lilac, etc.; dark grey nodes correspond to students whose year is unknown). First-year students are visually separated from the others, while the layout becomes denser if students have been at university for a longer time. On the right, it is also visible that first-year students show a higher tendency to mix with others they share residency with (dark grey: dorm unknown), which was also found in [120].

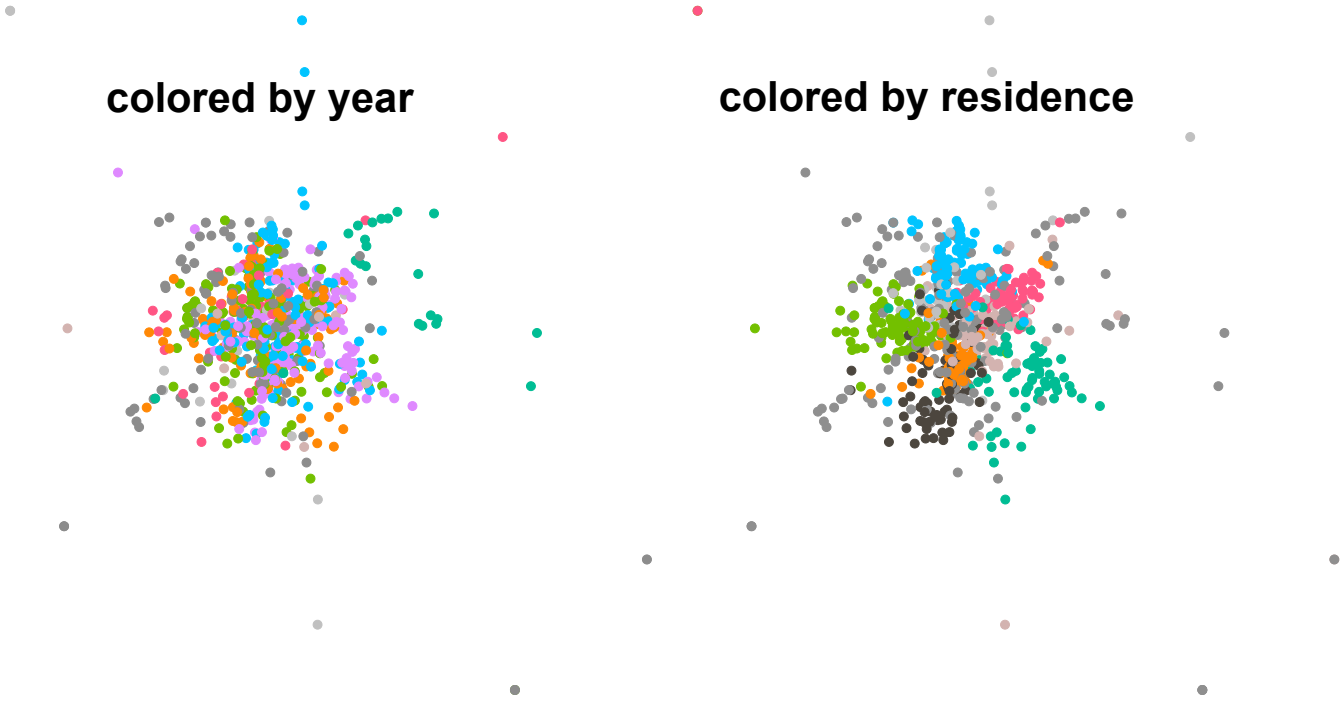


Figure 3.4: Facebook friendship network of Caltech, the network out of the Facebook100 data set with the highest assortativity with respect to residence, spatialised with Leipzig Layout and colored by year (left) and to dorm (right). Nodes are visibly placed according to dorm membership (dark grey: year/dorm unknown).

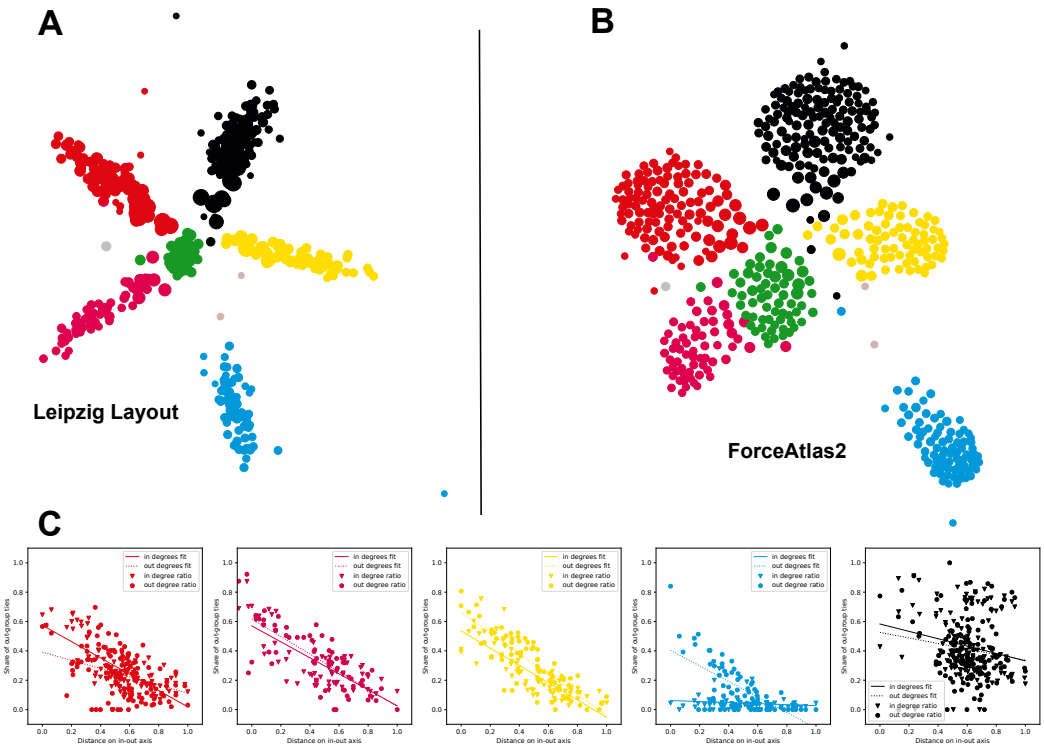


Figure 3.5: Leipzig Layout of the follower network of all German deputies that have a Twitter account (A). Members are colored according to their party. Clear division between parties, as well as a stronger division between the right-wing party *AfD* and the other parties is visible. All parties except the *Greens* are arranged on a one-dimensional axis. This is explained by a difference in cross-party ties between politicians of the same party: The further out a member on the party-internal axis, the fewer cross-party ties to and from them have been established (except for the *AfD*, which does not receive many ties from other parties no matter where the users are placed) (C, colored according to parties). ForceAtlas2, in comparison, has a stronger separation of nodes within party clusters due to its rejecting force being proportional to d^{-1} (B).

tions with the algorithms Yifan Hu and FruchtermanReingold can also be found in Appendix G. Moreover, we observe that several minima are inferred by both Leipzig Layout and the other FDLs – depending on the initial positions of the nodes. While the existence of different local minima is a general problem of FDLs, one can simply select the outcome with the highest likelihood with the present approach – a further advantage of an FDL grounded in an underlying model. A different, but less likely local minimum inferred with Leipzig Layout is also presented in Appendix G.[18]

Retweet network: Harper's letter In July 2020, Harper's magazine published an open letter signed by 153 public figures defending free speech which they saw endangered by 'forces of illiberalism.' Not only Donald Trump was denounced as contributing to illiberalism, but also some groups who advancing "racial and political justice," who had "intensified a new set of moral attitudes and political commitments that tend to weaken our norms of open debate and toleration of differences in favor of ideological conformity" [1]. On Twitter, the letter was controversially discussed subsequently.[19] The layout of the retweet network reproduces a division between critics and supporters of the letter: On the left side of Fig. 3.6, the account of Harper's magazine as well as prominent signees such as Thomas Chatterton Williams and Joanne K. Rowling are visible, while the right pole includes critics of the letter and its signees, such as Judd Legum, Astead W. Herndon and Julia Serano. Serano, a transgender activist, criticized that what the signees referred to as 'free speech' has prevented marginalized groups from speaking out. She explicitly mentioned 'transphobic "free speech"' which, she claimed, has had this effect in the past, and accused Rowling of having spread disinformation about trans children. That she was voicing rather specific criticism which aimed towards certain signatories of the letter is mirrored in her position close to the margin of the inferred space. Legum and Herndon are placed closer to the center: Legum noted in a relatively nuanced critique that the signees of the letter are not silenced in any way, while Herndon published several ironical tweets about the letter. Interestingly, the division of clusters visible in the layout is not as pronounced as in the spatialisation of the network with ForceAtlas2 and Yifan Hu (see Fig. G.5 in Appendix G), a finding that calls for further systematic investigation.

Survey data With the weighted layout, not only generic network data can be spatialised, but also surveys: There, evaluated items as well as respondents are nodes, and forces only exist between items and individuals.

In Fig. 3.7, we visualize a survey where respondents were asked about their attitude towards six different energy-generating technologies [134];[20] Gas and coal power stations, onshore and offshore wind stations, biomass power stations, and open-space photovoltaics (which we refer to as solar in Fig. 3.7).

[18] The more likely minimum displayed in Fig. 3.5 is also the politically more plausible one: In Appendix G, *SPD* is placed closer to *FDP* than *CDU/CSU*, while the latter two parties have more commonalities (especially when it comes to economic policy).

[19] See also `https://blog.twitterexplorer.org/post/harpers_letter/`

[20] The responses represent the initial attitudes of respondents with respect to the technologies before being confronted with several pro and counter arguments. Responses were initially given on a nine-point, and for our purposes aggregated to a three-point scale.

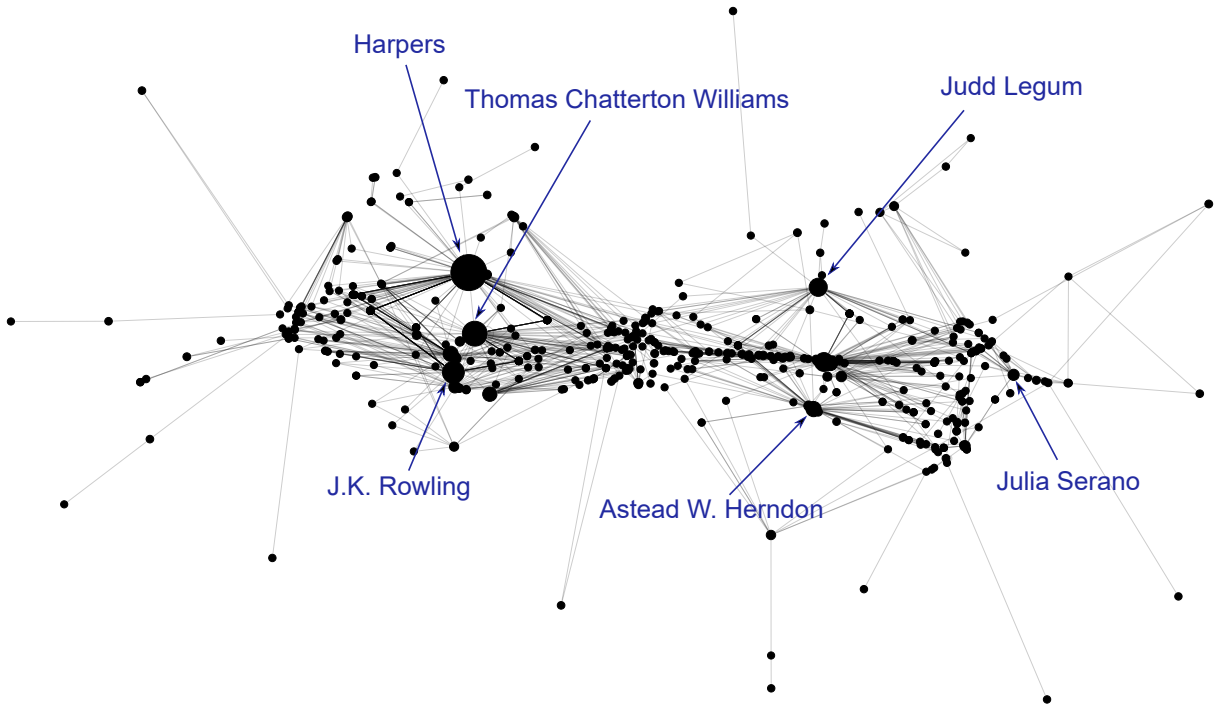


Figure 3.6: Retweet network of a debate on Twitter about a letter on free speech published by Harper's magazine. A two-camp division is visible, where the left pole includes the magazine as well as prominent signees, while the right pole contains critics. Comparison to ForceAtlas2 (Fig. G.5) reveals that the layout at hand here shows a less pronounced gap between poles.

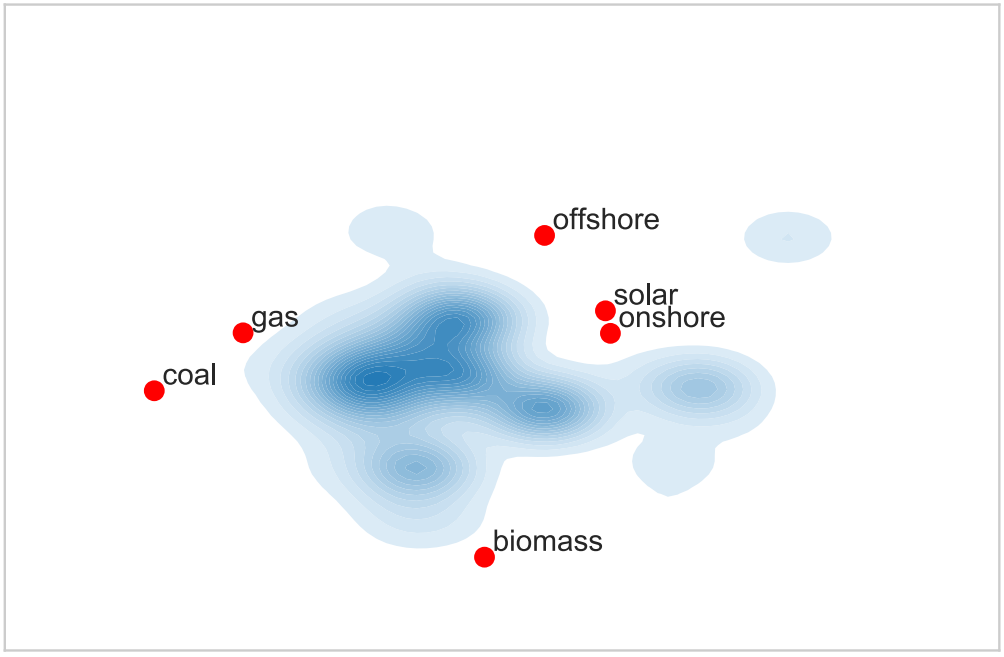


Figure 3.7: Visualization of a survey on six different energy-generating technologies. The distribution of respondents is plotted as a density in the background (the darker, the denser they are distributed in an area). Respondents are distributed close to gas, renewable energy-generating technologies, and between them. Two technological axes are visible: One from coal and gas to the renewables, and one among technologies using renewable sources of energy, with onshore and solar occupying central positions, while offshore and biomass are located opposite of each other.

The distribution of respondents over the inferred space, given by a density plot (the darker the color, the more respondents lie in a region of the layout), shows that the vast majority of respondents is located between gas and onshore, solar and offshore energy-generating technologies, while coal is placed far away from most respondents.

Several density peaks exist: One between gas and biomass, one placed rather centrally between gas, offshore, solar, onshore and biomass, two between biomass and solar/onshore, and one at the margin of the space, but closest to offshore and solar/onshore technologies. Even more interesting is the arrangement of technologies themselves, since it shows that collectively, response profiles of individuals create two orthogonal axes along which technologies are placed: One axis is visible from renewables towards technologies relying on fossil sources of energy (gas and coal). On the other hand, renewable sources of energy are distributed along an perpendicular axis. Onshore and solar occupy central positions there, while offshore and biomass are located opposite of each other. The respondents' distribution and the arrangement of technolgies are in line with the average ratings and rating correlations between individuals (see Appendix G.3). Average ratings for coal are reported to be significantly lower than for any other technology in [134], gas receives a neutral rating, renewable technologies (offshore, onshore, solar, biomass) are rated positively on average. Biomass receives the lowest average rating of the renewables, which is reflected in the distribution of respondents. Biomass has, among the renewables, the weakest correlations with the other renewables. On the other hand, ratings are not negatively correlated with coal or gas. This is mirrored by its placement in Fig. 3.7.

3.6 Discussion

While FDLs are frequently employed for network visualization across a variety of scientific disciplines, they lack theoretical grounding which allows to interpret their outcomes rigorously. We have presented a path towards interpretable FDLs based on latent space models of node interactions. We have derived force equations for a FDL that serve as maximum likelihood estimators of said models for different modes of interactions: Unweighted, cumulative, and weighted networks. Exemplary spatializations of several real-world networks show that important properties of the networks (assessed through different network measures) are reflected by node placement. Moreover, commonalities with, but also differences to existing algorithms, specifically ForceAtlas2 [73], have been pointed out: ForceAtlas2 exhibits a stronger separation of nodes within tightly connected clusters for the unweighted case and of retweet clusters between each other.

In general, the derivation of forces given can serve as a blueprint for the creation of FDLs based on alternative interaction models. The baseline interaction model might in principle even be accessible to empirical validation. The present approach can also be used to motivate parameter choices for already implemented FDLs, such as ForceAtlas2: The degree of influence of edge weights there, for example, can be arbitrarily chosen. But the choice could be guided by agreement with the algorithm implemented here.

Certain limitations remain: The convergence of FDLs to local minima is a problem

that can not be solved by the present approach. The Bundestag follower network, for instance, possesses several equilibria for which the parties were allocated in different order, both for Leipzig Layout as well as the three algorithms it was compared with. Nevertheless, the underlying model of Leipzig Layout allows a comparison of the log-likelihood of several equilibria, out of which the most likely can then be chosen (which was the one displayed in Fig. 3.5). This is not possible for existing FDLs. Moreover, the role of dimensionality for the outcomes of latent space inference in general has not been studied systematically [93] – with respect to network visualization, an extension to a three-dimensional latent space would be of interest in comparison to the two-dimensional case studied here.

3.7 Summary

Force-directed layout algorithms are ubiquitously-used tools for network visualization across a variety of scientific disciplines. However, they lack theoretical grounding which allows to interpret their outcomes rigorously. We proposed an approach building on latent space network models, which assume that the probability of nodes forming a tie depends on their distance in an unobserved latent space. From such latent space models, we derived force equations for a force-directed layout algorithm. With this approach, force-directed layouts become interpretable, since the forces infer positions which maximize the likelihood of the given network under the latent space model. We implemented these forces for (un-)directed unweighted and weighted networks. We spatialised different real-world networks, where we found central network properties reflected in the layout. Comparison to existing layout algorithms (not grounded in an interpretable model) revealed that node groups are placed in similar configurations, but that the established algorithms show a stronger intra-cluster separation of nodes, as well as a tendency to separate clusters more strongly in retweet networks.

Chapter 4

Voice, and silence, in public debate on Twitter

4.1 Twitter's reach beyond platform borders

Twitter is an immensely popular object of study for social scientists in a variety of contexts, ranging from politics [74] to popular culture [65] to crisis communication [29]. One reason for this popularity is that Twitter is open in a double sense: On the one hand, researchers can call Twitter data conveniently via an API. On the other hand – and more importantly – the content created on the platform is public by default. *In principle*, a user's activity is visible to everyone on the platform, and any user can interact with anyone else. Due to its open platform design, user interactions on Twitter might, of all major social media platforms, come closest to what is commonly referred to as 'public debate.' While not being representative of the general public [99, 100], Twitter provides a public arena for information gathering, opinion formation and persuasion. Since journalists incorporate the platform in their daily routines [27, 118], explicitly refer to content visible on Twitter as public opinion [97], and even tend to judge tweets as newsworthy as press agency reports [98], the standpoints that are prominently featured there are reinforced in traditional media. A better understanding of how different opinion groups shape debate on the platform is therefore highly important: The image created there not only affects how public opinion on certain issues is perceived by its users, but by society more generally. Certain standpoints – advanced by committed minorities in particular – might appear more prevalent than they actually are.

In this chapter, we attempt to assess to what extent group-level differences in willingness of opinion expression on the platform influence the visibility of certain standpoints. Our goal is both methodological and case-oriented: First, we propose a novel method to make systematic differences in the engagement of groups with different political leaning in public debate visible. Secondly, we employ this method in two case studies covering two political events in Germany.

To this end, we firstly choose a suitable theoretic underpinning for the concepts of public debate and public opinion in online environments, which we connect to find-

ings on user comments and their effects on readers. Then, we describe how relevant tweets are collected. After general comments on the data gathered from Twitter and ethical considerations in light of the data available and the analyses conducted, the data transformations which yield a social-structural view on interactions between users are expounded. The proposed method relies on an interplay of network representations of two types of user interactions on Twitter: Retweets and replies. Retweet networks are used to discern opinion groups, while reply networks make it possible to assess how these groups participate in public debate. We construct said networks from a user-centered data collection for two events: A state election in the German state of Saxony and a violent clash on New Year's Eve between police and parts of the population in the city of Leipzig. We show for both cases that while retweet networks are strongly polarized, debate between users of different opinion clusters is vivid. We also show that, while being a minority in number, Twitter users who mostly share content of right-wing parties and politicians are disproportionately active in debate and act more confrontational in the sense that they address users from different opinion groups more often.

4.2 Theoretical considerations

4.2.1 Public debate and public opinion

Before we quantitatively assess which opinion groups form public opinion and potentially dominate public debate on certain topics on Twitter, we first need to clarify what exactly we mean by these terms in the context of this work.

A communication-based view on public debate has been provided by Vincent Price. He describes public debate as "communication processes through which publics are constituted and within which opinions on public affairs are formed" [125, p. 74]. While he invokes the analogy of a big town meeting, the technical feasibility of creating such a meeting still seemed out of reach in the early nineties: "Modern communication technologies may have enabled the enlargement of public consciousness [...] but they have not come close to creating any sort of town meeting at large" [125, p. 78].

With the advent of social media, the analogy appears to have become a (digital) reality. As has been stressed in the introduction, Twitter is a public medium and allows its users to interact with potentially anyone else on the platform. Users can share others' thoughts, put out their own, and organize around hashtags, thereby creating publics and attracting attention of others [30]. These processes are strongly reinforced and amplified by traditional media: Journalists incorporate social media, especially Twitter, as an established news source in their daily routine [118, 27, 98], journalistic content or events on television are discussed on the platform in parallel [149, 58], and Twitter content is often explicitly used to represent public opinion, both in qualitative (by quoting certain tweets, e.g. to underline meta-narratives) and quantitative fashion [97]. Often, social media platforms themselves provide tools or even supply journalists with data or analyses in order to get mentioned in their articles [97].

Two different, basic paradigms of public opinion can be subsumed under the terms *discursive* and *demoscopic* public opinion [125, 132]. The former refers to public opin-

ion as a social-structural phenomenon and has a strong normative imprint. The process of arriving at public opinion – public debate – is understood as as a rational discourse between well-informed citizens [132], and should lead to the best possible decision with respect to the overall good. The latter is related to survey research where scientists seek to aggregate the attitudes of individuals towards certain issues in a representative fashion, which yields, by majority rule or a breakdown of percentages, public opinion.

For an understanding of online interactions, both conceptions are problematic. There have been attempts to replace classical voter surveys for elections by social media observations [150, 31], but the findings have been contested by others [74] or turned out to be incorrect (Burnap et al. [31] predicted a Labour win in the 2015 UK elections). Discursive public opinion as a normative concept, on the other hand, is in general hardly accessible to empirical research. And, after all, the internet does know many compulsions besides "the unforced force of the better argument" [62, p. 306].

Elisabeth Noelle-Neumann has proposed a social-psychological approach centered around *observable* opinion expressions.[1] She conceives public opinion – or rather, what people *perceive* as the opinion of 'the' public – as a force of informal social pressure and control manifesting itself in "approval and disapproval of publicly observable positions and behavior" [115, p. 64]. Her operational definition of public opinion incorporates "opinions on controversial issues that one can express in public without isolating oneself" [115, p. 63]. Especially for controversial topics, individuals, being social creatures and fearing social isolation, constantly and mostly sub-consciously monitor their social environment and the mass media. They estimate the majority opinion around them, employing some "quasi-statistical sense" [114], which they then refer to as public or popular opinion. The theory hence puts strong emphasis on the role of the subjective impression of public opinion of individuals.[2] Quantified interactions in online environments (Twitter displays the number of likes, retweets, and replies below each tweet) might suggest themselves as an objective foundation for the quasi-statistical impression of the opinion climate – but as we will show, these interactions themselves can be subject to strong biases.

In the following, we will be concerned with public debate as a (structural) communicative process, and public opinion in Noelle-Neumann's sense as publicly visible opinions.

4.2.2 Comment spaces and perceived public opinion

In order to capture public debate online, comment sections, predominantly of news websites, have been the target of attention since their introduction [139, 88, 87, 50, 146]. Studies found that user comments on news articles affected individuals' per-

[1] 'Expressions' is interpreted very broadly: Noelle-Neumann's account includes also non-verbal modes of communication, e.g. badges that support certain political parties or even subtle facial expressions such as raised eyebrows.

[2] As has already been addressed in Ch. 2, Noelle-Neumann's spiral of silence theory states that if people realize that they hold an opinion that differs from their impression of public opinion, they tend to be less willing to express their opinion publicly any longer. This, in turn, affects the perception of public opinion of others, potentially setting off a spiralling process in which certain groups become more expressive while others fall silent.

ceptions of public opinion [88, 87] – more so than simple comparisons of likes and dislikes.[3]

Allowing differing points of view to reach an audience with few formal constraints, comment spaces have also been interpreted as counterpublic spaces (or spheres) [146, 76], an expansion of the Habermasian concept of the public sphere. Nancy Fraser originally defined counterpublics as "parallel discursive arenas where members of subordinated social groups invent and circulate counterdiscourses to formulate oppositional interpretations of their identities, interests, and needs" [48, p. 67]. They arise in response to hegemonic "publics at large" [48, p. 67]. Comment sections in general (may they be the comment sections of newspapers, or the reply thread of a tweet of a public figure on Twitter) are very suitable for the formulation of oppositional interpretations: They are in the direct vicinity, nevertheless clearly demarcated from the interpretations and content they want to distance themselves from [146].

Thus, comment sections (i) have a significant effect on how people perceive public opinion on an issue – in Noelle-Neumann's terms, the "opinion climate" [115] – and (ii) they are hence, for all kinds of interest groups, important arenas for confrontation and contestation of certain opinions, standpoints and narratives, may they be hegemonic or not. Therefore, a careful investigation of comment sections and the views expressed there is important: Which standpoints are expressed, are they (and by whom are they) challenged, and which viewpoints (or users) remain silent?

That different groups strive for the award of being called *the* public is nothing new. As Baker notes of the pre-revolutionary times in France: "Indeed, one can understand the conflicts of the Pre-Revolution as a series of struggles to fix the sociological referent of the concept in favor of one or another competing group" [11, p. 186]. Online environments, which facilitate communication and decentralize information distribution, might appear to make this competition more transparent. But they also introduce additional potential for misperceptions, not least due to differing willingness of public opinion expression of different groups [78, 50, 103].[4] For certain opinion groups, this can lead to a False Consensus Effect [130], according to which individuals see their own opinions as more prevalent in society than they actually are. On the other hand, groups less willing to express their opinion might underestimate their size (False Uniqueness Effect, see [102]). The method in this chapter will allow an estimation of which opinion groups principally shape the impression of public opinion on the platform and the extent to which these opinion groups interact with each other.

[3] In the 20th century, scholars often distinguished between representatives of interest groups that debated publicly, and a large and more spectator-like 'body' which then reacted to the debate and approved or disapproved, hence formed public opinion [125, p. 27]. This relation is reproduced by the combination of newspaper articles and user comments below, which allow directly visible engagement in larger amount and with less control than the very limited number of redacted letters to the editor.

[4] [103] develop a method to investigate the difference in tweeting behavior between what they refer to as a 'silent majority' on the platform as opposed to a 'vocal majority'. They define the silent majority as users who only tweet once, while the vocal users tweeted more than 50 times. The method developed in the course of this chapter allows for an estimation of the size of the group of users who do not tweet at all (in the form of replies).

4.3 Political background

The two events under consideration were the Saxon state election which took place on September 1st, 2019, and a violent clash between police and parts of the population in the city of Leipzig on New Year's Eve four months later (in the following abbreviated with NYE). The events are complementary examples in the sense that the election was long-anticipated, while the latter occasion was a spontaneous incident.

The election was of special, nation-wide interest. Since Saxony had been the birthplace of the anti-Islam movement *Pegida* in 2014, which received international attention, the election was considered a litmus test for the mobilizing potential of extreme forces. Before the election, it was not clear whether there would be the possibility of forming a majority coalition without participation of the *Alternative für Deutschland* (*AfD*), a right-wing party founded in 2013, or the democratic socialist party *Die Linke*. While the Christian democratic party *Christlich Demokratische Union Deutschlands* (*CDU*) leading the polls ruled out coalitions with both of the parties, it was publicly discussed whether parts of the *CDU* were open towards collaboration with the *AfD* [128].

The NYE incident was not anticipated, but a spontaneous event which was subsequently discussed not only in Saxony, but also in national politics. On New Year's Eve, violent riots and attacks on at least one police officer occurred in the city of Leipzig's quarter Connewitz. The event was particularly polarizing: While some political actors framed the incident as an example of the violent potential of left-wing extremism, others accused the police of deliberate provocation [45].

4.4 On Twitter and data ethics

Twitter[5] is a microblogging and social networking service, where users can post short messages, called tweets, of a length of up to 280 characters. Twitter has become one of the most-used social networks in the world, with around 206 million monetizable active daily users on the platform as of July 2021.[6]

Twitter maintains an application programming interface (API), through which also scientists can receive data from the platform. Recently, an Academic Access track has been established, the free basic version of which allows the retrieval of 10 million tweets per month.[7] Nevertheless, tweet volume above a certain threshold is not freely available to researchers. At the time of data collection for this chapter (May 2019 to February 2020), data was retrieved via the Streaming API (now closed by Twitter), which allowed free streaming of at most 1% of all the tweets produced on Twitter at a given time. Requests which exceeded this limit were provided with a supposedly randomly chosen subsample of tweets capped at 1%. Problematically, Twitter's sampling techniques can produce misleading results (compared to data that has been collected

[5] `https://twitter.com/`

[6] According to Twitter's second quarter 2021 press release [2]. Twitter defines monetizable active daily users as people, organizations, or other accounts who logged in or were otherwise authenticated and accessed Twitter on any given day through `twitter.com` or Twitter applications that are able to show ads.

[7] `https://developer.twitter.com/en/products/twitter-api/academic-research`

Table 4.1: Modes of interaction on Twitter.

Interaction	Description
Retweet	Verbatim sharing of another tweet, indicated by 'RT @username' preceding the tweet.
Quote	Sharing of another tweet with a comment attached to the quoted content.
Reply	Tweet that is a direct response to a previous tweet.
Follow	Subscription to content of other users.
# Hashtag	Phrase preceded by '#', which usually serves as a tag for the content.
Direct message	Direct message to another user, not visible for others.

in its entirety) [101, 122]. The two cases under investigation in this chapter did not surpass this limit, which made it possible to analyse the complete data set returned by the search terms we employed.[8]

In the following, our analysis will be based on two types of user interactions on the platform: Retweets and replies. By retweeting, a user simply copies and shares an already existing tweet (primarily with her own followers), marked by 'RT @username' preceding the tweet. A reply is a tweet which constitutes a response to a previous tweet, and the immediate addressee is indicated by an '@username' preceding the reply. An overview of different interaction types is given in Table 4.4.

For this study, digital trace data produced by Twitter users was analysed in order to detect large-scale interaction patterns between opinion groups. Research conducted on the basis of Twitter data comes with certain ethical challenges. Users consent to their information being shared and used by third parties by agreeing to Twitter's Terms of Service agreement [3], and, as has been stated above, the actions of users on Twitter are public by default.[9] However, some users might not be fully aware of the potential reach of the content they post, as the distinction between public and private spaces online remains blurred for many [5]. Moreover, we classified users into different opinion groups due to their retweeting behaviour, which served as a proxy for political positions in political debates. Even if users agree in principle that their content is used for research, they might disagree with the assignment of a certain political stance on an issue by some method of inference. Therefore, we only identified influential political figures in our networks (such as leading politicians, institutions, or political parties). We interpreted our results only on a group level, and as with all statistical analyses, they allow only limited conclusions for the behaviour of concrete individuals, and are not intended for this purpose.

While privacy concerns are to be respected by scientists, reproducibility of results is of large importance, as well. In accordance with Twitter's guidelines, we provide the

[8] Still, a point must be noted here which should always be kept in mind when working with digital trace data: If private companies provide data to researchers, the data is first and foremost a product to these companies. Its collection as well as the environment in which the data is produced by users [147], is strongly shaped by their business interests, e.g. by the desire to extract identity-related information [152] in order to optimize ad revenue. Researchers have little influence on the data that they receive which, in turn, is influenced design and algorithmic choices of the platforms. Less so do researchers have insight into how the data is put together by the platforms [101].

[9] Users can opt out of this by setting their tweets as 'protected,' thereby only allowing a specific set of users to access them (and their profile information). Moreover, users can (auto-)delete their tweets after a certain period of time, and restrict who can reply to their tweets.

tweet IDs of all tweets used in this chapter, available at [52]. Since only the IDs of the tweets are given, the data set needs to be 'hydrated,' i.e. tweets must be collected again through the Twitter API. If tweets from the data set have been deleted in the meantime, they will not be retrievable, and hence, a user's decision to withdraw content from the platform is respected by this procedure.

4.5 Methods

4.5.1 Data acquisition

The data was collected in a user-centered approach – that is, all tweets that were produced by a seed set of users were gathered. Moreover, all tweets containing the Twitter handle of one or more of the seed users were collected – this included retweets, mentions, and replies to the users. With this method, not only first-order replies to a user in question could be collected, but any part of a reply tree that had been initiated by the user (see Fig. 4.2). In the months preceding the Saxon state elections in September 2019, a first seed set was constructed. Candidates in elections in Germany appear on electoral lists. We collected all names from these lists and checked whether the candidates had an actively maintained Twitter account. If this was the case, we included them. We also included

- state party accounts
- the leaders of the parliamentary groups in the state parliament and the local party organizations
- the Saxon members of the national parliament
- the Saxon members of the European parliament
- Saxon media accounts
- Saxony correspondents of national media

that had an active Twitter account. The seed set was expanded with a snowball-sampling method: In each sampling step, users that were not contained in the seed set but retweeted or mentioned at least once a week by users that were already in the seed set and were related to Saxony were included. The latter criterion was necessary to exclude nation-wide media accounts and national politicians. At the end of July, after seven iterations, the final seed set consisted of 270 users.

Tweets were gathered until February 2020. This allowed observation of other Saxony-related events, such as the NYE incident. Since the seed set was not perfectly tailored towards this event, we restricted the analysis to the subset of tweets containing certain event-specific keywords (*connewitz, antifa, polizei, polizist, le0101, linx, leipzig, Not-OP, notop, linke, chaoten, angriff, le3112, randal,* applied to the root tweets and incident-specific retweet network). The tweets used for the analysis of the election stem from the time period between the 25th of July and the 10th of September (364,626 tweets). For the NYE data set, tweets from December 31 until January

19 (130,685) were used. The two cases were chosen since one (NYE) represents a spontaneous event, while the other a long-anticipated election – a suitable test whether similar effects can be observed for both types of events. Moreover, the data sets were of considerable, but still maintainable size (especially with respect to visualization).

4.5.2 Network representations

Two types of user interactions in the data set were represented as networks in our analysis: retweets and replies. Retweet networks were used to discern different opinion communities on Twitter, while reply networks made it possible to assess how these groups participated in public debate.

Retweet networks

Retweet interactions are represented as a directed network in which every node is a user. A link is drawn from user a to user b every time a retweets b. It has become standard practice to employ community detection algorithms to find strongly connected clusters in a retweet network which are then interpreted as groups of users roughly sharing an opinion or political position [36, 37, 55].

It has already been expounded in the previous chapter that an alternative approach to community detection algorithms is the spatialisation of (e.g.) retweet networks via some force-directed layout algorithm (FDL). The advantage of FDLs compared to clustering methods such as modularity maximization is that nodes are assigned continuous positions in a layout, while the latter have to sort the nodes into discrete partitions. After spatialisation with FDLs, users might be partitioned into different groups, as well. But still, relative distances between these groups can be assessed. Moreover, one can, with FDLs, *visually* distinguish tightly-knit clusters and less dense in-between regions of nodes in the layout. This is useful especially in societal discourse: Public debate and the different opinion camps usually cannot be clearly demarcated from one another, and regions of transition between different groups, that do not clearly belong to any one of them, are politically meaningful and should hence be discernible. For the analysis, we used the spatialisation of the retweet networks with ForceAtlas2 (in its Gephi [17] implementation).[10] This makes analysis more coarse-grained than modularity maximization in one sense, but more nuanced in another: Usually, there are fewer communities, but the communities represent more fundamental divides and it is possible to discern the less dense regions between poles, which can be treated as separate clusters.

Retweet networks of both events showed polarized structure in the force-directed layout. Upon visual inspection, we divided the retweet networks into three different regions. We classified the two cohesive poles of the retweet network as two opinion clusters (the borders chosen as is visible in Fig 4.1 A and B), and assigned the in-between region to a different cluster.

[10] While it would have been desirable to spatialise the networks with the force-directed layout algorithm developed in Ch. 3, the current implementation does not allow the visualization of networks of the size under investigation here, with far more than 10,000 users in their giant components.

Reply trees and reply networks

Due to the user-centered data collection, it was also possible to retrieve an exhaustive collection of all replies that were initiated by posts of the seed users. A post together with all its replies can be represented by a reply tree (see Fig. 4.2). Only taking into account reply trees initiated by the prominent seed users corresponds quite naturally to the distinction of the previous section between representatives of interest groups debating publicly and the spectator-like body approving or disapproving subsequently [125]. These reply threads then function as spaces where different opinion groups can confront each other: They are widely visible due to the prominence of the creator of the tweet which spans up discussion, and can hence attract users of different opinion camps. Retweets, on the other hand, mainly serve to share information with one's followers. A retweet might point to a debate, but does not imply involvement in it.

In order to gain a global view on public debate, we aggregated the combined interaction structure of all reply trees into one reply network, assigning a directed edge between two users if one had directly replied to the other in a tree (see Fig. 4.1). Obviously, trees are networks, too. But if we in the following speak of reply networks, we mean the bigger networks constructed in this procedure.

Some works have taken similar routes by taking into account direct user interactions in the form of mentions [37] and replies [138, 7, 163]. Sousa, Sarmento and Mendes Rodrigues [138], and Yardi and boyd [163], use a keyword-based tweet collection. This approach is useful if one is solely interested in tweets that include a certain keyword, while full conversations in a reply thread between users are not accessible with the method. Aragón, Kappler, Kaltenbrunner, Laniado and Volkovich [7], and Nuernbergk and Conrad [117], employ a user-centered collection and construct a reply network, but only between politicians on Twitter and hence do not capture debate among a more general public. Since the data sets in the present contribution include the complete reply trees below each post of one of the seed users, it was possible to gain a more general perspective on public debate that did not only include certain elites.

The classification of users from the retweet network – the information about whether they belonged to the minority or majority pole, or the in-between region – was imported into the reply trees and networks. This made it possible to investigate how many users of the different retweet clusters were also involved in public debate, hence willing to express their opinion in discussion with others of possibly different opinions, and whether users of different opinion clusters debated mainly among each other or with others.

It must be noted here that not all users involved in debate were present in the retweet network. Hence, the classification in the reply trees and networks was not complete. Initially, around 47% of the users involved in debate in the election data set could be classified (33% for NYE). In order to include more users in the classifications, a larger retweet network was additionally constructed which included all retweets from July 2019 until the end of February 2020. The overall structure of the network was similar to the incident-specific retweet networks (see Appendix H). If a user was present in the reply trees, but not present in the incident-specific retweet network, it was checked whether the user was present in the large retweet network – if so, the user was assigned the classification from this network. With the use of the big retweet network, 63%

(election) and 67% (NYE) of users present in the reply trees could be classified.

Local assortativity

A useful measure describing the tendency of individuals in a network to link to others with similar properties or attributes is assortativity [107]. The value compares (for categorical attributes) the proportion of edges that connect nodes with the same type to the proportion expected if the links in the network at hand were randomly rewired. This is given (in the directed case) by

$$r_{\text{global}} = \frac{\sum_g e_{gg} - \sum_g a_g b_g}{1 - \sum_g a_g b_g}, \tag{4.1}$$

where $e_{gh} = \frac{1}{m} \sum_{i:y_i=g} \sum_{j:y_j=h} A_{ij}$ is the proportion of edges from nodes of type g to nodes of type h, and $a_g = \sum_h e_{gh}$ and $b_g = \sum_h e_{hg}$ are the proportions of edges originating from and arriving at nodes of type g. $a_g b_g$ is the expected proportion of edges between nodes of group g if one would randomly create edges between nodes, while keeping the total number of outgoing and incoming edges for each type constant. The denominator, also referred to as the maximum modularity $Q_{\max}$, serves as a normalization constant, so that r_{global} can only take values from -1 to 1. The latter means that all edges in the network only connect nodes of the same type, while for the former, the edges only connect nodes of different type (hence, the network is strongly disassortative). It has been argued that such a global view might obstruct insights into local differences between individuals or groups [120]. Networks of significantly different local mixing patterns might be subsumed under the same global assortativity coefficient.

In the examination of reply networks, we ask how opinion groups differ in their interaction patterns. Hence, we are in need of a measure of more granularity in order to compare assortativity coefficients *between* groups, and not only receive one coefficient for the whole network. One might propose to simply count how often a user interacts with others from her own group and compare that to the interactions with other groups. But, as will become visible shortly, the different opinion groups in the case studies are of different sizes. And if one group is significantly larger, one would expect that this group gets addressed more often than the others, even if interactions are random. The measure of *local assortativity* has been proposed in order make such differences visible – it takes into account these differences in group sizes by comparing the actual interaction patterns on a single-node level with the ones that would arise if the network was randomly rewired with node degrees preserved.

Each node l in a network is assigned a local assortativity score $r(l)$, such that differences in the score can be compared across all nodes. Local assortativity $r(l)$ is defined by the equation [120]

$$r(l) = \frac{1}{Q_{\max}} \sum_g (e_{gg}(l) - a_g b_g), \tag{4.2}$$

with $Q_{\max}$ again serving to normalize the assortativity coefficient. The proportion of edges $e_{gg}(l)$ that connect nodes of the same type g in a local environment of node l is,

just as for global assortativity, compared to $a_g b_g$. For arbitrary groups g and h, $e_{gh}(l)$ is given by

$$e_{gh}(l) = \sum_{i:y_i=g} \sum_{j:y_j=h} w(i;l) \frac{A_{ij}}{k_i^{\text{out}}}. \tag{4.3}$$

k_i^{out} is the out degree of node i.[11] The local environment of l does typically not only consist of its direct neighbors, but also by nodes only indirectly connected to l.[12] The extent to which they contribute to $e_{gh}(l)$ is governed by $w(i;l)$, a distribution of weights over nodes that defines the local environment of node l. We follow [120] in choosing the personalized PageRank vector (see Appendix I for the iterative equation as well as the solution for the following integral), which is the stationary distribution $w_\alpha(i;l)$ of a random walker starting at l that returns to l with a probability of $1-\alpha$ at each step. The choice of α is arbitrary – if $\alpha = 1$, the random walker never returns to l, while for $\alpha = 0$, only the immediate neighbors of l are taken into account. Either extreme can be problematic: the former is uniform across the network, while the latter might be based only on a small subset of nodes. In order to incorporate mixing patterns across all scales (i.e. an α value in the whole range of 0 to 1), a multicscale distribution over all possible values of α is calculated according to

$$w_{\text{multi}}(i;l) = \int_0^1 w_\alpha(i;l) d\alpha. \tag{4.4}$$

The local assortativity coefficient is also capable of including incomplete metadata – since in the current data set, not all users could be classified, this feature is beneficial. In the histograms displayed in Fig 4.4 and Fig 4.5, the node contributions to the histograms were adjusted according to the weight

$$z_l = \sum_{gh} e_{gh}(l), \tag{4.5}$$

the sum of local edge counts with *known* metadata.[13]

4.6 Results

4.6.1 Retweet networks and classification

The retweet networks for both cases are strongly polarized (see Fig 4.1 A/B) in the force-directed layout.

For the election data set (31,108 users in the giant component), seed users placed in the majority pole are politicians of the parties *SPD*, *Die Linke*, and *Bündnis 90/Die Grünen* (and one politician of the *CDU*), along with media accounts (e.g. *Bild Leipzig*,

[11] In an undirected network with the weights on the nodes given by the stationary distribution of a simple random walker on the network, $w(i;l)$ is given by $k_i/(2m)$, and we arrive at the global assortativity.

[12] Since not only direct neighbors might be taken into account, the term local 'neighborhood' used in [120] is a bit misleading, which is why we have replaced it here by 'local environment.'

[13] The implementation of this local assortativity computation is available at [119].

LVZ, or *MDR Sachsen*) and left-wing activists. In the region between the two clusters, politicians of the *CDU*, *Freie Wähler* and *FDP* are placed, as well as media accounts (e.g. *MDR Aktuell*, *Bild Dresden*, or *TAG24*). The minority pole, on the other hand, includes seed users from the *AfD*, *Freie Wähler*, *Blaue Partei* and the anti-Islam movement *Pegida*. The structure of the retweet network hence quite accurately mirrors the political constellations in the run-up to the elections. Left/left-wing and eco-friendly parties are placed in one cluster and the right to far-right parties in another, while politicians of the economically liberal *FDP* and the center-right *CDU* are located in-between the two.

The users of the majority cluster made up 64.5% (20,052) of the retweet network, 23.1% (7,195) were part of the minority cluster, and 12.4% (3,861) of the users were in the region between the two.

A very similar structure, both in proportions and in political leanings, is given for the NYE incident. Some differences occur, however. The set of users placed in-between the two clusters (711 or 7.9%) includes the official account of the city of Leipzig and the account of the Saxon police, as well as one politician of the *AfD*, one from *Die Linke* and one *SPD* politician. The majority (6,010 users, 66.6%) and minority cluster (2,301 users, 25.5% of the giant component) show similar composition as in the election retweet network.

The classification of users on the basis of their position in the force-directed layout of the retweet network was taken as a proxy for the political position of the users concerning the two issues. It must be noted, however, that users of one cluster should not necessarily be interpreted as holding exactly the same opinion or political position. Rather, the clusters reflect an issue-specific fundamental political difference which is then also reflected by the classification.

A randomly selected subset of 100 users was used check whether the groups assigned to the users were plausible in the sense that the users tweeted content sympathetic to one of the parties or political figures in their assigned cluster. Out of the 100 users, 96 acted consistently with their classification, while 4 did not.

4.6.2 Reply trees and engagement

All reply trees that had been initiated by a root tweet of one of the seed users were taken into account in the analysis of the reply interactions. The reply trees themselves can be seen as representations of discussions triggered by single statements, and can exhibit arbitrarily complicated tree shape (see Figure 4.2) and have arbitrarily many participants. Not every tweet receives replies. There are 23,221 posts from the seed users in the election data set that were not replies or retweets, out of which 8,033 received at least one reply. (NYE: 2,020 posts, 897 with at least one reply.)

Reply trees can be characterized by two quantities: Their size S, which is the overall number of tweets in the tree, and their depth D, which is the length of the longest branch of the tree (Fig 4.2). Fig 4.3 shows the cumulative distribution of sizes and depths of the reply trees in the two data sets. In both data sets, around 90 percent of all reply trees have a size smaller than 10 and a depth smaller than 5. Nevertheless, reply trees can be very large: The largest tree in the election data set has 1,936 replies (NYE: 1,475), and the maximum depths are 72 and 37, respectively.

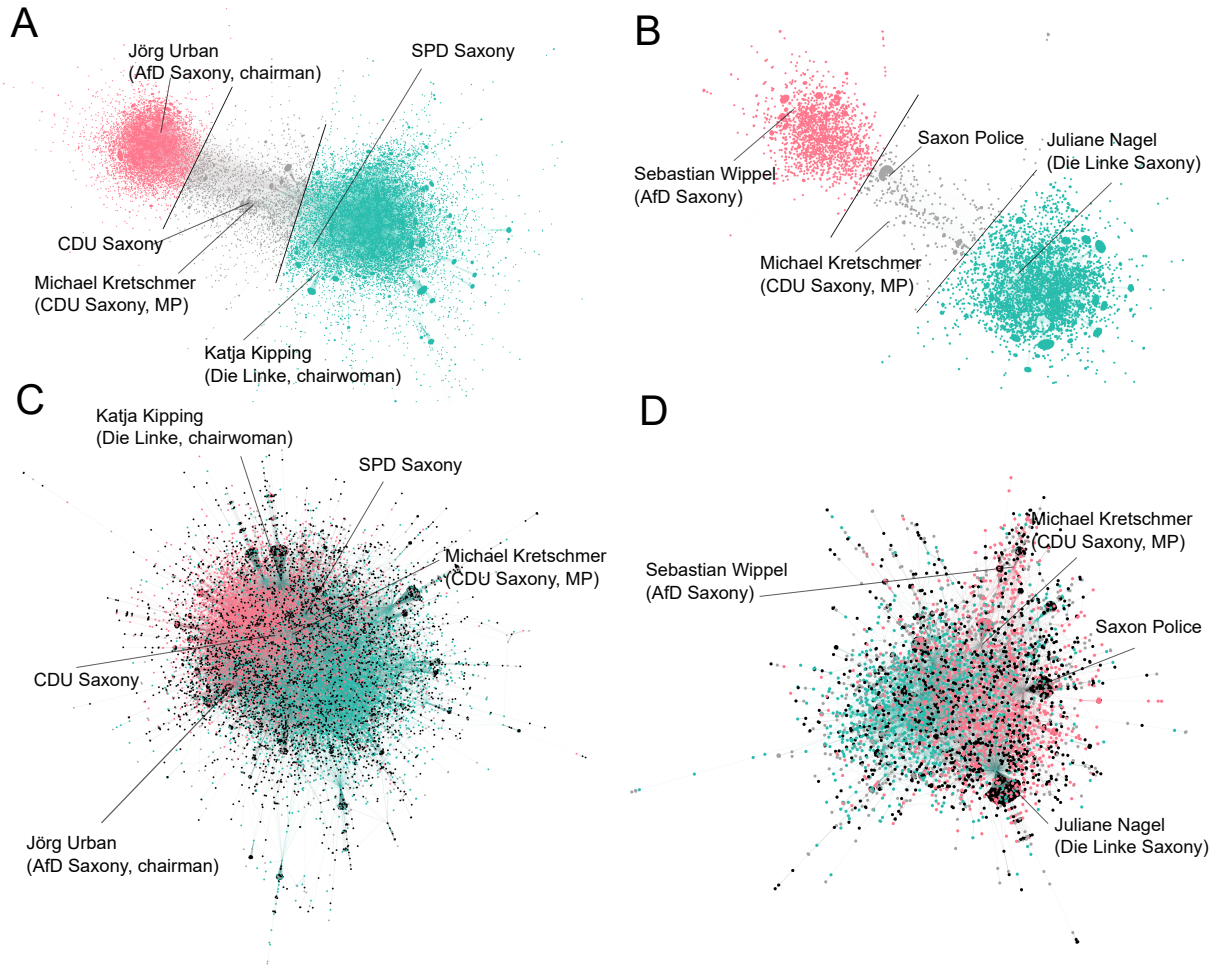


Figure 4.1: The giant component of the retweet networks (above) for the elections (A) and the NYE incident (B), and the giant component of the reply networks (C: elections, D: NYE). The retweet networks are both polarized, and users are classified according to their position in the force-directed layout. The borders for the classifications are included (black lines). The classifications are imported in the reply networks, which are not polarized. Black nodes indicate users that do not show up in the corresponding retweet network.

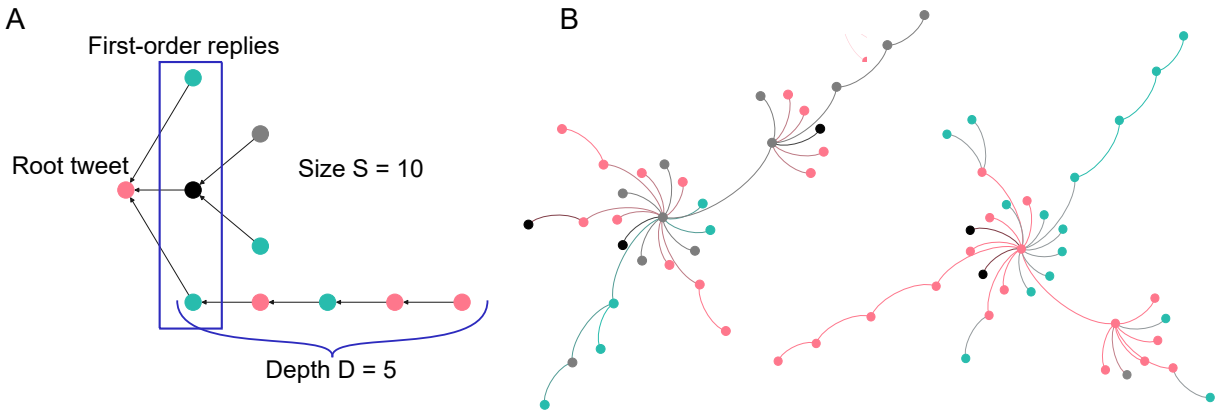


Figure 4.2: An exemplary dummy reply tree (A), along with two reply trees from the data set (B). Each node represents a tweet, and a directed edge between two nodes indicates a reply. If the users of the tweets appear in the retweet network, their replies were color-coded according to the cluster of the user (a black node indicates a reply by a user that does not appear in the retweet network). The root tweet is the original tweet by one of the seed users, while first-order replies are the direct replies to the root tweet.

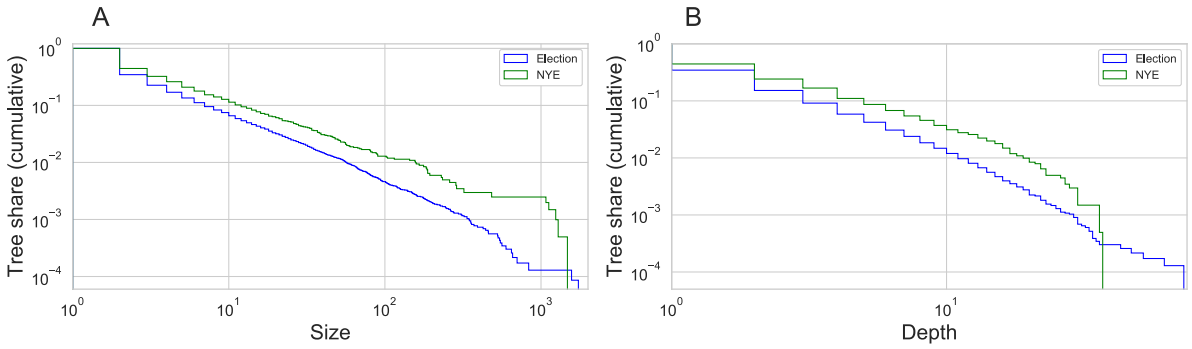


Figure 4.3: Reply tree size (A) and depth distribution (B) for the two events.

By importing the cluster classification from the respective retweet network, it was possible to compare engagement between groups in reply sections. Engagement proportions in the debate differ strongly from those in the retweet networks, both in number of users and number of tweets (see Table 4.2).

Table 4.2: Users and replies from the different retweet clusters involved in the reply trees.

	Election		NYE	
	Users	Replies	Users	Replies
Majority cl.	6,736 (33.3%)	30,615 (40.5%)	2,008 (32.4%)	6,403 (39.2%)
Minority cl.	4,696 (23.2%)	23,790 (31.5%)	1,727 (27.9%)	4,858 (29.8%)
Intermediate cl.	1,470 (7.0%)	6,785 (9.0%)	389 (6.3%)	1,235 (7.6%)
Not classified	7,395 (36.5%)	14,344 (19.0%)	2,072 (33.4%)	3,816 (23.4%)

In the election data set, participants from the majority and minority retweet clusters made up 33.3% and 23.2% of all replies, respectively. For the NYE incident, even more users from the minority retweet cluster participated in debate – 32.4% belonged the majority in the NYE retweet cluster, while 27.9% were minority users involved in debate (see Table 4.2). Users placed in-between the two poles in the retweet network also participated, but less so both in number of users and in number of replies. 36.5% (election) and 33.4% (NYE) of the users present in reply trees were not part of the corresponding retweet network (i.e. they did not retweet any of the seed users nor any tweets that contained the Twitter handle of one of the seed users). Interestingly, while these users make up the biggest number of users involved in debate, they do not constitute the majority in terms of replies. In both data sets, users from the poles of the retweet network, if involved in the debate, were most active. Users which did not retweet the seed users' tweets did not tend to debate in their reply sections – on average, in both data sets, they only wrote around two replies.

In comparison, the ratio between majority and minority pole in the corresponding retweet network was 64.5% to 23.1% (election) and 66.6% to 25.5% (NYE).

Table 4.3: First-order replies by retweet clusters. First-order replies from users of the two poles are roughly equal in number in both data sets and make up the majority of all first-order replies. The minority cluster is even more active in replies of first order than in reply trees in general.

	Election	NYE
Majority cl.	15,347 (36.0%)	2,445 (32.7%)
Minority cl.	14,911 (35.0%)	2,451 (32.8%)
Intermediate cl.	3,710 (8.7%)	572 (7.6%)
Not classified	8,629 (20.3%)	2,014 (26.9%)

First-order replies are of special interest since they (or at least a subset of them) are usually directly displayed below the root tweet on Twitter. Therefore, they most probably have a stronger impact on the perception of public opinion than tweets that are

Table 4.4: Percentage of users in the incident-specific retweet networks that are active in the reply networks (seed users excluded) by cluster. The share of users from the minority pole is, in both cases, around twice as big as the share of users from the majority pole. User share from the in-between region is slightly bigger than that of the majority pole in both cases.

	Election	NYE
Majority cl.	22.4% (4,644)	16.2% (977)
Minority cl.	48.2% (3,604)	32.2% (752)
Intermediate cl.	26.4% (1,050)	22.0% (707)

at the end of a long discussion branch. The amount of first-order replies by the different clusters is displayed in Table 4.3. First-order replies from users of the two poles are roughly equal in number in both data sets. Minority pole users hence produced an even larger proportion of highly visible replies. Users from the intermediate region in the retweet network only account for less than 10 percent of first-order replies, while users that were not present in the retweet networks produced 20.3% (election) and 26.9% (NYE) of replies of first order.

Comparing engagement in the form of replies and retweets makes it possible to assess whether the different opinion groups show different inclination to participate in the debate. To this end, we calculate the share of users present in the different clusters of the incident-specific retweet networks that were also present in the respective reply network (see Table 4.4). For both events, the groups showed significantly different behavior (election: $\chi^2 = 850.7$, $p < 0.001$; NYE: $\chi^2 = 138.2$, $p < 0.001$). Users of the minority cluster were roughly twice as likely to write at least one reply than users belonging to the majority cluster (election: z-score 41.0, $p < 0.001$; NYE: z-score 15.0, $p < 0.001$). Users from the in-between region were slightly more likely to get involved in debate than users from the majority pole (election: z-score 5.4, $p < 0.001$; NYE: z-score 4.0, $p < 0.001$), but still less than the minority pole (election: z-score 22.1, $p < 0.001$; NYE: z-score 46.0, $p < 0.001$).

Hence, two findings are worth stressing: (i) Minority pole users are disproportionately active in debate compared to the majority pole, both with respect to number of users involved and of replies written. This effect is even more pronounced in first-order replies that are, by platform design, most visible. And (ii), users occurring in the retweet network, if they take part in debate in the form of replies, do so more extensively than unclassified users.

4.6.3 Reply networks and global interaction patterns

Reply networks give a more comprehensive structural picture of public debate – it is possible to make visible patterns of interactions between different users and user groups beyond single reply trees. Each reply network was constructed by aggregating all replies in the reply trees into one network, where each node represents a user and a directed edge is created between two users if one has replied to the other.

The questions of interest here are whether the groups also exhibit large-scale po-

larization when they discuss among each other, and whether there are differences in discussion behavior between the groups. If public debate was fragmented in the sense that discussion ties were only existent amongst a certain subset of users, this would be visible in the force-directed layout of the reply network. But, as is displayed in Fig 4.1 C/D, this is the case for neither of the two events (again, spatialisation was carried out with ForceAtlas2). Users of clusters that are clearly separated in the force-directed layout of the retweet networks interact quite frequently in reply sections.

In Fig 4.4 and Fig 4.5, the local assortativity distributions of users of the different groups are displayed. The distributions are multimodal, i.e. they exhibit more than one peak. In the election data set, users of the majority cluster (A) have their largest peak at a local assortativity close to 1, i.e. they reply mainly to users of their own clusters (and, since local assortativity also takes into account the assortativity of the node environment beyond the direct neighborhood, also mainly interact with users who do the same). A second, yet smaller peak is visible with negative local assortativity (around -0.3). Hence, some majority cluster users mainly seek debate with users from other clusters. The reverse holds for the minority cluster (B): There, most of the users reply to users from the other groups. A smaller peak is visible at positive local assortativity (around 0.6). Users from the intermediate region (C) also exhibit a bimodal distribution, with one slightly negative peak and one peak at around 0.6. The NYE data shows an even more pronounced trend: Only few users of the minority cluster get assigned a positive local assortativity. Majority group users again show a bimodal distribution with one peak close to $r_l = 1$, the other peak is now at around $r_l = -0.6$.

Further insight into the interaction patterns between the different clusters is provided by Table 4.5, which shows counts of replies between and within the clusters. While users of the majority cluster of the retweet network address almost 60% (election) or 50% (NYE) of their replies to other users of their own cluster, the opposite holds for the other groups: Most of their replies are directed towards users from the majority cluster. For the election, users of the intermediate cluster replied more often to others from their own cluster than to the minority pole, while for the minority pole, replies to the own and replies to the intermediate cluster were almost equal in num-

Table 4.5: Reply interactions between and within the different clusters – the columns show what cluster the replies are from, the rows who the cluster replies to. The percentages (in brackets) are given with respect to the overall number of replies a cluster has given. Note that the column entries do not sum up to the overall number of given replies since there are also replies from users that do not show up in the retweet networks, hence could not be classified.

	Election			NYE		
to/from	Majority cl.	Minority cl.	Intermediate cl.	Majority cl.	Minority cl.	Intermediate cl.
Majority cl.	18,185 (59.4%)	10,590 (44.5%)	3,596 (53.0%)	3,234 (50.5%)	2,814 (57.9%)	821 (66.5%)
Minority cl.	5,778 (18.9%)	5,964 (25.1%)	1,035 (15.3%)	1,445 (22.6%)	827 (17.0%)	162 (13.1%)
Intermediate cl.	4,770 (15.6%)	6,025 (25.3%)	1,757 (25.9%)	1,041 (16.3%)	949 (19.5%)	191 (15.5%)
Overall	30,615	23,790	6,785	6,403	4,858	1,235

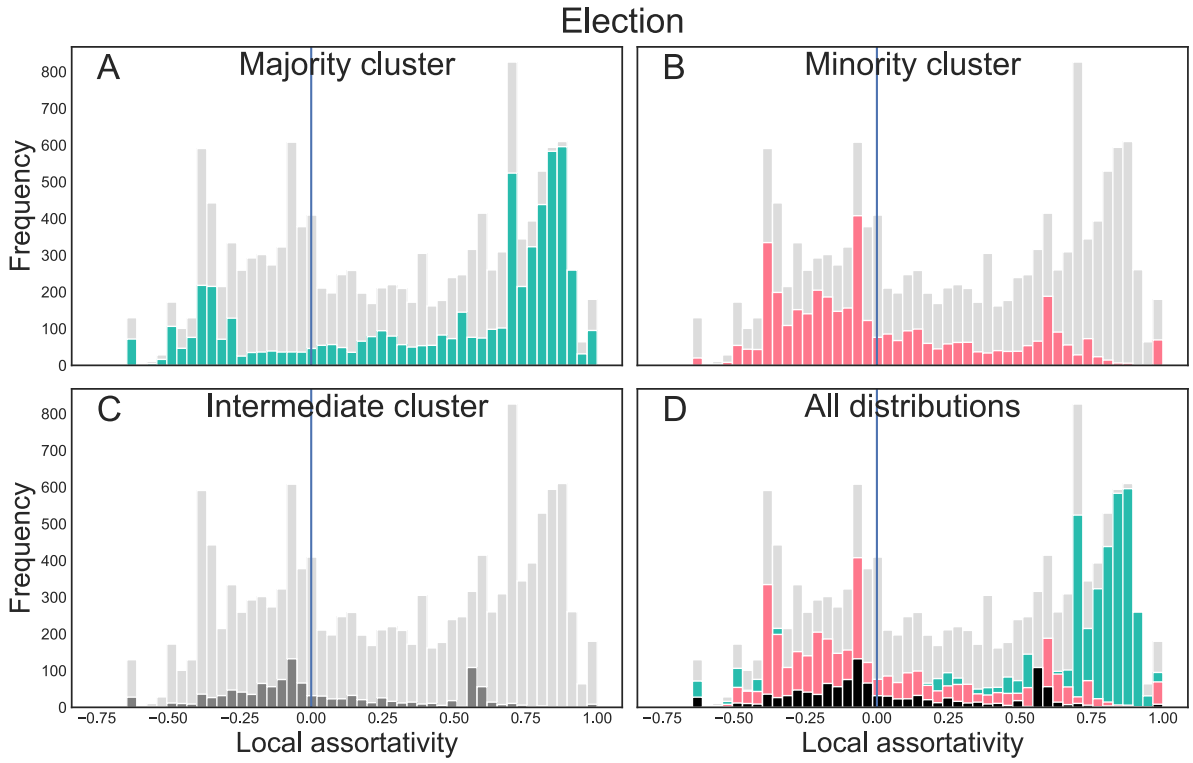


Figure 4.4: Local assortativity distribution for the reply networks of the election, split up by the users' respective retweet clusters (A: majority cluster, B: minority cluster, C: intermediate cluster, D: comparison of all distributions). The grey distribution in the background is the overall local assortativity distribution, the distribution of the respective subgroup(s) is plotted in the foreground and color-coded as before (in D, the intermediate cluster is displayed in black instead of dark grey).

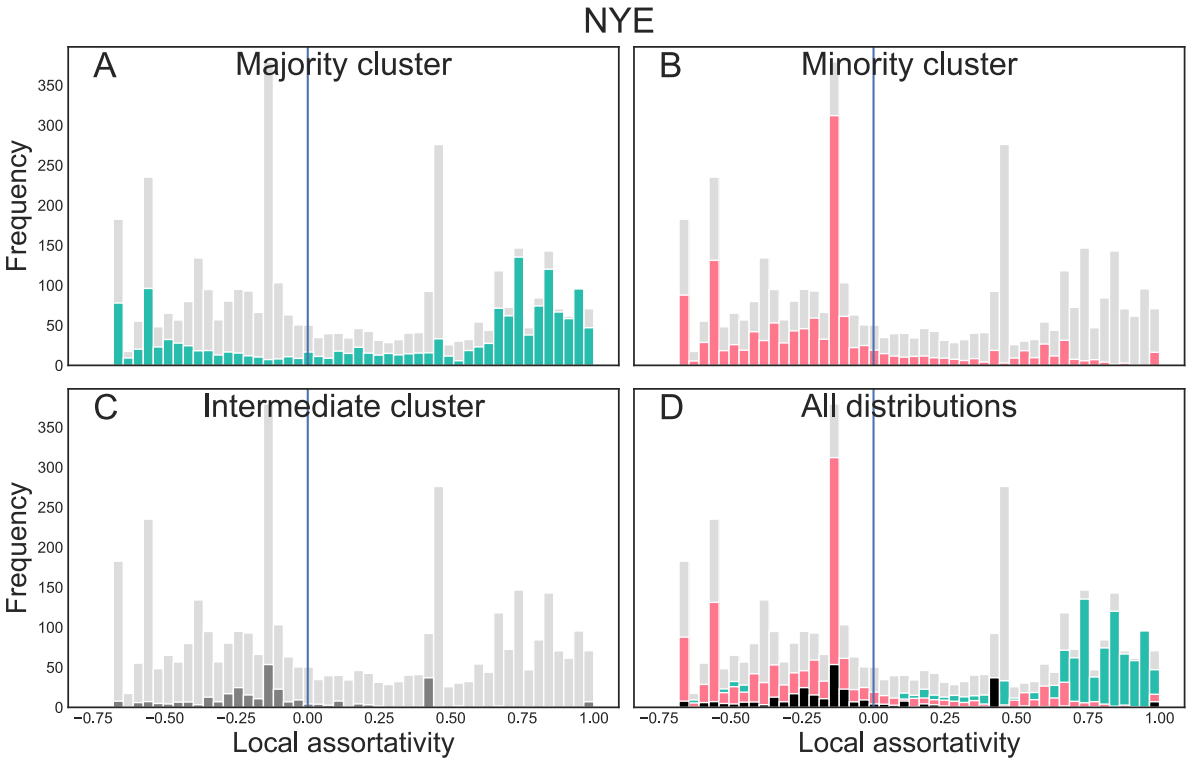


Figure 4.5: Local assortativity distribution for the reply networks of the NYE incident, split up by the users' respective retweet clusters (A: majority cluster, B: minority cluster, C: intermediate cluster, D: comparison of all distributions). The grey distribution in the background is the overall local assortativity distribution, the distribution of the respective subgroup(s) is plotted in the foreground and color-coded as before (in D, the intermediate cluster is displayed in black instead of dark grey).

ber. The number of tweets from intermediate users was considerably lower compared to tweets from the poles, hence it appears that users of the minority pole were motivated to challenge both intermediate and majority pole users. A similar tendency can be found for the NYE incident.

To sum up, interaction patterns within and between the different groups are heterogeneous: Some users from every group seek debate with the differently-minded and others show a tendency to discuss amongst their own cluster. Nevertheless, the minority pole by far shows a stronger tendency to reply to users from different clusters than the majority pole.

4.7 Discussion

Vincent Price's comparison of public debate and town meetings receives a refinement in the conclusion of [125]. He states that:

> "The democratic foundations of the concept of public opinion are indisputable; far less so are the democratic foundations of day-to-day political decisions, even when they are formed out of public debate. [...] We may well compare public debate to a town meeting – provided we keep in mind that although some town meetings enjoy free-flowing debate, there are other meetings for which almost no one shows up, at which powerful leaders and organized coalitions dominate, and at which people with minority viewpoints are shouted down or left standing outside." [125, p. 91]

The paragraph is quoted at length here since it illustrates aspects of public debate that might be amplified in online environments, specifically by social media. While the facilitation of communication bears the potential of enabling minorities and their concerns to gain public attention [146, 48], it can easily introduce systematic biases in the perception of public opinion. Online user comments are generally an important source of information for many and help in judging whether a product is good, a certain video is worth watching – or which standpoints are prevalent among a public in political debates. While the experiences and opinions of others can provide a very useful basis for decision in these contexts, naive reliance on what others express online can be collectively dangerous, especially in an era in which social media shapes politics to an unprecedented extent. If groups with certain minority opinions manage to become increasingly visible, their opinions might appear more socially acceptable and accepted than they actually are. Such disparities can be problematic since perceived public opinion, as different studies have shown, can have a persuasive and/or silencing effect [115, 114, 88, 112, 106]. Under certain circumstances, as we have investigated in Ch. 2, committed minorities might even be able to gain public predominance. In the context of Twitter, this phenomenon might be especially problematic due to its tight links with traditional media and news outlets, where Twitter content is often directly taken as representing public opinion [97] or used as a source in routine coverage [118, 27, 98], introducing either strongly biased data or leading to a potentially unjustified alarmism in coverage.

The present chapter has two main findings: In the two cases under investigation,

- disproportionately many replies come from users which constitute, if retweet networks are considered, a minority – composed of accounts by right to far-right parties, politicians, and users retweeting their content. It is hence probable that the content produced by this group influenced mere observers' perceptions of public opinion to a degree that did not reflect their real number.

- Users of different clusters also diverge in who they reply to. While users from the majority cluster tend to interact mainly amongst themselves, users from the intermediate and especially from the minority cluster are more keen on confronting differently-minded others (Fig. 4.4, Fig. 4.5, Table 4.5).

These findings can be connected to the theoretical considerations above. Specifically, an interpretation of reply sections on Twitter as *counterpublic spaces* [48] suggests itself. Not only are reply sections located in the direct vicinity of the content that some users want to challenge, but opposition can also be voiced immediately, which might serve as an incentive for the expression of fundamentally divergent opinions. Users retweeting mostly right-wing populist content – which is generally characterized by a fundamental opposition to the political 'mainstream' – exhibit a stronger willingness to express themselves in the form of replies in both case studies. First-order replies, as has been shown, are even nearly identical in number with those from users retweeting mostly left-leaning parties, politicians and contents.[14]

Previous findings parallel the results of this work: In a study from Switzerland, it has been shown that users with a right-wing political leaning engage more frequently in the comment sections of news pages [50], an effect also visible in Table 4.4 and in [146], where newspaper comment sections of articles covering the *AfD* are interpreted as counterpublic spaces. As an explanation for this effect in connection to the rise of right-wing populism, Schweiger [131] argues that misperceptions of the opinion climate, fuelled by news consumption through social media, lead to higher willingness of opinion expression for certain social groups – especially for those who already put less trust in established media sources, and those who lack awareness about the biases implicit in social media. Walter et al. [157] find that comments challenging assumptions related to anthropogenic climate change have a greater share in comment sections of newspapers for countries where public opinion converges towards the scientific consensus. They hypothesize that due to the marginalization of skeptics in broader public debate, they might, as a reaction, turn to comment sections. As opposed to these studies, which deal mostly with manual coding of the content of a limited number of comments below newspaper articles, the method presented here allows a more coarse-grained view of an in principle unrestricted amount of user engagement in reply sections.

Public opinion in the sense sketched in Sec. 4.2, assessed through reply sections, appears nearly balanced between the two poles, while retweet networks suggest that the actual opinion proportions on the platform are very different – users who share mainly left-leaning content are a majority. The model fleshed out in Ch. 2 is, in its

[14] While users retweeting center to center-right parties show similar activity patterns as left-leaning ones, they are comparatively small in number, which might reflect that this voter group rarely uses Twitter.

present state, too simplified allow connection to empirical data at hand: The interaction patterns between users from different clusters are heterogeneous (see Figs. 4.4 and 4.5). However, it has been argued above that reply sections incentivize the expression of strongly oppositional views. E.g., if users have a strong emotional reaction to certain content, they can express themselves within seconds and a few clicks. Such a reaction is more plausible for users that strongly oppose the political status quo. This could be modelled by reduced expression costs for the fundamental opposition, which enable the opinion group to speak out more easily, as we have seen in Ch. 2.[15]

With the proposed method, differences in willingness of opinion expression can be made visible. We stress that the method proposed here is not restricted to the specific cases studies. With a suitable seed set of users, any debate on Twitter can be analysed analogously. It is, nevertheless, limited in scope: It attempts to gain a comprehensive structural view on Twitter debate, but does not analyze tweet content. Moreover, a proportion of users in the data sets remains unclassified by the method since they did not appear in the retweet networks. Complementary methods of classification should be sought. Finally, Twitter is only one social media platform, which in addition is not representative of the general population [99]. A potentially insightful avenue for future inquiry might be the comparison of reply sections with comment sections of online newspapers [162]. Still, as we have argued, both Twitter's platform design as well as its echo in traditional media outlets at least implicitly award Twitter the role of the host of the big town meeting called public debate. We therefore deem it increasingly important to develop methods which enable a better understanding of which viewpoints are prominently featured on the platform, and which ones remain mostly unspoken or unheard.

4.8 Summary

We developed a method in order to investigate differences in activity and interaction patterns in public debate on Twitter. We argued, based on theoretical considerations and previous empirical studies, that the opinions that are visible in reply sections on Twitter have both a direct and mediated effect on the perception of public opinion. On the basis of a user-centered tweet collection, we investigated the network representations of two types of user interactions on Twitter: Retweets and replies. Retweet networks were used to discern opinion groups, while reply networks made it possible to assess the propensities of users from the different groups to get involved in debate.

We employed the method in two case studies, the Twitter echo of the Saxon state elections of 2019, and of violent riots in Leipzig on New Year's of in the same year. We found in both cases that (i) different opinion groups exhibit different propensities to get involved in debate, and therefore have unequal impact on public opinion. Users retweeting far-right parties and politicians are significantly more active, hence their positions are disproportionately visible. We also found that (ii) said users act significantly

[15] Connection to Noelle-Neumann's sprial of silence would imply that users fear social isolation upon their opinion expression online. An investigation of whether this is the case would require some type of survey to be conducted among commenters, which was beyond the scope of this work. The model does not incorporate this assumption, and hence can be related more easily to the findings.

more confrontational in the sense that they reply mostly to users from different groups, while the contrary is not the case.

Chapter 5

Summary and conclusion

> Was ich heut' nicht schaffe
> Schaffe ich tomorrow. *Haiyti*

This thesis dealt with modelling and observing group-level differences in opinion expression online. It was argued that the analysis of human behaviour in digital trace data needs to be conducted with a focus on explanation and theoretical considerations especially if one needs to understand how the collected data comes about – an issue which lies at the core of the question of how individual decisions of engaging in debate (or not) influence what becomes collectively visible online. Approaching this question is a necessarily interdisciplinary endeavour. Consequently, the contributions of this thesis gain their novelty through the development and innovative combination of approaches from different disciplines, based on plausible accounts of human behavior.

First, we developed a model of dynamics of collective opinion expression. That opinion exchange can not only affect opinions, but also the willingness of expressing them, has received only a marginal amount of attention in the opinion dynamics community. The model sheds light on this aspect of opinion exchange and identifies social-structural conditions for public opinion predominance of different groups. The approach was based on a game-theoretic formulation of opinion exchange between individuals. We argued that a focus on the incentive structures for opinion expression can offer a novel, insightful avenue of research. While it is an interesting framework for opinion dynamics in general (see [13]), it is especially useful with respect to opinion expression. Expression of opinion constitutes a conscious action rather than an (often partly unconscious) process like opinion formation; and in online environments, opinion expression is generally facilitated and can be rewarded almost instantly. Moreover, by grounding models in the rewards for certain actions, changes in the incentive structure, e.g. in the design social media platforms (such as the facilitation of opinion expression for certain opinion groups), can be addressed.

Starting with a game-theoretic formulation of a two-group systems on a network of stochastic blocks, where agreement (disagreement) is perceived positively (negatively), we investigated the Nash equilibria of the system. Depending on the proportion of in-group ties and the costs of opinion expression, equilibria arise in which a cohesive

minority can dominate public debate. In order to address questions of equilibrium selection and stability, we also provided a dynamical systems perspective: Using the reinforcement learning algorithm of Q-learning, we reduced the N-agent system in a mean-field approach to two dimensions which represented the two opinion groups. This two-dimensional system was analyzed in a comprehensive bifurcation analysis of its parameters. There, also a unilateral change in expression costs for one opinion group and its effect on willingness of opinion expression was analysed. Increased internal cohesion, as well as encouragement of a group to speak out can heave it into audibility. Increasing costs for all groups can lead to complete silence or an 'either-or' setting in which only one group (but not both) can be expressive. The model dealt with only one type of actions for individuals, namely the decision to express or withhold their opinion in public. A natural next step would be the combination of the model with models of opinion change, which potentially involves a differences in time scales for the two processes and differently shaped interaction networks (also illustrated by our findings of Ch. 4). In general, the network structure of stochastic blocks was useful for reasons of mathematical accessibility, but should be replaced with more realistic or empirical social networks. Moreover, more than two opinion groups could be modelled. This could be insightful in the further investigation of e.g. recent claims about a silencing of moderate voices in social media environments [10].

In order to compare groups of users in their activity patterns, one needs a method to distinguish between them. In Ch. 3, we developed a method concerned with the inference of latent social positions of individuals through interactions networks, e.g. subscribing to other users on social platforms or sharing their content. We identified that two different strands of research – latent space inference and force-directed layout algorithms – can be brought together: One can derive a novel force-directed layout algorithm from a latent space model which is then *interpretable* on the basis of said model. This is of increased relevance since the outcomes of existing force-directed layout algorithms, which are heavily used, lack clear interpretation.

To this end, one treats the negative log-likelihood of a network configuration – given the underlying model in which network interactions become more probable the closer two nodes are in a latent social space – as a potential energy. Its derivatives with respect to positions and model parameters can be seen as forces the move the nodes towards configurations that maximize the likelihood. We developed force terms for unweighted, weighted, and cumulative networks, and presented example layouts for each case. Comparison to existing layout algorithms (not grounded in an interpretable model) revealed that node groups are placed in similar configurations, but that the established algorithms show a stronger intra-cluster separation of nodes, as well as a tendency to separate clusters more strongly in retweet networks. The developed layout algorithm can be used for data exploration and illustration, but also, due to its grounding in a latent space model, as a rigorous basis for data analysis. Further work should include a more efficient implementation of the algorithm, so that also large-scale networks (larger than 10,000 edges) can be spatialised. Alternative force terms can be derived if alternative interaction models are proposed. For that task, this chapter can serve as a blueprint.

In Ch. 4, we focused on the observation of actual public debate online. We developed a novel method, employed in two case studies, which investigated public debate

on Twitter via network representations of retweets and replies. Through the interplay of the two networks, it was possible to identify group-level differences in willingness of opinion expression in comment sections below tweets from influential political figures. We argued that tweets observable there have both a direct and mediated effect on the perception of public opinion. While retweet networks were used to distinguish between opinion clusters on the basis of visual inspection of their force-directed spatialisation, the activity differences of users from these clusters were made visible for reply interactions. The measure of local assortativity was employed in order to investigate interaction patterns between opinion groups. The method was employed to observe public debate concerning two events: The Saxon state elections and violent riots in the city of Leipzig in 2019. We showed that in both cases, debate between clusters was vivid, but different groups exhibited different propensities to get involved in debate. Users retweeting far-right parties and politicians were significantly more active in both cases, hence their positions are disproportionately visible. Said users acted significantly more confrontational in the sense that they replied mostly to users from different groups, while the contrary was not the case. An especially striking feature of the local assortativity distributions was their bi-modal structure for both cases and all clusters. Further work should systematically investigate these differences. Moreover, as not all users that commented were also present in the retweet networks, alternative classification methods should be sought, possibly with methods of natural language processing (NLP).

The chapters of this thesis approached their common theme from different angles. Therefore, further work can be woven out of any of these strands individually, for which points of departure have been indicated in this chapter already – but it can also be developed from a combination of the methods and models presented. The model remains too simplified to allow connection to the empirical data at hand: The interaction patterns between users from different clusters are heterogeneous (see Figs. 4.4 and 4.5). However, one could qualitatively motivate that the increased visibility of a minority group found in Ch. 4 stems from the suitability of comment spaces to voice oppositional views. Users that fundamentally oppose 'mainstream' political parties might have a stronger incentive to write replies than others who, by and large, agree with the political status quo. This could be modelled by reduced expression costs for the fundamental opposition.

The collection and analysis of longitudinal digital trace data is a next step that suggests itself: Then one can ask not only who is visible at a specific point in time, but how visibility changes over time, which users change behaviour (or opinion), and what one should expect in the future. A need for more consistent data capturing dynamics of online interaction over time, instead of one-shot data, has been emphasized recently [75].

All in all, it is shown that group-level differences can (and do, as becomes clear in Ch. 4) have a strong influence on what becomes visible in public debate online. From the perspective of 'ordinary' users of online platforms, this work also allows an estimation of the opinion proportions on a platform that suggest themselves to them, and how this might be confounded by highly active user groups. Only taking into account what opinions are expressed online with no regard for activity differences between the groups that produce the content can be rather misleading if this is taken as representa-

tive of what users of a platform think about a certain issue. In that sense, naive reliance on what others express online can be collectively dangerous, especially in an era in which social media shapes social life and public discourse to an unprecedented extent.

Appendix A

Expected decrease of the difference in Q-values

We carry out the estimation for opinion group G_1, but the analogue holds for opinion group G_2. We can give an upper bound for the change in Q-value for the agent with the maximum Q-value of the group, $\dot{Q}^{\max}_{i\in G_1}$, and a lower bound for the change in Q-value for the agent with the minimum Q-value of the group, $\dot{Q}^{\min}_{i\in G_1}$ due to the monotonicity of the function $\frac{1}{1+e^{-x}}$:

$$\dot{Q}^{\max}_{i\in G_1} = \alpha'\Big(\frac{\gamma}{\gamma+1}\frac{1}{N_1-1}\sum_{\substack{j\in G_1\\ j\neq i}}\frac{1}{1+e^{-\beta Q_j}} - \frac{1}{\gamma+1}\frac{1}{N_2}\sum_{j\in G_2}\frac{1}{1+e^{-\beta Q_j}} - Q^{\max}_{i\in G_1} - c\Big) \leq$$
$$\alpha'\Big(\frac{\gamma}{\gamma+1}\frac{1}{N_1}\Big(\sum_{\substack{j\in G_1\\ j\neq i}}\frac{1}{1+e^{-\beta Q_j}} + \frac{1}{1+e^{-\beta Q^{\max}_{i\in G_1}}}\Big) - \frac{1}{\gamma+1}\frac{1}{N_2}\sum_{j\in G_2}\frac{1}{1+e^{-\beta Q_j}} -$$
$$Q^{\max}_{i\in G_1} - c\Big), \tag{A.1}$$

$$\dot{Q}^{\min}_{i\in G_1} = \alpha'\Big(\frac{\gamma}{\gamma+1}\frac{1}{N_1-1}\sum_{\substack{j\in G_1\\ j\neq i}}\frac{1}{1+e^{-\beta Q_j}} - \frac{1}{\gamma+1}\frac{1}{N_2}\sum_{j\in G_2}\frac{1}{1+e^{-\beta Q_j}} - Q^{\min}_{i\in G_1} - c\Big) \geq$$
$$\alpha'\Big(\frac{\gamma}{\gamma+1}\frac{1}{N_1}\Big(\sum_{\substack{j\in G_1\\ j\neq i}}\frac{1}{1+e^{-\beta Q_j}} + \frac{1}{1+e^{-\beta Q^{\min}_{i\in G_1}}}\Big) - \frac{1}{\gamma+1}\frac{1}{N_2}\sum_{j\in G_2}\frac{1}{1+e^{-\beta Q_j}} -$$
$$Q^{\min}_{i\in G_1} - c\Big). \tag{A.2}$$

If we now look at the change in time in the difference of $Q^{\max}_{i\in N_1}$ and $Q^{\min}_{i\in N_1}$, we can conclude by the above inequalities that the difference decreases at least exponentially in expectation by substracting the right hand-sides of (A.1) and (A.2).

$$\frac{d}{dt}(Q^{\max}_{i\in G_1} - Q^{\min}_{i\in G_1}) \leq -\alpha'(Q^{\max}_{i\in G_1} - Q^{\min}_{i\in G_1}). \tag{A.3}$$

The analogue holds for opinion group G_2:

$$\frac{d}{dt}(Q^{\max}_{i\in G_2} - Q^{\min}_{i\in G_2}) \leq -\alpha'(Q^{\max}_{i\in G_2} - Q^{\min}_{i\in G_2}). \tag{A.4}$$

Appendix B

Exploration rate bifurcation

The parameter β determines how sensitive agents are in their actions towards the current evaluation of their expected reward. A high β-value indicates a choice of the agent similar to a best response to their current evaluation of the expected reward, while $\beta = 0$ means that each available action is chosen with equal probability.

As is visible in Figure B.1, for very low β, there is only one fixed point available with a very low Q-value for both opinion groups. With β (≈ 5), further fixed points arise in a supercritical pitchfork bifurcation, and then, at $\beta > 6$, another (now subcritical) pitchfork bifurcation arises, such that we arrive at three stable fixed points (one in which both groups are in an expressive mode, and one for opinion dominance for each group) and two unstable ones in-between. Hence, if the action selections is close to a best response, we get more possible equilibria in the system. In the intermediate region, we have a situation in which only one of the two groups can be expressive, despite them both being internally well-connected.

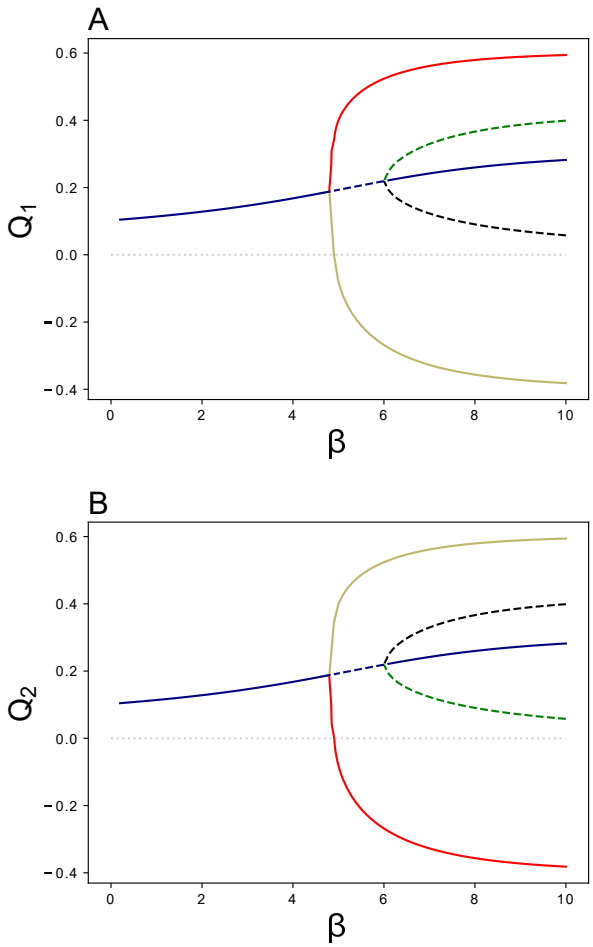


Figure B.1: The development of the fixed points with β given $c = 0.1$ and $\gamma = \delta = 2.36$. Since γ and δ are the same, the plots (A) and (B) are symmetric.

Appendix C

Force derivation: Cumulative networks

For cumulative networks, we stipulate the probability of establishing *a single tie upon action k from user i to user j* by

$$p(a_{ij}^k = 1) = \frac{1}{1+\exp(-\alpha_i - \beta_{jk} + d_{ij}^2)}, \tag{C.1}$$

The derivation of the forces for this case is analogous to the unweighted case, except that we sum over tweets instead of user pairs.

The log-likelihood for a given network can be written as

$$LL(G) = \sum_j \sum_{k=1}^{m_j} \Big(\sum_{(i,j)\in E_{jk}} (\alpha_i + \beta_{jk} - d_{ij}^2) - \sum_{\substack{i \\ i\neq j}} \log(1 + e^{-\alpha_i - \beta_{jk} + d_{ij}^2}) \Big) \tag{C.2}$$

The force on the position of node i' exerted by node j' is given by

$$\partial_{x_{i'}}^{j'} LL(G) = 2(\mathbf{x}_{i'} - \mathbf{x}_j) \Bigg[\sum_{k=1}^{m_{j'}} \Big(-a_{i'j'} + \frac{1}{1+\exp(-\alpha_{i'} - \beta_{j'k} + d_{i'j'}^2)} \Big) + \sum_{k=1}^{m_{i'}} \Big(-a_{j'i'} + \frac{1}{1+\exp(-\alpha_{j'} - \beta_{i'k} + d_{i'j'}^2)} \Big) \Bigg]. \tag{C.3}$$

The force on $\alpha_{i'}$ exerted by j' is given by

$$\partial_{\alpha_{i'}}^{j'} LL(G) = \sum_{k=1}^{m_{j'}} \Big(a_{i'j'} - \frac{1}{1+\exp(-\alpha_{i'} - \beta_{j'k} + d_{i'j'}^2)} \Big), \tag{C.4}$$

the force on each $\beta_{i'k}$ due to j' by

$$\partial_{\beta_{i'k}}^{j'} LL(G) = a_{j'i'} - \frac{1}{1+\exp(-\alpha_{j'} - \beta_{i'k} + d_{i'j'}^2)}. \tag{C.5}$$

Appendix D

Force derivation: Weighted networks

A possibility of extending the model to weighted graphs, i.e. an adjacency matrix $A = \{a_{ij} = k | k \in \mathbb{N}_0\}$, is the ordered logit or proportional odds model. There, the response variable has levels $0, 1, ..., n$ (e.g.: people rate their relationships to others on a scale from 0 to 6). The probability of the variable being greater than or equal to a certain level k is given by:

$$p(a_{ij} \geq k) = \frac{1}{1 + \exp(-c_k - \alpha_i - \beta_j + d_{ij}^2)} \tag{D.1}$$

($c_0 = \infty$, $c_{n+1} = -\infty$). The probability of a_{ij} equal to a certain k is given by

$$p(a_{ij} = k) = p(a_{ij} \geq k) - p(a_{ij} \geq k + 1). \tag{D.2}$$

The log-likelihood of the graph is then given by

$$LL(G) = \sum_{\substack{i,j \\ i \neq j}} \log\left(\frac{1}{1 + \exp(-c_{a_{ij}} - \alpha_i - \beta_j + d_{ij}^2)} - \frac{1}{1 + \exp(-c_{a_{ij}+1} - \alpha_i - \beta_j + d_{ij}^2)}\right) \tag{D.3}$$

As the simplest example, we turn to the case with three levels, where the probabilities are given by[1]

$$p(a_{ij} = 0) = 1 - \frac{1}{1 + \exp(-c_1 - \alpha_i - \beta_j + d_{ij}^2)} \tag{D.4}$$

$$p(a_{ij} = 1) = \frac{1}{1 + \exp(-c_1 - \alpha_i - \beta_j + d_{ij}^2)} - \frac{1}{1 + \exp(-c_2 - \alpha_i - \beta_j + d_{ij}^2)} \tag{D.5}$$

[1] In practice, one might often encounter data sets where individuals have rated a limited set of others, e.g. politicians, which themselves might not participate in the rating. This yields a bi-partite graph for which β only applies to the rat*ed* individuals, while α-values are present only for the rat*ing* individuals.

$$p(a_{ij}=2)=\frac{1}{1+\exp(-c_2-\alpha_i-\beta_j+d_{ij}^2)} \tag{D.6}$$

The derivative of Eq. (D.3) with respect to $\mathbf{x}_i$, α_i and β_i gives us the forces on the position and the parameters of node i. Additionally, the derivative with respect to c_1 and c_2 estimates the cut points. We introduce the following abbreviations:

$$C_1=-c_1-\alpha_{i'}-\beta_{j'}+d_{i'j'}^2,$$

and

$$C_2=-c_2-\alpha_{i'}-\beta_{j'}+d_{i'j'}^2.$$

The first part of the force on the position of node i' exerted by node j' is given by

$$\partial_{x_{i'}}\log(\mathrm{p}(a_{i'j'}=0))=2(\mathbf{x}_{i'}-\mathbf{x}_{j'})\frac{e^{C_1}}{(1+e^{C_1})^2-(1+e^{C_1})}=2(\mathbf{x}_{i'}-\mathbf{x}_{j'})\frac{1}{1+e^{C_1}} \tag{D.7}$$

if $a_{i'j'}=0$. If $a_{i'j'}=1$, the force is given by

$$\partial_{x_{i'}}\log(\mathrm{p}(a_{i'j'}=1))=2(\mathbf{x}_{i'}-\mathbf{x}_{j'})\frac{(1+e^{C_2})^{-2}e^{C_2}-(1+e^{C_1})^{-2}e^{C_1}}{(1+e^{C_1})^{-1}-(1+e^{C_2})^{-1}}. \tag{D.8}$$

Or, if $a_{i'j'}=2$,

$$\partial_{x_{i'}}\log(\mathrm{p}(a_{i'j'}=2))=-2(\mathbf{x}_{i'}-\mathbf{x}_{j'})\frac{1}{1+e^{-C_2}}. \tag{D.9}$$

For the second part of the force on i', caused by $a_{j'i'}$, one simply needs to take the appropriate term out of the three above and switch i' and j' for α and β.

The force on α caused by j' is given by

$$\partial_{\alpha_{i'}}\log(\mathrm{p}(a_{i'j'}=0))=-\frac{1}{1+e^{C_1}} \tag{D.10}$$

or

$$\partial_{\alpha_{i'}}\log(\mathrm{p}(a_{i'j'}=1))=\frac{(1+e^{C_1})^{-2}e^{C_1}-(1+e^{C_2})^{-2}e^{C_2}}{(1+e^{C_1})^{-1}-(1+e^{C_2})^{-1}} \tag{D.11}$$

or

$$\partial_{\alpha_{i'}}\log(\mathrm{p}(a_{i'j'}=2))=\frac{1}{1+e^{-C_2}}. \tag{D.12}$$

Now, there is no second force part – this is the only contribution by the pair i' and j'. The force on $\beta_{i'}$ is given by the analogous term where i' and j' are, again, switched for α and β, and the level of $a_{j'i'}$ is considered.

The forces on c_1 and c_2 by the pair are given by

$$\partial_{c_1}\log(\mathrm{p}(a_{i'j'}=0))=\partial_{\alpha_{i'}}\log(\mathrm{p}(a_{i'j'}=0)), \tag{D.13}$$

$$\partial_{c_1}\log(\mathrm{p}(a_{i'j'}=1))=\frac{e^{C_1}(1+e^{C_1})^{-2}}{(1+e^{C_1})^{-1}-(1+e^{C_2})^{-1}}, \tag{D.14}$$

$$\partial_{c_1}\log(\mathrm{p}(a_{i'j'}=2))=0. \tag{D.15}$$

(For the second part, switch i' and j' for α and β and consider $a_{j'i'}$.)

$$\partial_{c_2}\log(\mathrm{p}(a_{i'j'}=0))=0, \tag{D.16}$$

$$\partial_{c_2}\log(\mathrm{p}(a_{i'j'}=1))=-\frac{e^{C_2}(1+e^{C_2})^{-2}}{(1+e^{C_1})^{-1}-(1+e^{C_2})^{-1}}, \tag{D.17}$$

$$\partial_{c_2}\log(\mathrm{p}(a_{i'j'}=2))=\partial_{\alpha_{i'}}\log(\mathrm{p}(a_{i'j'}=2)). \tag{D.18}$$

(For the second part, switch i' and j' for α and β and consider $a_{j'i'}$.)

Appendix E

Bayesian correction of force term

For the inference of the model parameters we have only one sample. Although our model is a probabilistic model, our empirical distribution contains either $p(a_{ij} = 1) = 1$ if there is an edge, or $p(a_{ij} = 1) = 0$, if not. This can lead to divergences of the model parameters in the maximum likelihood solution. If one node is connected to all the other nodes, for instance, i.e. the graph contains a (N-1)-star subgraph, α_i will diverge in the maximum likelihood solution ($\alpha_i \to \infty$). One way to avoid this problem is to formulate is as a Bayesian inference problem. This is a well-defined problem, even with a single data point. Lets denote the entries of the adjacency matrix of our graph G as a_{ij} and the corresponding random variables A_{ij} or $\mathbf{A}$, respectively. Moreover, lets call the parameter vector of our model $\theta = \{\alpha, \beta, \mathbf{x}\}$, with $\theta_i = \{\alpha_i, \beta_i, \mathbf{x}_i\}$ and the corresponding random variable X. Then the posterior distribution of the parameters is given as

$$p(X|\mathbf{A}) = \frac{p(\mathbf{A}|X)p(X)}{p(\mathbf{A})} \tag{E.1}$$

Here $p(\mathbf{A}|X)$ is the likelihood (see Eq. (3.2)), $p(X)$ comprises all prior distributions and $p(\mathbf{A})$ is the marginal likelihood of the data. Instead of asking for the parameters that maximize the likelihood we can now ask for the parameters that maximize the posterior $p(X|\mathbf{A})$ or, equivalently, its logarithm. If one considers the gradients of the logarithm of the posterior again as forces, the likelihood term produces the same forces as in the maximum likelihood case. However, we may get additional forces from the prior term, which, for instance, can prevent the activity parameters from diverging.[1]

[1] [14] assumes normal priors for all parameters of the model. Nevertheless, it is noted in the main text that flat priors are used except for α and the positions $\mathbf{x}$. The mean of the prior distribution of α is set to 0 and the prior distribution for the positions is $\mathcal{N}(0,1)$.

Appendix F

Z-scores for Gaussian distribution of nodes

Table F.1: Average z-scores for a Mantel test for layout of a network generated from a Gaussian distribution of two groups of nodes with a sigma of 1/12 and a distance of 5/6 between the groups with varying node number (averaged over three runs).

Node number	100	300	600	900
Unweighted	36.65	77.49	90.56	94.91
Cumulative	43.07	77.04	84.58	92.12
Weighted	45.60	83.32	92.84	95.69

Appendix G

Real-world networks and comparison to other layout algorithms

G.1 German parliament

In the main text, the Bundestag follower network layout was only compared to ForceAtlas2. In Fig. G.1, we also show the follower network spatialised with Yifan Hu [69] (lower left) and Fruchterman Reingold [51] (lower right). All layout algorithms roughly reproduce party divisions. Fruchterman Reingold tends to distribute nodes homogeneously in spaces. *Die Linke* and the *Greens* have some overlap in this layout. A less pronounced overlap is also visible for Yifan Hu, which produces a comparably dense spatialisation.

The central placement of the *Greens* in the Leipzig Layout (as well as in the other layout algorithms) can be explained by their homogeneous usage of the platform: Each user of the party has an in/out degree of at least 50 (which is not the case for the other parties, see Fig. G.2). Moreover, the party members are followed and follow all other parties in a quite uniform fashion (see Fig. G.3).

Different (local) minima exist for this network, one of which is displayed in Fig. G.4. There, the *FDP* is placed between *SPD* and *Die Linke*. The inference of local minima is a general problem of FDLs – nevertheless, in the present framework, one can compare the log-likelihood of different equilibria and take the spatialisation with the highest likelihood (i.e. the lowest negative log-likelihood). The log-likelihood of Fig. G.4 is around 61,700, while it is roughly 57,700 for Fig. 3.5. The minimum with the higher likelihood is also the politically more plausible one: *SPD* and *Die Linke* are, especially when it comes to economic policy, oftentimes strongly opposed to the market-liberal *FDP*. On the other hand, *FDP* and *CDU/CSU* have often stressed that they are parties that have many things in common, such that Fig. 3.5 seems to be closer to political reality.

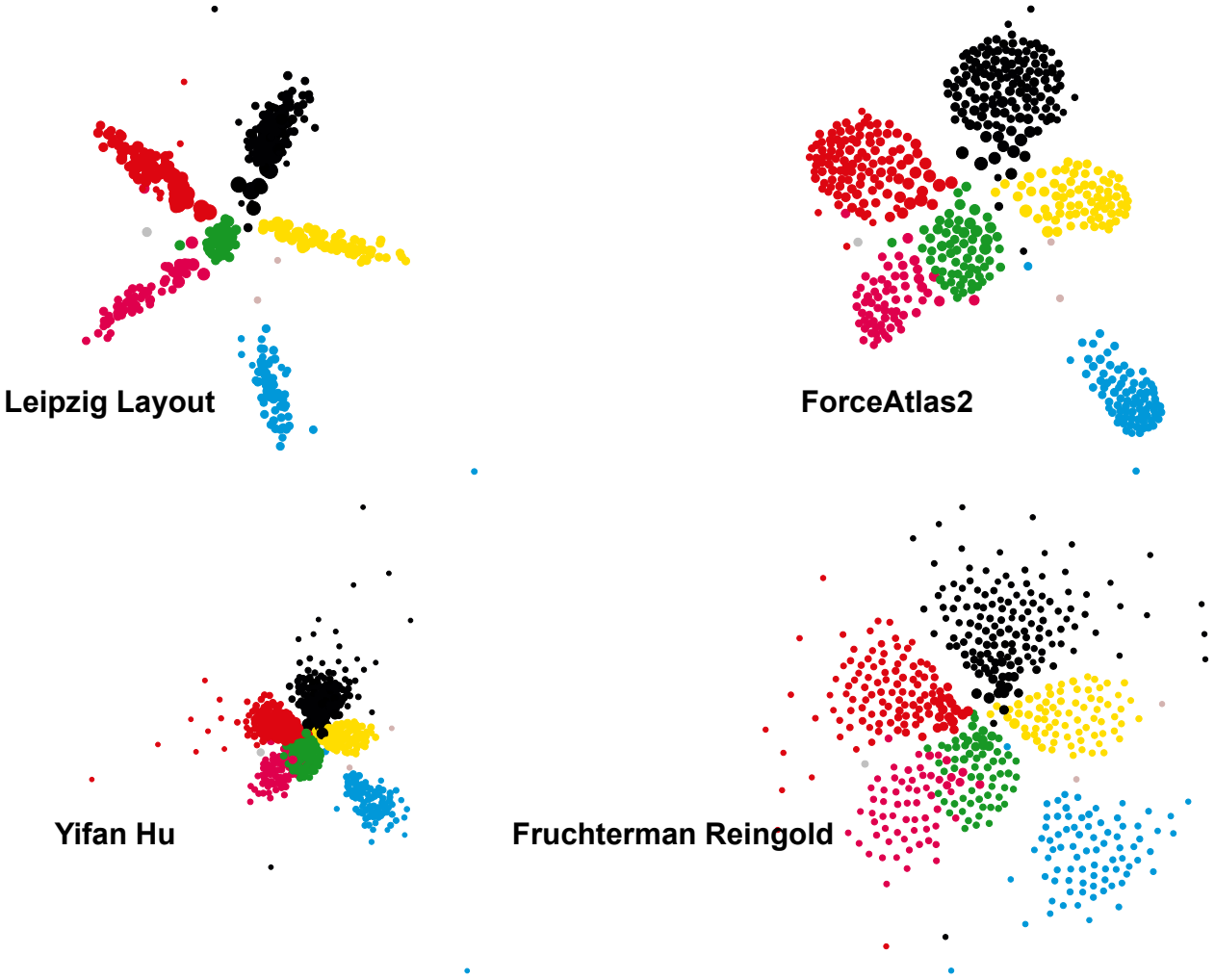


Figure G.1: Bundestag follower network, comparison of Leipzig Layout (top left), ForceAtlas2 (top right), Yifan Hu (lower left), and Fruchterman Reingold (lower right). Overall placement of parties is similar, but Leipzig layout, which allows closer placement of nodes, arranges all parties (except the *Greens*) along a one-dimensional axis.

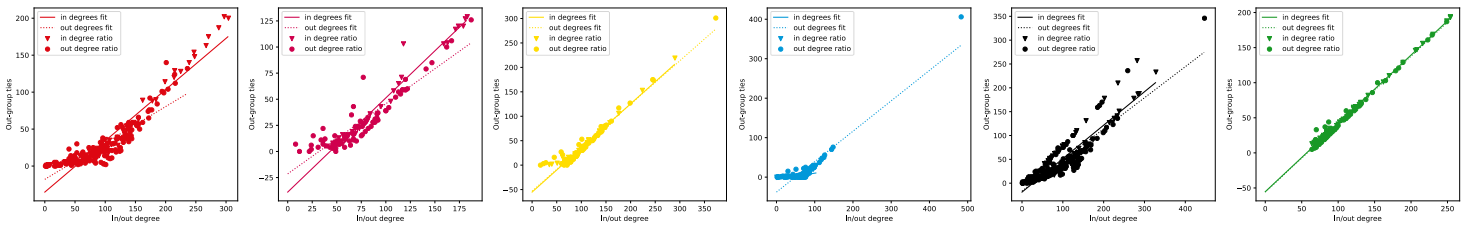


Figure G.2: Scatter plot of in/out degree and number of out-group ties for each party. No user from the *Green* party has in in or out degree of less than 50. They use Twitter quite homogeneously, which explains their central placement in the force-directed layout (as well as the homogeneously distributed incoming and outgoing links among parties, see next Fig.).

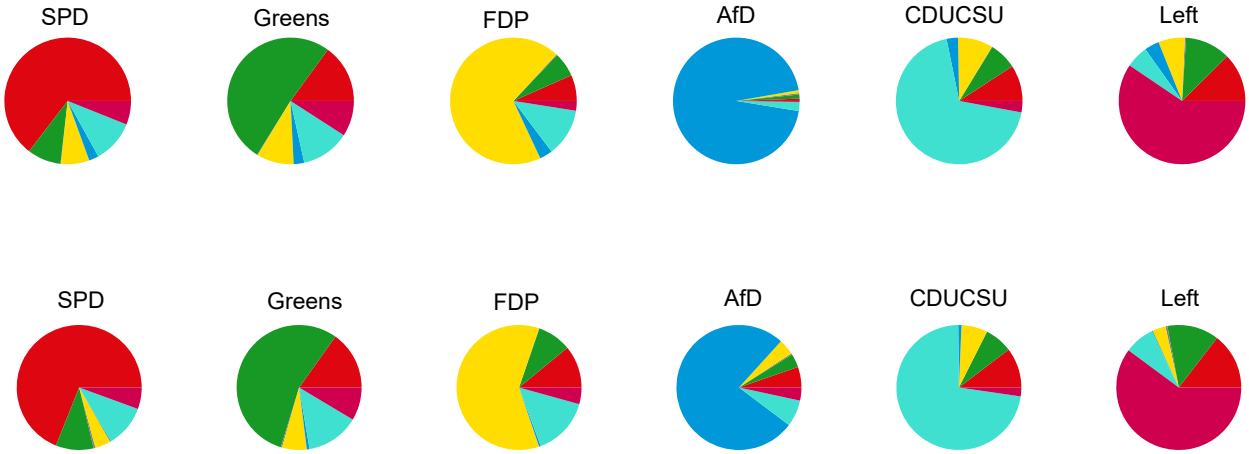


Figure G.3: Incoming and outgoing links from/to different parties by party. *Greens* are followed and follow all other parties (except *AfD*) quite uniformly, which explains their central placement.

Figure G.4: Local minimum of the follower network of German deputies. Parties are still visibly separated, but *FDP* is now placed between *SPD* and *Die Linke*.

G.2 Harper's letter

Fig. G.5 shows the Harper's letter retweet network spatialised with the four layout algorithms. Fruchterman Reingold, again, arranges the nodes rather uniformly in space. Interestingly, Yifan Hu and ForceAtlas2 produce a much more polarized spatialisation than Leipzig Layout.

G.3 Survey on energy-generating technologies: Correlations

Pairwise Pearson correlation coefficients between ratings of different energy-generating technologies (for the aggregated 3-point scale) can be found in Fig. G.6. Solar and onshore technologies exhibit the strongest correlation. Gas and coal, as well as offshore and onshore technologies are also correlated relatively strongly. Biomass has, among the renewables, the weakest correlations with the other renewables. On the other hand, ratings are not negatively correlated with coal or gas. This is mirrored by its placement in Fig. 3.7 at a certain distance from solar and onshore, and also offshore.

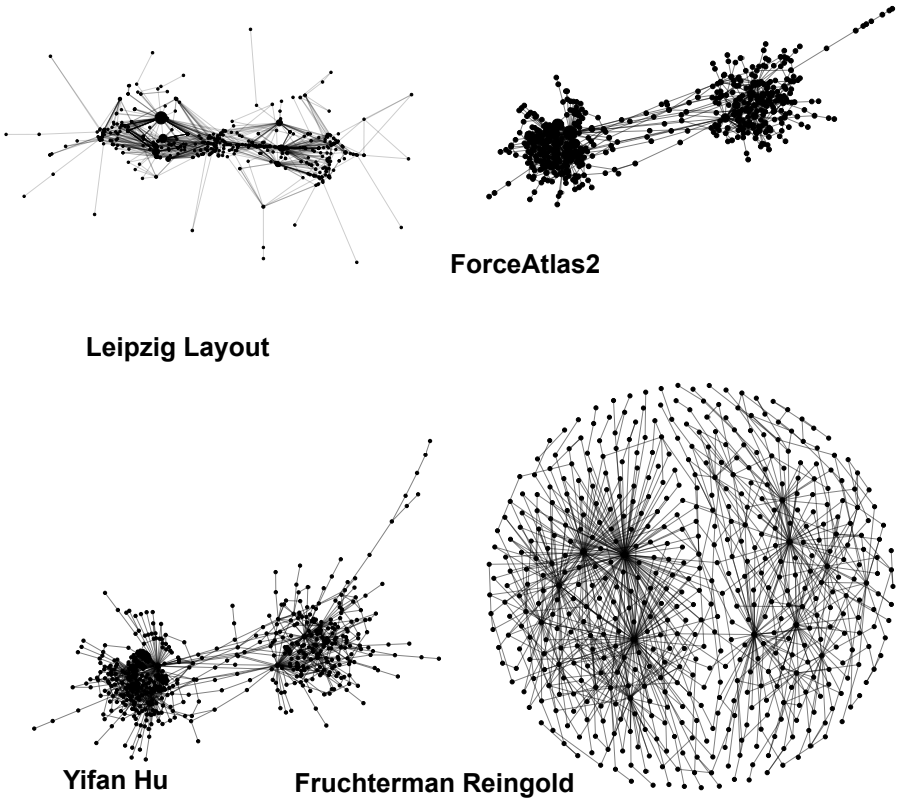


Figure G.5: Harper's letter retweet network comparison of Leipzig Layout (cumulative case, top left), ForceAtlas2 (top right), Yifan Hu (lower left), and Fruchterman Reingold (lower right).

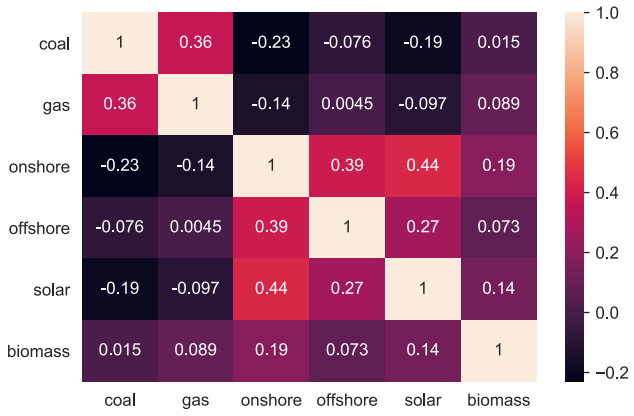


Figure G.6: Correlations between ratings of different energy-generating technologies for the aggregated 3-point scale.

Appendix H

Large retweet network

Figure H.1 shows the retweet network constructed out of all retweets in the data set between the 1st of July 2019 and then 24th of February 2020. If users were not present in the incident-specific retweet network, it was checked whether this large retweet network contained the users to increase the amount of classified users. Figure H.1 exhibits a very similar shape as the incident-specific retweet networks. Out of 88,167 users, 71.8% belonged to the majority cluster, 18.4% to the minority cluster and 9.8% to the intermediate region.

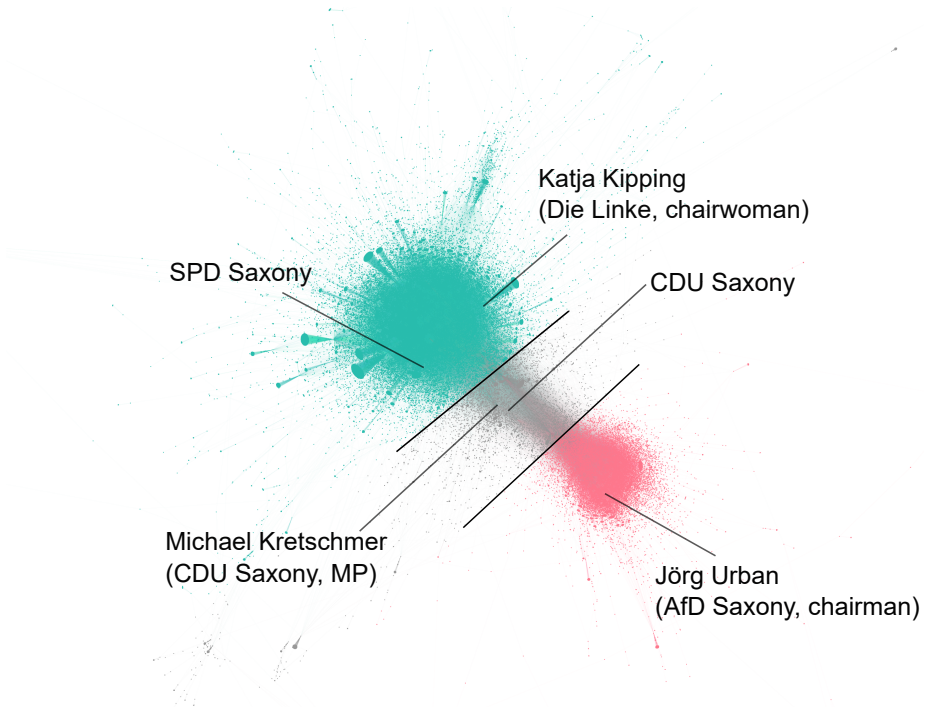


Figure H.1: Largest retweet network, constructed out of all retweets in the data set from July 2019 to the end of February 2020.

Appendix I

Local assortativity: Personalized PageRank

The stationary distribution of the Personalized PageRank can be calculated using the so-called Power Method, where the random walk process is directly simulated and the probability distribution over nodes i at step $s+1$ given by

$$w_\alpha(i;l)_{s+1} = \alpha \sum_j \frac{A_{ji}}{k_j^{\text{out}}} w_\alpha(j;l)_s + (1-\alpha)\delta_{i,l}, \tag{I.1}$$

with k_j^{out} being the out degree of node j and A_{ji} denotes a tie from j to i. This, at convergence, yields the distribution $w_\alpha(i;l)$ [120]. In order to solve this approximately, one can truncate the series of Eq. (I.1) at a certain step η. In [23], it is shown that the approximation of PageRank computed at step η of the Power Method coincides with the truncation of the power series of PageRank at step η; moreover, they show that one can calculate the distribution for a given $\alpha = \alpha_0$ and use the sequence of approximations to calculate the distribution for any other α. This yields

$$w_\alpha(i;l)_\eta = \delta_{i,l} + \sum_{s=1}^{\eta} \frac{\alpha^s}{\alpha_0^s} (w_{\alpha_0}(i;l)_s - w_{\alpha_0}(i;l)_{s-1}). \tag{I.2}$$

Integrating over α in the range of 0 to 1 then amounts to (the integral was carried out firstly for the TotalRank algorithm, see [22])

$$w_{\text{multi}}(i;l) = \int_0^1 w_\alpha(i;l)d\alpha = \delta_{i,l} + \sum_{s=1}^{\eta} \frac{w_{\alpha_0}(i;l)_s - w_{\alpha_0}(i;l)_{s-1}}{(s+1)\alpha_0^s}. \tag{I.3}$$

Appendix J

Acitivity share and possible data biases

In Table 4.2, we assess whether users from different clusters have different probabilities to get involved in debate. We use the incident-specific retweet network for the statistics there. One can also use the incident-specific retweet cluster combined with the large retweet cluster. The effect then is even more pronounced, as we show in Table J.1. We must note here, however, that the number of seed users in the majority pole is, for both data sets, larger than the number of seed users in the minority pole. While we do believe that the snowball-sampling method yielded a representative sample of important figures of Saxon politics and Saxon media outlets on Twitter, if this was not the case, a bias could have been introduced in the activity share of users if many retweets were of replies *and* minority pole users mostly commented in reply trees that were initiated by majority pole users. In both data sets, around 20% of the retweets were of replies. One can also construct a retweet network without retweets of replies to make sure this bias is eliminated. The effect still holds (Table J.2).

Table J.1: Percentage of users in the respective retweet network (not only incident-specific) that are active in the reply networks (seed users excluded) by cluster.

	Election	NYE
Majority cl.	10.3% (6,737)	3.1% (2010)
Minority cl.	28.1% (4,697)	10.2% (1727)
Intermediate cl.	14.5% (1,431)	4.3% (390)

Table J.2: Percentage of users in the respective retweet networks (without replies) that are active in the reply networks (seed users excluded) by cluster.

	Election	NYE
Majority cl.	21.5% (4,040)	16.5% (954)
Minority cl.	44.1% (2,173)	25.6% (386)
Intermediate cl.	24.7% (827)	22.3% (116)

Appendix K

Data losses, keywords, election results

Data losses Between the 26th and the 31st of August 2019, some data losses due to API problems occurred in the collection of tweets, see Figure K.1.

Keywords for the NYE data *connewitz, antifa, polizei, polizist, le0101, linx, leipzig, Not-OP, notop, linke, chaoten, angriff, le3112, randal* (applied to the root tweets and incident-specific retweet network.)

Election results *CDU* 32.5%, *AfD* 28.4%, *Die Linke* 12.3%, *Bündnis 90/Die Grünen* 8.9%, *SPD* 7.7%, *FDP* 4.7%, *Freie Wähler* 4.6%.

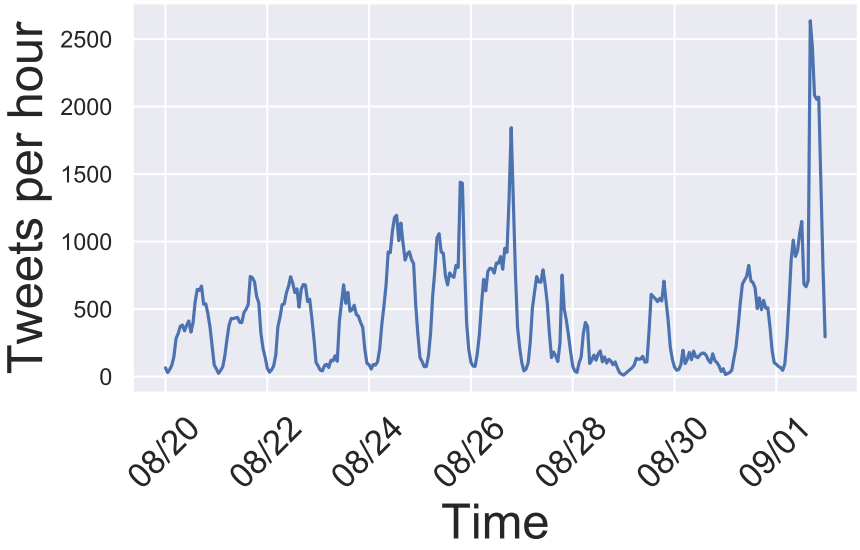


Figure K.1: Time series of the collected tweets. As is visible, some tweets were lost due to API problems between the 26th and the 31st of August.

Bibliography

[1] A letter on justice and open debate. https://harpers.org/a-letter-on-justice-and-open-debate/, 2020. (accessed 01.08.2021).

[2] Twitter announces second quarter 2021 results. https://s22.q4cdn.com/826641620/files/doc_financials/2021/q2/Q2'21-Earnings-Release.pdf, 2021. (accessed 01.08.2021).

[3] Twitter Terms of Service. https://twitter.com/en/tos, 2021. (accessed 01.08.2021).

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

[5] Wasim Ahmed, Peter A Bath, and Gianluca Demartini. Using Twitter as a data source: An overview of ethical, legal, and methodological challenges. *The ethics of online research*, 2017.

[6] Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07, 2008.

[7] Pablo Aragón, Karolin Eva Kappler, Andreas Kaltenbrunner, David Laniado, and Yana Volkovich. Communication dynamics in Twitter during political campaigns: The case of the 2011 spanish national election. *Policy & internet*, 5(2):183–206, 2013.

[8] Solomon E Asch. Opinions and social pressure. *Scientific American*, 193(5):31–35, 1955.

[9] Vasco Asturiano. force-graph. https://github.com/vasturiano/force-graph, 2018. (accessed 01.10.2021).

[10] Chris Bail. *Breaking the Social Media Prism*. Princeton University Press, 2021.

[11] Keith Michael Baker. *Public opinion as political invention*, page 167–200. Inventing the French Revolution: Essays on French Political Culture in the Eighteenth Century. Cambridge University Press, 1990.

[12] S. Banisch and E. Olbrich. Opinion polarization by learning from social feedback. *The Journal of Mathematical Sociology*, 43(2):76–103, October 2018.

[13] Sven Banisch, Felix Gaisbauer, and Eckehard Olbrich. How social feedback processing in the brain shapes collective opinion processes in the era of social media, 2020.

[14] Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91, 2015.

[15] Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26(10):1531–1542, aug 2015.

[16] L. Barnett, E. Di Paolo, and S. Bullock. Spatially embedded random networks. *Phys. Rev. E*, 76:056115, Nov 2007.

[17] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 3, 2009.

[18] Carlo Batini, L Furlani, and Enrico Nardelli. What is a good diagram? a pragmatic approach. In *Proceedings of the Fourth International Conference on Entity-Relationship Approach*, pages 312–319, 1985.

[19] Fabian Baumann, Philipp Lorenz-Spreen, Igor M. Sokolov, and Michele Starnini. Modeling echo chambers and polarization dynamics in social networks. *Phys. Rev. Lett.*, 124:048301, Jan 2020.

[20] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53:659–697, 2015.

[21] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[22] Paolo Boldi. Totalrank: Ranking without damping. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 898–899, 2005.

[23] Paolo Boldi, Massimo Santini, and Sebastiano Vigna. A deeper investigation of pagerank as a function of the damping factor. In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2007.

[24] Mike Bostock. d3-force. `https://github.com/d3/d3-force`, 2015. (accessed 01.10.2021).

[25] Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

[26] Ulrik Brandes. Drawing on physical analogies. In *Drawing graphs*, pages 71–86. Springer, 2001.

[27] Marcel Broersma and Todd Graham. Twitter as a news source: How Dutch and British newspapers used tweets in their news coverage, 2007–2011. *Journalism practice*, 7(4):446–464, 2013.

[28] Axel Bruns. Faster than the speed of print: Reconciling 'big data' social media analysis and academic scholarship. *First Monday*, 18(10):1–5, 2013.

[29] Axel Bruns and Jean Burgess. Crisis communication in natural disasters: The Queensland floods and Christchurch earthquakes. *Twitter and Society [Digital Formations, Volume 89]:*, pages 373–384, 2014.

[30] Axel Bruns and Jean Burgess. Twitter hashtags from ad hoc to calculated publics. *Hashtag publics: The power and politics of discursive networks*, pages 13–28, 2015.

[31] Pete Burnap, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams. 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. *Electoral Studies*, 41:230–233, 2016.

[32] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, May 2009.

[33] Damon Centola, Robb Willer, and Michael Macy. The emperor's dilemma: A computational model of self-enforcing norms. *American Journal of Sociology*, 110(4):1009–1040, 2005.

[34] Chun Cheng and Changbin Yu. Opinion dynamics with bounded confidence and group pressure. *Physica A: Statistical Mechanics and its Applications*, 532:121900, 2019.

[35] Hyunyoung Choi and Hal Varian. Predicting the present with google trends. *Economic record*, 88:2–9, 2012.

[36] M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the political alignment of Twitter users. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 192–199, Oct 2011.

[37] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on Twitter. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[38] Ross Cressman, Christopher Ansell, and Ken Binmore. *Evolutionary dynamics and extensive form games*, volume 5. MIT Press, 2003.

[39] Jesper Dall and Michael Christensen. Random geometric graphs. *Physical Review E*, 66(1):016121, 2002.

[40] Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528, 2003.

[41] Mathias Decuypere. Visual network analysis: a qualitative method for researching sociomaterial practice. *Qualitative research*, 20(1):73–90, 2020.

[42] Peter Duggins et al. A psychologically-motivated model of opinion change with applications to american politics. *Journal of Artificial Societies and Social Simulation*, 20(1):1–13, 2017.

[43] Peter Eades. A heuristic for graph drawing. *Congressus numerantium*, 42:149–160, 1984.

[44] Achim Edelmann, Tom Wolff, Danielle Montagne, and Christopher A Bail. Computational social science and sociology. *Annual Review of Sociology*, 46:61–81, 2020.

[45] Maria Fiedler and Frank Jansen. Was geschah an Silvester in Leipzig-Connewitz? https://www.tagesspiegel.de/politik/angriff-auf-polizisten-wirft-fragen-auf-was-geschah-an-silvester-in-leipzig-connewitz/25386832.html (last accessed: 20 June 2020), 2020.

[46] Andreas Flache, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, 20(4):2, 2017.

[47] Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *Proceedings of the national academy of sciences*, 104(1):36–41, 2007.

[48] Nancy Fraser. Rethinking the public sphere: A contribution to the critique of actually existing democracy. *Social text*, (25/26):56–80, 1990.

[49] Daniel Friedman. On economic applications of evolutionary game theory. *Journal of evolutionary economics*, 8(1):15–43, 1998.

[50] Thomas N Friemel and Mareike Dötsch. Online reader comments as indicator for perceived public opinion, 2015.

[51] Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

[52] Felix Gaisbauer. saxony-twitter. `https://github.com/fgais/saxony-twitter`, 2020. (accessed 01.08.2021).

[53] Michael T Gastner, Beáta Oborny, and Máté Gulyás. Consensus time in a voter model with concealed and publicly expressed opinions. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(6):063401, 2018.

[54] Michael T Gastner, Károly Takács, Máté Gulyás, Zsuzsanna Szvetelszky, and Beáta Oborny. The impact of hypocrisy on opinion formation: A dynamic model. *PloS one*, 14(6):e0218729, 2019.

[55] Noé Gaumont, Maziyar Panahi, and David Chavalarias. Reconstruction of the socio-semantic dynamics of political activist Twitter networks—Method and application to the 2017 French presidential election. *PLOS ONE*, 13(9):1–38, 09 2018.

[56] P. Gawronski, M. Nawojczyk, and K. Kulakowski. Opinion formation in an open system and the spiral of silence. *Acta Physica Polonica A*, 127, 2015.

[57] Daniel Gayo-Avello. No, you cannot predict elections with Twitter. *IEEE Internet Computing*, 16(6):91–94, 2012.

[58] Sherice Gearhart and Seok Kang. Social media in television news: The effects of Twitter and Facebook comments on journalism. *Electronic News*, 8(4):243–259, 2014.

[59] Benjamin H Good, Yves-Alexandre De Montjoye, and Aaron Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106, 2010.

[60] Mark Granovetter and Roland Soong. Threshold models of diversity: Chinese restaurants, residential segregation, and the spiral of silence. *Sociological Methodology*, 18:69–104, 1988.

[61] Jürgen Habermas. *Strukturwandel der Öffentlichkeit*. 1990.

[62] Jürgen Habermas. *Between facts and norms: Contributions to a discourse theory of law and democracy*. John Wiley & Sons, 2015.

[63] Mark S. Handcock, Adrian E. Raftery, and Jeremy M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, mar 2007.

[64] Frank E Harrell Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[65] Stephen Harrington. Tweeting about the telly: Live TV, audiences, and social media. *Twitter and society [Digital Formations, Volume 89]*, pages 237–247, 2014.

[66] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, dec 2002.

[67] Jake M Hofman, Duncan J Watts, Susan Athey, Filiz Garip, Thomas L Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J Salganik, Simine Vazire, et al. Integrating explanation and prediction in computational social science. *Nature*, pages 1–8, 2021.

[68] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[69] Yifan Hu. Efficient, high-quality force-directed graph drawing. *Mathematica Journal*, 10(1):37–71, 2005.

[70] Chung-Yuan Huang and Tzai-Hung Wen. A novel private attitude and public opinion dynamics model for simulating pluralistic ignorance and minority influence. *Journal of Artificial Societies and Social Simulation*, 17(3):8, 2014.

[71] Kosuke Imai, James Lo, Jonathan Olmsted, et al. Fast estimation of ideal points with massive data. *American Political Science Review*, 110(4):631–656, 2016.

[72] Mathieu Jacomy. *Situating Visual Network Analysis*. PhD thesis, 2021.

[73] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one*, 9(6), 2014.

[74] Andreas Jungherr. Analyzing political communication with digital trace data. *Cham, Switzerland: Springer*, 2015.

[75] Andreas Jungherr. Normalizing digital trace data. In *Digital discussions*, pages 9–35. Routledge, 2018.

[76] Jonas Kaiser. Public spheres of skepticism: Climate skeptics' online comments in the German networked public sphere. *International Journal of Communication*, 11, 2017.

[77] Antonis Kalogeropoulos, Samuel Negredo, Ike Picone, and Rasmus Kleis Nielsen. Who shares and comments on news?: A cross-national comparative analysis of online and social media participation. *Social Media and Society*, 3(4), 2017.

[78] Antonis Kalogeropoulos, Samuel Negredo, Ike Picone, and Rasmus Kleis Nielsen. Who shares and comments on news?: A cross-national comparative analysis of online and social media participation. *Social media+ society*, 3(4), 2017.

[79] Fariba Karimi, Mathieu Génois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. Homophily influences ranking of minorities in social networks. *Scientific reports*, 8(1):1–12, 2018.

[80] Ardeshir Kianercy and Aram Galstyan. Dynamics of Boltzmann Q learning in two-player two-action games. *Physical Review E*, 85(4):041145, 2012.

[81] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.

[82] Michael A. Krassa. Social groups, selective perception, and behavioral contagion in public opinion. *Social Networks*, 10(2):109–136, June 1988.

[83] Timur Kuran. Sparks and prairie fires: A theory of unanticipated political revolution. *Public choice*, 61(1):41–74, 1989.

[84] Timur Kuran. *Private truths, public lies*. Harvard University Press, 1997.

[85] David MJ Lazer, Alex Pentland, Duncan J Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, et al. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062, 2020.

[86] Eun Lee, Petter Holme, and Sang Hoon Lee. Modeling the dynamics of dissent. *Physica A: Statistical Mechanics and its Applications*, 486:262–272, 2017.

[87] Eun-Ju Lee. That's not the way it is: How user-generated comments on the news affect perceived media bias. *Journal of Computer-Mediated Communication*, 18(1):32–45, 2012.

[88] Eun-Ju Lee and Yoon Jae Jang. What do others' reactions to news on internet portal sites tell us? effects of presentation format and readers' need for cognition on reality perception. *Communication research*, 37(6):825–846, 2010.

[89] Simon Lindgren. *Data Theory: Interpretive Sociology and Computational Methods*. John Wiley & Sons, 2020.

[90] Jan Lorenz. Continuous opinion dynamics under bounded confidence: A survey. *International Journal of Modern Physics C*, 18(12):1819–1838, December 2007.

[91] REJ Lucas. Econometric policy evaluation: a critique, w: The phillips curve and labor mar kets,(red.: K. brunner and ah meltzer), 1976.

[92] Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220, 1967.

[93] Catherine Matias and Stéphane Robin. Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proceedings and Surveys*, 47:55–74, 2014.

[94] Joerg Matthes, Johannes Knoll, and Christian von Sikorski. The "spiral of silence" revisited: A meta-analysis on the relationship between perceptions of opinion support and political opinion expression. *Communication Research*, 45(1):3–33, 2018.

[95] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[96] Dan McGinn, David Birch, David Akroyd, Miguel Molina-Solana, Yike Guo, and William J Knottenbelt. Visualizing dynamic bitcoin transaction patterns. *Big data*, 4(2):109–119, 2016.

[97] Shannon C McGregor. Social media as public opinion: How journalists use social media to represent public opinion. *Journalism*, 20(8):1070–1086, 2019.

[98] Shannon C McGregor and Logan Molyneux. Twitter's influence on news judgment: An experiment among journalists. *Journalism*, 21(5):597–613, 2020.

[99] Jonathan Mellon and Christopher Prosser. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3):2053168017720008, 2017.

[100] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. Understanding the demographics of Twitter users. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[101] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from Twitter's streaming API with Twitter's firehose. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[102] Brian Mullen, John F Dovidio, Craig Johnson, and Carolyn Copper. In-group-out-group differences in social projection. *Journal of Experimental Social Psychology*, 28(5):422–440, 1992.

[103] Eni Mustafaraj, Samantha Finn, Carolyn Whitlock, and Panagiotis T Metaxas. Vocal minority versus silent majority: Discovering the opionions of the long tail. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 103–110. IEEE, 2011.

[104] Philip R Neary. *Multiple-group games*. PhD thesis, UC San Diego, 2011.

[105] Philip R. Neary. Competing conventions. *Games and Economic Behavior*, 76(1):301 – 328, 2012.

[106] German Neubaum and Nicole C Krämer. Monitoring the opinion of the crowd: Psychological mechanisms underlying public opinion perceptions on social media. *Media Psychology*, 20(3):502–531, 2017.

[107] Mark EJ Newman. Mixing patterns in networks. *Physical Review E*, 67(2), 2003.

[108] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[109] Andreas Noack. Unified quality measures for clusterings, layouts, and orderings of graphs, and their application as software design criteria. 2007.

[110] Andreas Noack. Modularity clustering is force-directed layout. *Physical Review E*, 79(2):026102, 2009.

[111] Elisabeth Noelle-Neumann. The spiral of silence a theory of public opinion. *Journal of Communication*, 24(2):43–51, June 1974.

[112] Elisabeth Noelle-Neumann. *Die Schweigespirale. Öffentliche Meinung–Unsere soziale Haut.* Riper [ie Piper], 1980.

[113] Elisabeth Noelle-Neumann and Thomas Petersen. The spiral of silence and the social nature of man. In *Handbook of political communication research*, pages 357–374. Routledge, 2004.

[114] Elisabeth Noelle-Neumann and Thomas Petersen. The spiral of silence and the social nature of man. In *Handbook of political communication research*, pages 357–374. Routledge, 2004.

[115] Elizabeth Noelle-Neumann. Public opinion. our social skin, 1984.

[116] Ann Nowé, Peter Vrancx, and Yann-Michaël De Hauwere. Game theory and multi-agent reinforcement learning. In *Reinforcement Learning*, pages 441–470. Springer, 2012.

[117] Christian Nuernbergk and Julia Conrad. Conversations and campaign dynamics in a hybrid media environment: Use of Twitter by members of the German Bundestag. *Social Media+ Society*, 2(1), 2016.

[118] Steve Paulussen and Raymond A Harder. Social media references in newspapers: Facebook, Twitter and YouTube as sources in newspaper journalism. *Journalism practice*, 8(5):542–551, 2014.

[119] Leto Peel. Multiscalemixing. `github.com/piratepeel/multiscalemixing`, 2018. (accessed 01.08.2021).

[120] Leto Peel, Jean-Charles Delvenne, and Renaud Lambiotte. Multiscale mixing patterns in networks. *Proceedings of the National Academy of Sciences*, 115(16):4057–4062, 2018.

[121] Mathew Penrose. *Random geometric graphs*. Number 5. Oxford University Press, 2003.

[122] Jürgen Pfeffer, Katja Mayer, and Fred Morstatter. Tampering with Twitter's sample API. *EPJ Data Science*, 7(1):50, 2018.

[123] Pablo Porten-Cheé and Christiane Eilders. Spiral of silence online: How online communication affects opinion climate perception and opinion expression regarding the climate change debate. *Studies in communication sciences*, 15(1):143–150, 2015.

[124] Armin Pournaki, Felix Gaisbauer, Sven Banisch, and Eckehard Olbrich. The Twitter Explorer: A framework for observing twitter through interactive networks. *Journal of Digital Social Research*, 3(1):106–118, 2021.

[125] Vincent Price. *Public opinion*. 1992.

[126] Tom Quilter. Noise matters in heterogeneous populations. *ESE Discussion Papers*, 169, 2007.

[127] Veronica Red, Eric D Kelsic, Peter J Mucha, and Mason A Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53(3):526–543, 2011.

[128] Christine Reissing. Sachsens CDU-Kandidaten schließen Koalition mit AfD aus. https://www.mdr.de/nachrichten/politik/regional/cdu-schliesst-koalition-afd-aus-sachsen-100.html (last accessed: 20 June 2020), 2019.

[129] Björn Ross, Laura Pilz, Benjamin Cabrera, Florian Brachten, German Neubaum, and Stefan Stieglitz. Are social bots a real threat? an agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, 28(4):394–412, 2019.

[130] Lee Ross, David Greene, and Pamela House. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3):279–301, 1977.

[131] Wolfgang Schweiger. *Der (des)informierte Bürger im Netz*. Springer, 2017.

[132] Wolfgang Schweiger and Miriam Weihermüller. Öffentliche meinung als onlinediskurs–ein neuer empirischer zugang. *Publizistik*, 53(4):535–559, 2008.

[133] Indira Sen, Fabian Floeck, Katrin Weller, Bernd Weiss, and Claudia Wagner. A total error framework for digital traces of humans. *arXiv preprint arXiv:1907.08228*, 2019.

[134] Hawal Shamon, Diana Schumann, Wolfgang Fischer, Stefan Vögele, Heidi U Heinrichs, and Wilhelm Kuckshinrichs. Changing attitudes and conflicting arguments: Reviewing stakeholder communication on electricity technologies in germany. *Energy research & social science*, 55:106–121, 2019.

[135] Yilun Shang. Resilient consensus for expressed and private opinions. *IEEE Transactions on Cybernetics*, 2019.

[136] Dongyoung Sohn. Spiral of silence in the social media era: A simulation approach to the interplay between social networks and mass media. *Communication Research*, June 2019.

[137] Dongyoung Sohn and Nick Geidner. Collective Dynamics of the Spiral of Silence: The Role of Ego-Network Size. *International Journal of Public Opinion Research*, 28(1):25–45, 03 2015.

[138] Daniel Sousa, Luís Sarmento, and Eduarda Mendes Rodrigues. Characterization of the Twitter replies network: are user ties social or topical? In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 63–70, 2010.

[139] Nina Springer and Anna Sophie Kümpel. User-generated (dis) content. In *Journalismus im Internet*, pages 241–271. Springer, 2018.

[140] Barrett Steinberg and Marc Ostermeier. Environmental changes bridge evolutionary valleys. *Science advances*, 2(1):e1500921, 2016.

[141] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[142] Richard S Sutton, Andrew G Barto, and Ronald J Williams. Reinforcement learning is direct adaptive optimal control. *IEEE Control Systems Magazine*, 12(2):19–22, 1992.

[143] William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clusters. *The Journal of Chemical Physics*, 76(1):637–649, January 1982.

[144] Daiki Takeuchi, Gouhei Tanaka, Ryo Fujie, and Hideyuki Suzuki. Public opinion formation with the spiral of silence on complex social networks. *Nonlinear Theory and Its Applications, IEICE*, 6(1):15–25, 2015.

[145] Roberto Tamassia, Giuseppe Di Battista, and Carlo Batini. Automatic graph drawing and readability of diagrams. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1):61–79, 1988.

[146] Florian Toepfl and Eunike Piwoni. Public spheres in interaction: Comment sections of news websites as counterpublic spaces. *Journal of Communication*, 65(3):465–488, 2015.

[147] Petter Törnberg and Anton Törnberg. The limits of computation: A philosophical critique of contemporary Big Data research. *Big Data & Society*, 5(2):2053951718811843, 2018.

[148] Amanda L Traud, Peter J Mucha, and Mason A Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

[149] Damian Trilling. Two different debates? investigating the relationship between a political debate on tv and simultaneous comments on twitter. *Social science computer review*, 33(3):259–276, 2015.

[150] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*, 2010.

[151] Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. A selection-mutation model for q-learning in multi-agent systems. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 693–700, 2003.

[152] José Van Dijck. 'You have one identity': Performing the self on Facebook and LinkedIn. *Media, Culture & Society*, 35(2):199–215, 2013.

[153] Livia van Vliet, Petter Törnberg, and Justus Uitermark. The Twitter parliamentarian database: Analyzing Twitter politics across 26 countries. *PLOS ONE*, 15(9):e0237073, sep 2020.

[154] Tommaso Venturini, Mathieu Jacomy, and Pablo Jensen. What do we see when we look at networks: Visual network analysis, relational ambiguity, and force-directed layouts. *Big Data & Society*, 8(1):20539517211018488, 2021.

[155] Loup Verlet. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*, 159(1):98–103, July 1967.

[156] Annie Waldherr and Marko Bachl. Simulation gesellschaftlicher Medienwirkungsprozesse am Beispiel der Schweigespirale. In *Rezeption und Wirkung in zeitlicher Perspektive*, pages 235–252. Nomos Verlagsgesellschaft mbH & Co. KG, 2011.

[157] Stefanie Walter, Michael Brüggemann, and Sven Engesser. Echo chambers of denial: Explaining user comments on climate change. *Environmental Communication*, 12(2):204–217, 2018.

[158] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[159] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[160] Duncan J Watts. Should social science be more solution-oriented? *Nature Human Behaviour*, 1(1):1–5, 2017.

[161] B.M. Waxman. Routing of multipoint connections. *IEEE Journal on Selected Areas in Communications*, 6(9):1617–1622, 1988.

[162] Tom Willaert, Sven Banisch, Paul Van Eecke, and Katrien Beuls. Facilitating online opinion dynamics by mining expressions of causation. The case of climate change debates on The Guardian. *arXiv:1912.01252*, 2019.

[163] Sarita Yardi and danah boyd. Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of science, technology & society*, 30(5):316–327, 2010.

[164] Mengbin Ye, Yuzhen Qin, Alain Govaert, Brian DO Anderson, and Ming Cao. An influence network model to study discrepancies in expressed and private opinions. *Automatica*, 107:371–381, 2019.

Bibliographische Daten

Voice and silence in public debate: Modelling and observing public opinion expression online
((Un-)Sichtbarkeit in öffentlichen Debatten online: Modellierung und Beobachtung der Bereitschaft zur Meinungsäußerung im Netz)
Gaisbauer, Felix
Universität Leipzig, Dissertation, 2021
125 Seiten, 30 Abbildungen, 164 Referenzen

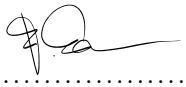
Daten zum Autor

Name:	Felix Gaisbauer
Geburtsdatum:	28.05.1993 in Passau
04/2012 - 07/2018	Studium der Physik (BSc, MSc) Universität Regensburg und Ludwig-Maximilians-Universität München
10/2013 - 03/2017	Studium der Philosophie (BA) Ludwig-Maximilians-Universität München
seit 11/2018	Doktorand am Max-Planck-Institut für Mathematik in den Naturwissenschaften

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 22.09.2022



.....
(Felix Gaisbauer)