DISS. ETH Nr. 23979

Opinion Polarization in Online Communities

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
ADIYA ABISHEVA

M. Sc. ETH Zurich in Computer Science
B. Eng. & ACGI Imperial College London in Computing

born on August 8, 1987

citizen of Kazakhstan

accepted on the recommendation of
Prof. Dr. Dr. Frank Schweitzer
Prof. Dr. Ricardo Baeza-Yates

2016

ETH zürich

Acknowledgements

I want to express my profound gratitude to my advisor, Prof. Dr. Dr. Frank Schweitzer, for supporting me in identifying my best professional skills and strengths and finding the right motivational words that only encouraged me to improve further in my academic performance. For me he has truly been both an academic supervisor and a supportive mentor.

I want to gratefully thank my second advisor, Prof. Dr. Ricardo Baeza-Yates. His detailed and structured comments helped to improve my thesis significantly. I feel honoured that he agreed to become my thesis co-examiner.

I want to deeply thank Dr. David Garcia for bringing me in to the scientific world. He believed in my abilities when I was a Computer Science student assistant five years ago. Throughout the studies he gave me iteratively an excellent feedback and valuable suggestions. It was my great honour and pleasure to work with him.

I want to thank my colleagues with whom I had scientific and not only scientific talks, and with some of whom we became also very good friends. Thanks to my colleagues, Antonios, Emre, Ingo, Nicolas and Rebekka, for your senior advises. Thanks to PhD colleagues, Giaccomo, Giona, Simon, Vahan and Yan. Thanks to former colleagues, Mario, Pavlin, Rahel and Corneel. It was always a pleasure to have either game evenings with you, or tea breaks with biscuits, or watch funny videos on YouTube during the lunch break.

I would like to thank my collaborators in the political polarization project, Dr. Uwe Serdült and Dr. Tomislav Milic.

Final gratitude goes to our secretary, Nadja, for being great not only in administrative issues but also in making exceptional cakes.

Contents

Abstract

This thesis investigates the instances of collective opinion polarization in various domains in online participatory media and the role of emotional interactions in the polarization of opinions. We adopt an interdisciplinary approach. First, we leverage on theoretical knowledge and empirical evidence from social and political science and psychology to formulate the research questions. Second, we collect the digital traces of manifestation of opinions and emotions, such as comments and posts, or likes and dislikes on social networking sites, blogs or video-sharing websites, and analyze them by means of novel computational methods like the automated classification of emotions expressed in text, machine learning and statistical modeling, and analysis of large-scale data.

Conceptually, we divide our work into three parts. In the first part, we reveal statistical regularities and patterns of opinion polarization in political and non-political domain at the collective level. In political domain, our contribution is two-fold. From methodological point of view, we assess the level of polarization in the multi-party system, moving away from the tradition of applying US two-party polarization metrics. While this metric is the commonly accepted measure polarization, the two-party political system is not the most dominant political configuration in the most democratic systems. From conceptual point of view, we address political polarization not from the ideological stances of individuals, but from the social interactions of political actors by employing the network science approach. In non-political domain, we explore in a quantitative way one of the aspects of the filter bubble phenomenon, namely the backlash of negativity after the threshold of collective attention is reached. We show statistically the presence of the dual regimes of human appraisals of online items, namely local and global popularity of online content, such that global popularity almost always entails extreme polarization of opinions. We also reveal that the expression of strong arousal and negative emotions in online content drives more polarized responses, which is in line with the theoretical findings in psychology.

In the second part, we elaborate the study of emotional reactions at the level of individual users. By statistically analyzing 65 millions response messages from three online communities, we test the predictions of one of the most influential psychological theories of emotions – the negativity bias. We build the complete picture of reaction tendencies to emotional expressions, and find that what makes us react online is more salient than

when we react, which highlights the differences between emotional interaction in offline face-to-face interaction and in online computer mediated communication.

In the third part, motivated by the duality of the relationship between positive and negative collective evaluations, we develop a data-driven statistical model of human appraisals governed by the law of proportionate effect and the decay in collective attention. The model reproduces well the statistical parameters of the empirical distribution of collective evaluations as well as certain characteristics of the burst of the filter bubble. However, the complete trajectories of the dynamics of collective evaluations cannot be recovered with the developed model. We propose that this limitation can be later overcome by developing an agent-based model where the emotional dynamics of agents is incorporated.

Finally, based on a five-year continuous computer crawl, we obtained and analyzed large-scale datasets of digital traces of opinions and emotions, which allowed us to perform all our studies in an unobtrusive, observational setting. This way, instead of conducting surveys and experiments, we move away from the traditional "forced exposure to content" approach of learning human behaviour to a "natural exposure" approach of studying polarization of opinions and emotional reactions in vivo.

Kurzfassung auf Deutsch

Die vorliegende Arbeit untersucht Beispiele von kollektiver Meinungspolarisation auf partizipativen Online-Medien über verschiedenen Bereiche hinweg, und ergründet die Rolle emotionaler Interaktionen in der Meinungspolarisation. Unser Ansatz ist interdisziplinärer Natur. Zuerst verwenden wir theoretisches Wissen und empirische Evidenz aus den Sozial- und Politikwissenschaften sowie der Psychologie, um unsere Forschungsfragen zu formulieren. Danach sammeln wir digitale Spuren von Meinungsbildung und Emotionen, wie zum Beispiel Kommentare und Posts oder Neigungs- und Abneigungsäusserungen auf sozialen Netzwerkseiten, Blogs oder online Video-Platformen. Wir analysieren diese Daten mithilfe neuer rechnergestützter Methoden, wie zum Beispiel automatisierter Klassifikation von Emotionen, die in Texten zum Ausdruck gebracht werden, maschinellem Lernen und statistischer Modellierung sowie Analyse gross angelegter Datensätze.

Konzeptionell gliedert sich unsere Arbeit in drei Teile. Im ersten Teil zeigen wir statistische Regularitäten und Meinungsbildungsmuster in politischen sowie nichtpolitischen Gebieten auf kollektiver Ebene auf. Im politischen Bereich besteht unser Beitrag aus zwei Hauptpunkten. Aus methodologischer Perspektive erfassen wir das Polarisationsniveau im Vielparteiensystem, wobei wir uns von der Tradition abheben, US-typische Zwei-Parteien-Polarisationsmetriken zu verwenden. Auch wenn solche Zwei-Parteien-Polarisationsmetriken allgemein akzeptiert sind, so reflektieren diese nicht die Tatsache, dass die meisten demokratischen Systeme aus mehr als zwei Parteien bestehen. Konzeptuell behandeln wir politische Polarisation nicht ausgehend von ideologischen Standpunkten von Individuen, sondern auf Grundlage der sozialen Interaktionen politischer Akteure, indem wir einen Netzwerkansatz verwenden. Im nicht-politischen Bereich explorieren wir auf quantitative Art und Weise einen Aspekt des sogenannten Filterblasenphänomens, nämlich den Negativitätsbacklash, nachdem die kollektive Aufmerksamkeitsschwelle erreicht ist. Wir weisen statistisch die Existenz zweier Regime der menschlichen Bewertung von Onlinegegenständen auf, nämlich lokale und globale Popularität von Onlineinhalten, wobei mit globaler Popularität zumeist extreme Meinungspolarisation einhergeht. Ausserdem zeigen wir, dass der Ausdruck starken emotionalen Arousals und negativer Valenz in Onlineinhalten zu stärker polarisierte Antworten führt. Dieses Ergebnis ist mit theoretischen Erkenntnissen in der Psychologie im Einklang.

Im zweiten Teil elaborieren wir die Studie emotionaler Reaktionen auf der Ebene individueller User. Indem wir 65 Millionen Antwortmitteilungen dreier Online Communities statistisch analysieren, testen wir die Prognosen einer der einflussreichsten psychologischen Theorien der Emotionen - den Negativity Bias. Wir entwerfen ein vollständiges Bild von Reaktionstendenzen zu emotionalen Ausdrücken und finden heraus, dass Emotionen online staerker beeinflussen ob wir reagieren, als wann wir reagieren. Dies hebt

Unterschiede in der emotionalen Interaktionen zwischen direkter Offline- und indirekter Online-Kommunikation hervor.

Die Dualität der Beziehung zwischen positiven und negativen kollektiven Evaluationen motiviert den dritten Teil dieser Arbeit. In diesem entwickeln wir ein datengestütztes statistisches Modell menschlicher Bewertungen, das von Gibrats Gesetz und dem Abfall kollektiver Aufmerksamkeit geprägt wird. Das Modell reproduziert die statistischen Parameter der empirischen Verteilung kollektiver Evaluationen und gewisse Charakteristiken des Platzens der Filterblase. Dennoch können die vollständigen Trajektorien der Dynamik kollektiver Evaluationen nicht mit dem dargestellten Modell beschrieben werden. Wir legen dar, wie solche Modelllimitationen später überwunden werden können, indem emotionale Dynamiken der Akteure in das Modell eingebaut werden.

Abschliessend haben wir durch fünf Jahre langes kontinuierliches Crawlen einen gross angelegten Datensatz digitaler Spuren von Meinungen und Emotionen gewonnen. Dies hat uns ermöglicht, all unsere Studien mit einem nicht-reaktiven Boebachtungs-Design durchzuführen. Anstatt also Umfragen oder Experimente durchzuführen, d.h. Vesuchspersonen künstlichen Stimuli auszusetzen, haben wir einen Ansatz der natürlichen Exposition gewählt, um Meinungs- und Emotions-Polarisierung in vivo zu untersuchen.

Summary

For Chapters 2 to 5, a detailed summary is presented on the first page.

Chapter 1: Introduction

We start with the rationale behind studying collective opinion polarization and provide an elucidating example to vividly demonstrate the topic. We briefly review existing empirical and theoretical literature in social science and psychology on opinion formation, opinion polarization and emotions, and describe the methodology of sentiment detection and analysis of online texts. We also list the research questions and hypotheses which are addressed by this dissertation.

Chapter 2: Political Polarization in Online Communities

We study the extent and evolution of political polarization based on social interactions of politicians in `politnetz.ch`, a Swiss online platform focused on political activity.

Chapter 3: Emotional Reactions in Online Communities

We test the outcomes of the negativity bias theory of emotions on 65 millions pairs of human responses in online communities. We reveal the patterns of emotional expressions in messages that trigger replies in users.

Chapter 4: Emotional Polarization in Online Communities

We quantify the traces of the burst of the filter bubble on the Internet. We also study statistical regularities in the collective evaluations online, and measure the extend to which emotions play role in opinion polarization.

Chapter 5: Multiplicative Growth Model of Collective Evaluations

We develop a statistical model of coupled evaluations that explains the growth in evaluations by activation of positive and negative evaluation and the decay in collective attention. We also reproduce the statistical parameters of the empirical distributions of evaluations.

Chapter 6: Conclusions

We outline contributions of each chapter in brief details. We present our past studies which provide an additional support to results in the main chapters. We give the direction of the future works on studying polarization. And we list the scientific contribution of this work and its relevance in other fields.

Appendix

We provide supplementary figures and tables for certain chapters.

List of Publications

The present dissertation is based on the following publications:

- Adiya Abisheva, David Garcia, and Frank Schweitzer. (2016) "When the Filter Bubble Bursts: Collective Evaluation Dynamics in Online Communities". In: *Proceedings of the 8th ACM Conference on Web Science*. ACM, pp. 307–308.
- David Garcia, Adiya Abisheva, Simon Schweighofer, Uwe Serdült, and Frank Schweitzer. (2015). "Ideological and temporal components of network polarization in online political participatory media". *Policy & Internet* 7(1), pp. 46–79.
- Adiya Abisheva, Kiran Venkata Rama Garimella, David Garcia, and Ingmar Weber. (2014) "Who watches (and shares) what on Youtube? and when?: using Twitter to understand Youtube viewership". In: *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, pp. 593–602.

and the following unpublished working papers:

- Adiya Abisheva, David Garcia, and Frank Schweitzer. (2016) "Response patterns to emotional expression in online interaction".
- Adiya Abisheva, David Garcia, and Frank Schweitzer. (2016) "Modelling the filter bubble".

Chapter 1

Introduction

Technological advancements of the recent two decades affected our every day life at different speed. We have witnessed services, like `Uber` or `AirBnB`, that disrupted fast and overt the existing markets and changed classical economy. However, we have just realised the significance of the services, like online social media, that over the last decade have been increasingly penetrating into our daily life. Rewind a decade, `Facebook` was just another social network for students, `LinkedIn` was essentially a digital resume and `Twitter` was not even understood at first. Fast forward to 2016, `Facebook` now connects not only former classmates, but it is also a messaging and photo sharing service for families and friends, `Twitter` is our daily news aggregator with personalised newsfeeds, and `LinkedIn` is a job market. Online social media is officially embedded into our personal life and businesses.

Through gradual changes, like one time authentication, one account for all, simpler and friendlier user design and many more facilitations, staying connected has become seemless and eventually essential. Nowadays, through digital bits we share bits of us – identity, looks, connections, emotions and opinions. What started off as only consumption of information has eventually turned into active interaction and participation in online media.

In the offline world, we are able to say but not to be heard by everyone. In the online world, however, we are able to say and potentially be heard by millions of different kind, of different hierarchy, social status, geographical location, religion or culture. The opportunity for every individual to have their say on the Internet is what turns online participatory media into active discussion platforms. This development leads to an increasing transformation of our societies into "digital democracies" in which online participatory media play a central role in the provision with information on political and societal issues as well as in the political discourse. The increase of citizens' engagement in the public discourse is undeniably a positive aspect. However, there is an increasing number of evidences that - under certain circumstances - the exchange of opinions between individuals can lead to an increasing polarization of opinions. The latter process is characterized by a growing distance between users in their opinion space to the extent that it becomes harder for users

to reconcile. Not only these developments affect social media, but they also influence the culture of online discourse.

The importance of collective opinion formation and dynamics, and mechanisms underlying it have recently become the subject of intense interdisplinary research, *e.g.* computational social science. This thesis aims at studying different instances of opinion polarization, such as mobilisation of politicians before and during election seasons or polarized reactions of users to controversial online posts, as well as the factors that are behind the increasing polarization, such as elicitation of mixed and negative emotional reactions as a response to such controversial stimuli.

As an illustrative example of how opinion polarization starts and evolves in online participatory media and what factors influence it's development, we commence with an anectodal example from the Internet – the inane debate that polarized the online community.

The Great Dress Debate On February 25th, 2015, a `Tumblr` user posted a photo of a dress (Figure 1.1) asking the online community to help her figure out its colours, and mentioning that one half of her friends perceived the dress colour on the photo as blue and black, and the other half saw the colour as white and gold. Within 48 hours the post went suddenly viral and gained over 400,000 notes[1] and about 840,000 views per minute, way above the normal traffic rates for other content on the site.[2] The colour dispute went worldwide and crossed various platform borders. The article about the dress gained attention in several news websites, including Wired, CNBC, Business Insider and Time, as well as among celebrities. Users expressed their opinion on *"#TheDress"* in online polls, tweets, `Facebook` comments and other social media channels. `On Buzzfeed`, the original post collected more than 38 millions views surpassing all known traffic records of the site, and resulted in 1.8 millions voters making their voice heard on the dress, with 72% of votes given for white and gold and 28% for blue and black. On `Twitter`, users divided into opinion camps by creating hashtags for opposing sides *"#WhiteAndGold"* or *"#BlackAndBlue"*, and within 24 hours, the dress image received over 1.2 millions tweets with users fiercely expressing their stance for one team or another. The dress image gained absurd popularity, went viral within a day and clashed together a lot of people, such that some articles labelled the dispute as "the drama that divided the planet" that potentially can harm interpersonal relationships (McCoy,

Figure 1.1: The image of the dress that sparked fierce online debates. Note: the colours of the image on a computer display may appear differently than on a hard printed version.

[1] `http://knowyourmeme.com/memes/thedress-what-color-is-this-dress`
[2] `https://en.wikipedia.org/wiki/The_dress`

2015). Some users even compared the "Great Dress Debate of 2015" with political elections: "This is so crazy! It is like we are in the middle of an election season judging by how divided people are over this issue. LOL" (Ford, 2015). Ironically, the dress controversy on the Internet, indeed, obtained chacteristic attributes of any online election campaign, like the armies of supporters on both sides, mobilising users to gather under hashtags *"#teamwhiteandgold"* or *"#teamblackandblue"*, or presence of doubting users gathering under *"#teamIDK"* ("I don't know") as well as users obviously not interested in "elections" like *"#teamIDGAF"* ("I don't give a f."), and finally emergence of new "parties" like *"#teamLOL"* (Wilson, 2015). While the optical mechanism behind the colour perception is an interesting question by itself, not less of an interesting question is how such a trivial matter can create such a big argument (Conway, 2015). Forgetting about the science of "collective retina" illusion (Ford, 2015), the truth *is* that even a trivial online item *can polarize* the Internet community.

However, one might ask oneself whether the seemingly trivial might not be so trivial in the end. One overlooked factor is the direct connection of colour to our emotional states. Across different people, certain colours are usually associated with certain *emotional states*, for instance numerous findings argue that red carries negative valence, *e.g.* the warning signal, while green usually yields positivity, *i.e.* the signal of security (Elliot, 2015; Gil and Le Bigot, 2015; Kuhbandner and Pekrun, 2013). The complex interplay between subjective colour perception and emotional reactions evoked by visual stimuli like colour is what makes the collective obsession with this viral hit (Conway, 2015).

This apparently anectodal story provides a literal illustrative example of how *polarization based on emotions* can arise. Additionally, we learn several facts of human behaviour:

- We observe collective behaviour and in particular the traces of opinion polarization in online participatory media. Individuals form opinions not only on matters of global scale and of surely big importance, like political elections, but even on such seemingly trivial local issues.
- Emotional reactions are evoked from various stimuli. In addition to visual stimuli, like the ones in the dress conundrum, communication and participation platform such as the Internet offers plethora of textual stimuli – words, phrases or sentences – which can also evoke emotional reactions in human beings.
- Interactions based on emotional responses can lead to unintended outcomes, for instance polarizing the online community. Thus, strong opinion polarization may appear as a result of emotional interactions.

The mentioned characteristics of online human behaviour constitute the core of this thesis and each will be examined in detail in the related chapters. In the following sections of the Introduction chapter we go through the empirical and theoretical background and the state-of-the-art research in each of the areas.

1.1 Opinion Polarization as Collective Social Phenomenon

One of the research lines in sociology is to understand *collective behaviour*, a social phenomenon in which large group of individuals are engaged, which emerges in a "spontaneous" way but lives short, and where societal norms are either absent or unclear, or contradict each other (Blumer, 1951; Wikipedia, 2015). Examples of collective behaviour are numerous. They include phenomena like social panic (human crush at Hillsborough football stadium or the Kentucky Beverly Hills Supper Club fire), sudden interest in fashion garment (the case of "WristStrong" silicone bracelets), peaceful social movements like Greenpeace or violent social riots, or the rapid spread of rumours and scandals (Wikipedia, 2015). Additionally, due to the power of online communication new examples of collective behaviour appear, such as the mobilization of crowds during political protests, *i.e.* Twitter revolution (Wikipedia, 2016),[3] or sudden "bursts" of interest in the website or online posts that later become online viral hits, *e.g.* "#TheDress" debate.

In the social science literature, polarization is seen as both a state and a process (DiMaggio *et al.*, 1996). One of the contributions of this thesis is to distinguish between the state of a system which is *polarized* and a *polarization* process. According to Lang and Lang (1962), polarization is one the five collective processes that produces collective behaviour. Collective process is described as a "collective response growing out of unstructured interaction", and polarization is further specified as "a solidary response of each subgroup at the expense of over-all consensus". The response of the social system may appear as a result of either exogeneous (external) or endogeneous (internal) *system shock*. Opinion polarization process, therefore, is a response of social system to some event, like viral hit online or the upcoming elections campaign, which ignites interactions among individuals, which eventually leads to a more polarized state of opinions than it was initially. Furthermore, it can be characterized by two distinctive features: a) *inter-communication* (between subgroups) diminishes, b) *intra-communication* (within subgroup) becomes confined and strengthens.

Models of opinion formation based on social influence The central concept in opinion polarization is the presence of some opinion groupings in social system or the formation of the opinions and their further divergence from each other. Lang and Lang (1962) clarifies that the process of polarization may develop from the existing antagonisms, *e.g.* from existing ideological differences or social divisions, which increasingly strengthen the opposing viewpoints into partisan commitments.

[3]The term refers to different revolutions and protests, most of which had the social networking site Twitter be used by protestors and demonstrators in order to communicate. The set of events included Arab spring, Euromaidan and Moldova civil unrest (Wikipedia, 2016).

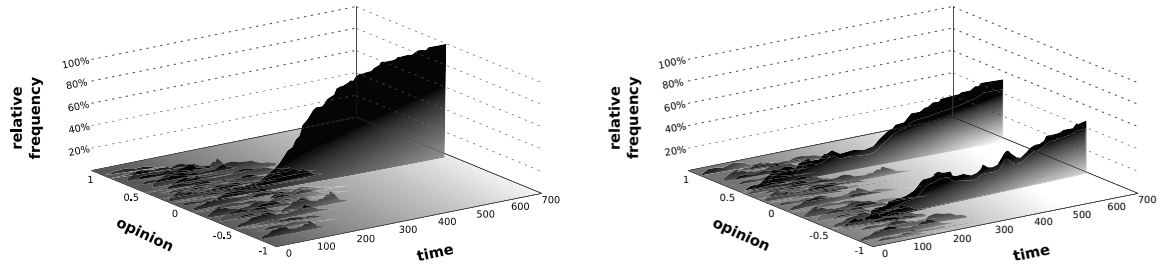


Figure 1.2: Left: Positive homophily and social influence result in opinion consensus. Agents update their opinion to the average opinion of their contacts. Right: Positive homophily and social influence with interaction noise result in opinion polarization. Figures are adapted from Mäs and Flache (2013).

Choice homophily, or the tendency of individuals to interact more with similar others, has been empirically found responsible for the social group formation (Easley and Kleinberg, 2010). However, there are relatively few formal models (Axelrod, 1997a,b; Macy *et al.*, 2003) that demonstrate under which conditions homophily alone keeps the social diversity. Centola *et al.* (2007) and others postulate that homophily alone is not enough for the social groupings to be preserved. This is due to the principle of *social influence* (Abelson, 1964; DeGroot, 1974), which asserts that the more interaction among individuals leads to more similarity between them. When projected to opinion formation, this means that interaction between individuals under social influence will lead to individuals updating their opinions to opinions of their contacts, and this development eventually results in opinion consensus, see Figure 1.2.

Mathematical approaches towards the modeling of opinion formation based on the principle of social influence have been developed. Additionally, they have focused on exploring the conditions under which the opinion diversity is induced. For instance, *bounded confidence model* (Deffuant *et al.*, 2000; Hegselmann *et al.*, 2002; Lorenz, 2007) introduces an influence interval that agents have on other agents, such that outside of this interval agents do not influence opinions of each other. Depending on the weight of the influence interval, polarization is emergent. The model however suffers from the two major weaknesses: a) opinion clustering under this model is not stable, *i.e.* if a certain amount of interaction noise is introduced, agents converge their opinions, and b) it cannot generate opinion diversity from initial consensus of opinions. Another agent-based model, known as the *model of negative influence* (Flache and Mäs, 2008), is based on four core mechanisms – homophily, social influence, heterophobia (disliking of dissimilar others) and negative influence (distancing or rejection, or the tendency to become more dissimilar to the disliked ones). Even though opinion polarization can be reached under this model, the empirical confirmation for the negative influence mechanism is mixed (Mäs, 2012). The latest introduced agent-based model known as *persuasion model* (Mäs *et al.*, 2013) is based on the fact that a) opinions are influenced through the exchange of arguments and b) the information exchange between individuals of the similar opinion tends to intensify pre-existing

attitudes *before* the discussion. This development results in a social phenomenon known as *group polarization* and have been empirically confirmed by Myers and Lamm (1976).

So far, empirical studies of opinion polarization (Moscovici and Zavalloni, 1969; Myers and Lamm, 1976; Van Swol, 2009) as well as mathematical models discussed earlier have focused on *social* mechanisms underlying the formation of opinions. While these *external* factors, like social interaction, mass media or social influence, have a defining role in opinion formation, there are also *internal* factors, like *subjective* experience and emotions, that have a direct access to shaping the opinion of an individual. In the next section we present theoretical and empirical studies that have revealed the influence of emotions on opinions and decisions – the question that still remains an active research area.

The role of emotions in decision-making and opinion formation Opinions are cognitive states which are a product of subjective experience and emotions. Zaller (1992) relates opinion to "a marriage of information and predisposition", where information serves to form a mental representation of the given issue, and predisposition helps to motivate some conclusion about it. Emotions are subjective states that exhibit a much faster relaxation towards neutral states (Kuppens *et al.*, 2010), and can be mapped to a predisposition in opinion formation, and therefore can be viewed as subjective cues that help individuals to decide in favor or against the issue.

For a long time, the role of affect (or emotion) in the existing theories of choice, decision-making and information processing have been underestimated (Shafir *et al.*, 1993). However, Zajonc (1980) argued that affective reactions to stimuli are often the very first reactions that occur automaticaly and subsequently guide information processing and judgement. He gives the following example: "Quite often 'I decided in favor of X' is no more that 'I liked X'". Later, Isen *et al.* (1978) proposed the theory of affect priming which states that emotional associations help reducing the effort of tasks' evaluation. Finally, Finucane *et al.* (2000) have proposed a theory of affective heuristics which suggests that emotional states reduce the cognitive complexity. In order to not complete an extensive search for information when making a decision and shorten the decision-making process, individuals resort to mental shortcuts, one of which is the affect heuristic. It is described as "a swift, involuntary response to a stimulus that speeds up the time it takes to process information", and serves as a first and fast rather than informed reponse mechanism in the decision making.

In the empirical studies on opinion formation, emotional dimensions – like *arousal*, or *valence* (Mano, 1999) – and their influence on an individual's opinion have been investigated (Kühne, 2012). Furthermore, based on the theory of *core affect* (Russell and Barrett, 1999), a number of empirical studies (Gonzalez-Bailon *et al.*, 2010; Gorn *et al.*, 2001; Kühne *et al.*, 2012) in computational social psychology have shown that emotions foster opinion polarization. In Section 1.2 we will explain in greater details the mentioned theory of core affect, emotional dimensions, and the existing theories of emotional interactions.

1.2 On Emotions and Emotional Interactions

1.2.1 Emotions Theory

The dimensional model of emotion in psychology distinguishes the two orthogonal axes of affect, also known as *core affect*: valence and arousal (Russell, 1979, 1980; Russell and Barrett, 1999). *Valence* characterizes a feeling of pleasure or displeasure, and summarizes at the level of subjective experience how well an individual is doing (Russell and Barrett, 1999). *Arousal* encompasses a feeling of activation or deactivation, and decsribes at the level of subjective experience a sense of mobilization of energy (Russell and Barrett, 1999). There are also evidences of other dimensions, such as *dominance* (Fontaine *et al.*, 2007; Russell and Mehrabian, 1977) or *potency* (Osgood, 1969), but their consistent inclusion in psychological research about opinions is still to be explored.

Evaluative space Experimental findings show that there are differences in the structure of affective experience – most of the individuals are more valence focused than arousal focused (Feldman, 1995), which indicates that valence is at the center of emotional experience (Chmiel *et al.*, 2011a). While valence was originally conceptualized as a single dimension, Cacioppo *et al.* (1997) have argued for a bivariate evaluative space, with separable positive and negative dimensions, see Figure 1.3 (Cacioppo and Berntson, 1994; Cacioppo and Gardner, 1999; Cacioppo *et al.*, 1997). An individual's evaluation of an object or an event can therefore have both positive and negative components at the same time. Furthermore, these components are represented in anatomically different subsystems of the brain (Norman *et al.*, 2011). As a consequence, this give rise to a possibility of a simultaneous co-activation of both positive and negative affect, which leads to a feeling of *ambivalence* or that of mixed emotions (Berrios *et al.*, 2015). And, additionally, an important characteristic of bivariate evaluations is their asymmetric influence on behaviour, known as the *negativity bias*.

The Negativity Bias theory Cacioppo's model links evaluations to behaviour in the following way: while stronger evaluations generally lead to stronger reactions, the relative influence of negative and positive evaluations determines the *type of behavioral reaction*. This simple mapping of evaluations to behavior is distorted by three principles: positivity offset, negativity bias, and evaluative ambivalence (Cacioppo *et al.*, 1997; Miller, 1961; Norman *et al.*, 2011). In the absence of strong positive or negative stimulus information, an organism will tend to exhibit behavior characteristic of positive evaluations. Thus, the function mapping positive information to behavior has a positive offset. On the other hand, negative stimulus information will draw more cognitive resources, and lead to stronger and faster behavioral reactions than positive information of the same strength. In other words, the function mapping negative information to behavior has a steeper slope, see Figure 1.3 and Equation 1.1.

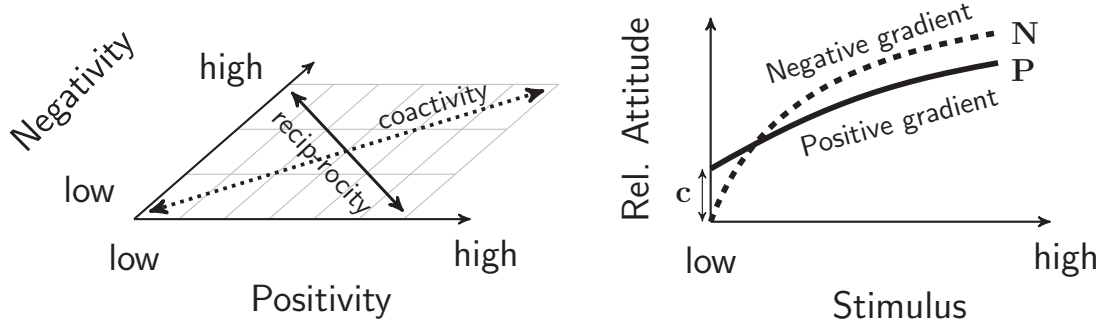


Figure 1.3: Left: The bivariate evaluative plane. The horizontal axis represents the level of positive evaluation and is labelled as *positivity*, the slanted axis – the level of negative evaluation, or *negativity*. The solid diagonal line maps to the bi-polar reciprocal evaluation, and the dotted diagonal line represents the ambivalence or co-activation of both emotions. Right: The negativity bias and the positivity offset. y-axis shows the relative strength of attitude, x-axis represents the absolute value of activation (or the function of evaluation). The strength of the negative attitude compared to the positive attitude has a steeper slope. However, at a low activation level the positive gradient has a higher intercept than the negative gradient. Figures are adapted from Cacioppo *et al.* (1997) and Norman *et al.* (2011).

The mapping of behavior to evaluation

$$\begin{aligned} \text{Strength of Behaviour} &= f(\text{Strength of Attitude}), \\ \text{Strength of Attitude} &= f(\text{Strength of Evaluation}), \\ \text{Strength of Evaluation} &= f(\text{Strength of Stimulus}), \text{ and} \\ \text{Strength of Attitude} &= \mathbf{P} \cdot f(i_+) + c - \mathbf{N} \cdot f(i_-), \end{aligned} \tag{1.1}$$

where i_+ and i_- are the units of positive and negative stimuli, and therefore $f(i_+)$ and $f(i_-)$ are the activation functions of positivity and negativity respectively, and $\mathbf{P}$ and $\mathbf{N}$ are the weighting coefficients of evaluations. In Equation 1.1, the negativity bias is instantiated when $\mathbf{N} > \mathbf{P}$, such that with each unit of respective stimulus, the change in negative attitude is larger than the change in positive attitude. And the positivity offset is ensured by $c \geq 0$, such that in the absence of stimulus when $f(i_+) = 0$ and $f(i_-) = 0$, there is a tendency for a weak positive attitude.

Both principles are likely to have an evolutionary origin: while the positivity offset allows an organism to explore its environment in the absence of negative stimuli, the negativity bias enhances an organism's reaction to threat (Norman *et al.*, 2011). Finally, evaluative ambivalence occurs when positive and negative behavioral tendencies balance each other out.

The role of arousal and negative valence in creating polarized behavioural reactions Cacioppo has proposed the mapping between valence and asymmetrical behavioural reactions. Empirical research has shown that emotional arousal can also result in behavioral reactions and in particular in polarized responses (Reisenzein, 1983). The *theory of misattribution* (Reisenzein, 1983; Zillmann, 1971) explains the emergence of extreme reactions in terms of the transfer of residual emotions between events, which subsequently intensifies the reaction to the second event. For instance, men in the state of high emotional arousal, for example from physical exercises, give more extreme ratings of attractiveness to women in comparison to the situation in which raters are in a calm emotional state. Further empirical evidence have found that the subjective experience of arousal motivates evaluation on the extremes (Paulhus and Lim, 1994). For example, the ratings of famous figures by students are found to be more polarized right before taking an exam, in comparison to weeks before or after. This extreme reactions are especially salient when arousal is experienced along with negative valence, for instance the stress before an exam. Thus, in light of these theoretical and empirical studies, we may expect that the activation of negativity and arousal can result in more polarized behavioural responses.

Emotional interactions So far, we have presented studies on the role of emotional evaluations on behavioral reactions of a single individual. Numerous previous works in social psychology have shown the influence of emotions on interactions between individuals and specifically on creation and maintenance of social relations (Christophe and Rimé, 1997; Heath *et al.*, 2001; Rimé, 2009).

One of the fundamental theories in emotional interactions, *the theory of social sharing* (Christophe and Rimé, 1997; Rimé, 2009), states that we not only experience emotions but we also share our emotional experiences. Therefore, emotions can not be only viewed as individual *intrapersonal* episodes, but they are also fundamentally social *interpersonal* experiences (Kappas, 2013) that motivate social interaction (Rimé, 2009) and shape our communication (García *et al.*, 2012a). It has already been shown experimentally that emotional stories are more attractive to the listeners than neutral ones, and that they elicit emotions in listeners (Zech *et al.*, 2004). However, while surveys and experiments can shed light on reactions to emotions in controlled scenarios and particular contexts (Rimé, 2009; Zech *et al.*, 2004), a broader understanding of human emotional behavior requires us to take into account *natural exposure* (McPhee, 1963), *i.e.* the elicitation of emotional reactions *in vivo*. In Chapter 3, we do this by conducting an unobtrusive, observational study Webb *et al.* (1966) of a large-scale dataset of digital traces of emotional interaction in 223 millions conversational samples from three social media platforms. By measuring how we react when others express their emotions through online texts, we address this open research question of the influence of communication on emotional experience and vice versa.

We have briefly mentioned that our studies on emotional interactions are based on *digital traces of emotional expressions*. In Section 1.2.2, we explain in great detail the research on online emotions and how emotions on the Internet are detected.

1.2.2 Emotions in Online Communities

Emotional expression through online text has been analyzed in earlier research on data from `MySpace` (Thelwall *et al.*, 2010b), `Yahoo` answers (Kucuktunc *et al.*, 2012), IRC channels (Garas *et al.*, 2012), `Gentoo` (García *et al.*, 2013b), `Wikipedia` (Iosub *et al.*, 2014), BBC (Chmiel *et al.*, 2011b), `Digg`, `YouTube` (García *et al.*, 2012b) and `Twitter` (Bollen *et al.*, 2011; Thelwall *et al.*, 2012). Availability of large-scale quantitative datasets allows us to understand emotions and their role in various domains, from health, to politics, to finance. For instance, the role of emotions in finance has been assessed in relation to the mood (Bollen *et al.*, 2011) and expressions of worry (Gilbert and Karahalios, 2010) in stock markets.

Studies that monitor mental health and depression leverage extensively on quantifying emotions through text. For instance, subjective well-being is manifested in `Facebook` status updates (Wang *et al.*, 2014), and it exhibits the pattern of assortativity in social networks (Quercia *et al.*, 2012) in relation to the feelings of loneliness (Burke *et al.*, 2010). Theories of periodic mood oscillations (Golder and Macy, 2011) have been validated with the help of `Twitter` data. Furthermore, psycholinguistic analysis of emotions on `Twitter` has displayed the traces of depression in some participants (De Choudhury and De, 2014; De Choudhury *et al.*, 2013). Finally, De Choudhury *et al.* (2012) have shown the relation between the mood of participants on `Twitter`, measured in terms of valence and arousal, and their online interaction and participation on the platform.

Geographic and demographic components of emotional experience have also been addressed by computational social scientists. Emotional reactions to urban aesthetics can be measured through crowdsourcing methods (Quercia *et al.*, 2014a), leading to technologies that automatically produce pleasant travel maps (Quercia *et al.*, 2014b). Emotional expressions have exhibited segregation patterns in geographical space (Lin, 2014) and between genders (Kivran-Swaine *et al.*, 2012; Thelwall *et al.*, 2010b). Furthermore, collective emotions are analyzed through digital traces (Schweitzer and García, 2010). These are are emotional states shared by large amounts of individuals due to their social interaction. The properties and traces of collective emotions have been demonstrated in real-time chats (Garas *et al.*, 2012) and product reviews (García *et al.*, 2011).

Lastly, an information-centric role (García *et al.*, 2012a) of online emotions has been studied through blogs (Miller *et al.*, 2011), on `Twitter` (Pfitzner *et al.*, 2012), and on `Yahoo` answers (Kucuktunc *et al.*, 2012). Emotional interactions lead to creating certain social network structures (Tan *et al.*, 2011; West *et al.*, 2014). Negative emotional posts

have been shown to be the drivers of communication among users and are responsible for the extension of the lifetime of online discussions in forums (Chmiel *et al.*, 2011b). Finally, online emotional reactions like valence, arousal and dominance, synchronize with political outcomes (Gonzalez-Bailon *et al.*, 2010), which is in line with the findings that political discussions are emotionally charged (Hoang *et al.*, 2013), in particular during election periods (García *et al.*, 2012b).

Sentiment detection To study the emotional expression on the Internet requires the usage of tools from sentiment analysis (Pang and Lee, 2008). While most of these tools are developed for opinion mining, part of the sentiment analysis focuses on the extraction of emotional content from text. Some supervised approaches can be trained on large datasets (De Choudhury *et al.*, 2012). The current state-of-the-art tools apply unsupervised lexicon-based analysis techniques (Taboada *et al.*, 2011; Thelwall *et al.*, 2012, 2010a). These tools use lexica of annotated emotional-bearing terms and syntax rules, building on previous survey studies of emotional words (Bradley and Lang, 1999; Pennebaker *et al.*, 2001).

Throughout the thesis, we employ one of such lexicon-based method called `SentiStrength` classifier (Thelwall *et al.*, 2012, 2010a), which is considered to be the state-of-the-art (Abbasi *et al.*, 2014; Kucuktunc *et al.*, 2012) sentiment detection tool. It has already been successfully applied in earlier research on the online data from `MySpace` (Thelwall *et al.*, 2010b), `Yahoo` (Kucuktunc *et al.*, 2012), IRC channels (Garas *et al.*, 2012), BBC, `Digg`, `YouTube` (Thelwall *et al.*, 2012), `Twitter` (Pfitzner *et al.*, 2012; Thelwall *et al.*, 2011, 2012) and `Wikipedia` (Iosub *et al.*, 2014).

`SentiStrength` identifies the hidden internal state of the author from their text, and not the emotion concerning the opinion of the text (Thelwall *et al.*, 2010a). This lexicon-based approach distinguishes itself from the other known methods of sentiment analysis, namely full-text machine learning and linguistic analysis (Thelwall *et al.*, 2011). The core of the algorithm is to predict the sentiment of a text, based upon the occurrences of the words from a lexical corpora, which contains the set of terms with known sentiment of a text. Initially, the lexicon contained 298 positive 456 negative human classified terms (Thelwall *et al.*, 2010a), later it was extended (Thelwall *et al.*, 2012) for negative terms by adding the negative General Inquirer terms (Stone *et al.*, 1966). Furthermore, the lexicon is adapted specifically towards the social web by including the "nonstandard" English terms and emoticons common to the Internet slang, *e.g.* luv, xoxo, lol, haha, muah, derived from `MySpace`, the social network website that was popular in 2009 (Thelwall *et al.*, 2010a). Seed words in the dictionary can be further sentiment fine-tuned towards specific domain, for instance politics. The strength of this classifier is an incorporation of various rules, which strengthens or weakens sentiments of the lexicon words detected in the short text message. Among the rules are syntactic rules, *e.g.* exclamation marks and punctuation, language modifiers and intensifiers, *e.g.* negation and booster words, and

spelling correction rules, such as deleting repeated letters. The final sentiment score of a text message consists of the most positive and the most negative emotion identified in it. Therefore, the classification returns both a positive E_p and negative E_n valence for each text, *i.e.* two discrete values in the range of $[+1, +5]$ and $[-5, -1]$ respectively. The two values (E_n, E_p) characterize the two-dimensional emotional charge of a single text, which agrees with the theory of the evaluative bivalence (Cacioppo and Berntson, 1994; Cacioppo and Gardner, 1999; Cacioppo *et al.*, 1997) in human beings and a recent evidence of the elicitation of mixed emotions and the feeling of ambivalence (Berrios *et al.*, 2015).

While `SentiStrength` returns the negative and positive emotional charge of the text, to detect arousal or dominance we employ a *lexicon of affective norms* of nearly 14,000 English words by Warriner *et al.* (2013).

Finally, since we focus our analysis on English texts only, for all online data we apply the *language classification* (Nakatani, 2010–current) and filter out all non-English posts. The default output of non-English texts from `SentiStrength` is $(-1, +1)$, therefore their removal ensures the validity of the measured sentiment and does not introduce an additional bias towards online posts with no emotional expression.

1.3 Research Questions

Collective phenomena in online mediated communication In the last years, much research has focused on the topic of online communities, or large groups of individuals that interact through an online medium. Collective phenomena such as the dynamics of trends (Wang and Huberman, 2012; Wu and Huberman, 2007), or viral marketing (Leskovec *et al.*, 2007) are usual topics that can be assessed with data from online communities. This weaker definition of community allows the testing of previous hypotheses from the social science such as Dunbar's number (Goncalves *et al.*, 2011), and the strength of weak links (Szell and Thurner, 2010).

Experimental studies like *e.g.* the one by (Salganik *et al.*, 2006) show the existence of social influence in online cultural markets. Other empirical studies reveal the existence of social influence on `Twitter` (Varol *et al.*, 2014), on movie ratings on `Imdb.com` (Lorenz, 2009), among collective response patterns on `Youtube` (Crane and Sornette, 2008), and also the different regimes of social influence on `Facebook` (Onnela and Reed-Tsochas, 2010).

Online communities exhibit differences in exchange of opinions and ideas than in offline communication. Empirical evidence shows that the mechanisms leading to group polarization are different in computer-mediated communication than in face-to-face discussions (Taylor and Macdonald, 2002). Another form of opinion manifestation prevalent in online communities is ratings and collective evaluations, *i.e.* `likes` or `dislikes`. Both high

amount of `likes` and high amount of `dislikes` show a highly polarized attitude of individuals towards the content. A recent finding by Van Mieghem (2011) suggests nontrivial relations between collective evaluations on `Reddit`, calling for psychological explanations for this phenomenon.

A currently emergent field of research is the study of political science from online data. Initial works showed the relevance of blogs (Adamic and Glance, 2005; Dodds and Danforth, 2010) in political discussions, and the evolution of *memes* related to political topics (Leskovec *et al.*, 2009). While the usage of these online data sources to predict the outcome of elections is still under debate (Gayo Avello *et al.*, 2011), recent works show that user behavior on `Twitter` can predict political alignment (Conover *et al.*, 2011a,b), and the party asymmetries in social interaction (Conover *et al.*, 2012). Furthermore according to Sobkowicz and Sobkowicz (2010) emotional interaction in political fora also reveals patterns of behavior regarding hate and political topics, with users taking the lead in controversial discussions.

We have outlined the state-of-the-art works on collective phenomena in online media and underlined the areas of active research. In Sections 1.3.1–1.3.4 we will introduce research questions in the mentioned areas and we will refer to the chapters of this thesis where we address each of the questions through the data-driven quantitative analysis and modelling.

1.3.1 Political Polarization in Multi-party Systems

Political polarization is an important ingredient in the functioning of a democratic system, but too much of it can lead to gridlock or even violent conflict. An excess of political homogeneity, on the other hand, may render democratic choice meaningless (Sunstein, 2003). It is therefore of fundamental interest to understand the factors shifting this delicate balance to one of the two extremes. Consequently, polarization has long been a central topic for political science. However, certain aspects in addressing political polarization so far have been disregarded. First, large parts of political polarization literature are concerned with two-party systems, particularly in the context of the United States (Waugh *et al.*, 2009). And second, polarization is generally conceptualized on the basis of positions in ideological space (Hetherington, 2009). Commonly used measures of political polarization are derived from those two premises, and therefore conceptualize *polarization* as two ideological blocks (*i.e.* parties) drifting apart on one political dimension, while increasing their internal agreement.

Towards network polarization Many democratic systems are characterized however by *more than two relevant parties*, making it problematic to apply standard ideology-based measures of polarization to them (Waugh *et al.*, 2009). Besides methodological concerns, we argue that an exclusive focus on ideological positions does not capture all aspects of

the term "polarization". It has been claimed that a sensible definition of polarization would have to comprise not only the ideological stances of the polarized set of individuals or parties, but also the *interactions* between them (Baldassarri and Bearman, 2007; Blau, 1977; Conover *et al.*, 2011a; Gruzd and Roy, 2014). At the same level of ideological polarization, we would ascribe a higher level of polarization to a set of political actors when: i) they display a bias towards positive interaction between actors of similar political positions, and ii) they tend towards negative interaction between actors with dissimilar political positions (Gruzd and Roy, 2014; Guerra *et al.*, 2013).

Assuming a feedback between opinion and network polarization, a polarized society is divided into by a small number of groups with high internal consensus and sharp disagreement between them (Flache and Macy, 2011). When establishing a *social link* means to agree on opinions to certain extent (Guerra *et al.*, 2013), *network polarization* is defined as a phenomenon in which the underlying social network of a society is composed of highly connected subgroups with weak inter-group connectivity (Conover *et al.*, 2011a; Guerra *et al.*, 2013). The terms *clusters*, *modules* or *communities* are often used as synonyms to define *groups* of individuals in a (social) network based on their link topology. Such assignments are usually not unique, *i.e.* individuals can be counted in different communities. Various algorithms for network partitioning exist to optimize this assignment, often based on the optimization of modularity metrics (Newman, 2006). By adopting a network science approach, we are able to capture and analyse *the interaction aspect of polarization*, and at the same time to expand the study of polarization to multi-party systems.

In Chapter 2, therefore, one of our research questions is to observe the evolvement of the network of interactions, *e.g.* supports, likes and comments, of *politicians* from Switzerland, a country that provides a representative example of a multi-party political framework. Additionally, this analysis is novel as it is performed within the context of *elite polarization* (or politicians) (Fiorina and Abrams, 2008; Hetherington, 2009), as a complementary view to previous studies of *mass polarization* through blogs and `Twitter` (Gruzd and Roy, 2014; Guerra *et al.*, 2013).

Furthermore, previous works on the dynamics of polarization use agent-based modelling approaches to simulate and analyze polarization in *opinion* dynamics. Here, we can distinguish between two different model classes. First, models with *binary* opinions, often called voter models, already imply that these are opposite opinions. The question is then about the *share* of opposite opinions in a population of agents (Schweitzer and Behera, 2009). Some scenarios show the emergence of a majority favoring one opinion, or even the convergence of the whole population toward the same opinion, called consensus. A polarization scenario in binary opinion models results in the *coexistence* of opposite opinions, often with almost equal share. Second, models with *continuous* opinions focus on the (partial) convergence of "neighboring" opinions such that groups of agents with the same opinion emerge. Polarization in such models can occur if two of these groups coexist,

without any possibility to reach consensus (Groeber *et al.*, 2009). In most cases, opinion dynamics models assume an interaction between opinions and the underlying communication structure, with agents of similar opinions communicating more frequently with each other than with dissimilar agents. This can be a result of homophily (*i.e.*, opinion difference influencing structure) or of social influence (structure influencing opinion). We will explore if those model assumptions also hold in the empirical social network of Swiss politicians.

Social interaction within and across parties Previous works centered around the US found strong polarization between left-leaning and right-leaning politicians (Saunders and Abramowitz, 2004). However, left- and right-leaning attitudes generally coincide with party membership in the US two-party system. This way, it cannot be distinguished whether polarization is created along the left-right dimension or along party lines, which can only be differentiated in a multi-party system. Furthermore, the two competing parties of the US elections show asymmetries in their online interaction, as shown in previous works using blogs (Adamic and Glance, 2005), `Twitter` (Conover *et al.*, 2011a), and `Youtube` (García *et al.*, 2012b). In both cases, the users supporting the Republican party, while less active and less numerous than the supporters of the Democratic party, show stronger social interaction. This can be attributed to mobilization efforts to encourage supporters to "win the online debate", *i.e.* become more involved in political online discussions. Inspired by these findings and hypothetical mechanisms, we aim at a generalization of these results, testing whether this mobilization pattern appears related to other political system, for instance in Swiss political system. With this we formulate the following hypothesis:

Hypothesis 1 *Compared to left-aligned communities, right-aligned communities are less influenced by content created by political campaigns, but have stronger social interaction regarding political topics.*

In Chapter 2, we test this hypothesis with data from online interactions of politicians on Swiss platform and we provide further evidence on US data from `YouTube` political channels in Chapter 6.

1.3.2 Priority Assignment and the Origin of Bursts in Human Activity

From the temporal aspect, one the most characteristic features of human interaction is its bursty nature. The temporal dynamics of human activity have been studied through different channels: written letter correspondence (Oliveira and Barabási, 2005), email communication (Barabasi, 2005; Malmgren *et al.*, 2008), short message exchange (Wu *et al.*, 2010), telephone calls (Jiang *et al.*, 2013) and IRC chats (Garas *et al.*, 2012). Remarkable patterns of temporal human activity were revealed through the analyses of *interevent time* τ, the time between consecutive individual-driven events such as sending emails or making calls, and *waiting* or *response time* τ_w, the time between receiving and

replying to a message. It has been shown that the distribution $P(\tau)$ of the interevent time and the distribution $P(\tau_w)$ of the waiting time follows the power law distribution, $P(\tau) \propto \tau^{-\alpha}$. This result implies that human activities are organized into the periods of frequent actions – bursts of activity – followed by long periods of inactivity.

Several mechanisms have been proposed to explain *the origin of bursts*, such as priority-queue processes (Barabasi, 2005), Poisson processes modulated by circadian and weekly cycles (Malmgren *et al.*, 2008), or a combination of both (Jo *et al.*, 2012). Recently, Wu *et al.* (2010) proposed a modified version of the priority-queue model (Barabasi, 2005) and provided evidence for bimodality of human activity: power law (bursty nature) and Poisson distribution. The reasoning behind the priority-queue models is that each message is assigned a *priority*, and those with high priority are answered soon after receipt. Prioritization of correspondence allows individuals to deal with multiple communication tasks. We suggest that attention and the priority to respond to messages and content depend on the emotions involved. Individuals therefore assign high priority to responding to a message dependent on the emotional charge of a message. Previous research in psychology (Miller, 1961) show that the strength of positive and negative stimuli differ in their gradient and reach. Experimental results find that the attraction strength of positive stimuli reaches farther than the repulsion strength of negative stimuli, a phenomenon called *positivity offset*. Additionally, the strength of the influence in the emotions of an individual was shown to increase faster with proximity to negative stimuli than to positive ones, which is commonly knowns as *negativity bias*. These asymmetric phenomena of emotional experience are integrated in Cacioppo's theory of evaluative bivalence. In the context of online communication and responses to posts on the Internet, therefore we formulate the following hypothesis:

Hypothesis 2 *Following Cacioppo's theory, user emotional expression has a positive baseline, however users respond to comments by others assigning higher priority to the response to negative stimuli.*

In Chapter 3, we test this hypothesis by analyzing a large number of collective responses from three popular online communities. Additionally, we formulate the priority to reply not only in terms of the speed of response but also in terms of the likelihood of response. To complete the full picture of emotional interactions between individuals, we analyze the types of emotional reactions elicited from stimuli of different valence or arousal and answer the questions such as whether positivity or negativity results in positive or negative replies, or which emotional stimuli lead to ambivalent or polarized emotional reactions.

1.3.3 Local and Global Collective Attention and Online Popularity

The dynamics of collective popularity of online content have been shown to be closely connected to the dynamics of *positive collective evaluations*. Namely, that more online `likes`, or more positive ratings or upvotes, lead to higher popularity of an item. Furthermore, popularity distribution in online communities has been observed to have statistical regularities, such as a large variance and heavy tails. For instance, popularity in online communities measured in the amount of votes for `Digg` stories (Van Mieghem *et al.*, 2011; Wu and Huberman, 2007), in the number of `likes` on `Reddit` (Van Mieghem, 2011), in the number of tweets in trending topics (Asur *et al.*, 2011; Wang and Huberman, 2012), or in the number of views on `YouTube` (Szabo and Huberman, 2010), have been consistently shown to be fit to the *log-normal distribution* (Mitzenmacher, 2004).

An interesting aspect of social influence on popularity of `Facebook` applications has been studied by Onnela and Reed-Tsochas (2010). Two modes of social influence have been distinguished, with one mode stemming from *local* signal and the other from *global* signal. While the local factor indicates how friends and local community influence an individual's behaviour, the global factor shapes the behaviour of an individual towards the product from the aggregate popularity.

Building on the notion of *local* and *global* social influence, we hypothesize that there exists a threshold of collective attention, such that above this threshold the amount of negative evaluations towards online item, *i.e.* `dislikes`, grow faster than that of positive evaluations, *e.g.* `likes`. This development leads to greater amount of `likes` and greater amount of `dislikes`, which essentially means the polarization of opinions among online community. We suggest that this regime shift in the collective evaluations is forced by social filtering mechanisms and recommender systems, the concept coined as *the filter bubble* (Pariser, 2011). And importantly, we assume that different emotional signatures might accompany local and global regime, for instance more negative and emotionally aroused discussions might be characteristic of a polarized (global) state. With this we state the following two hypotheses:

Hypothesis 3 *Popular discussions can be divided in two classes: one where the amount of users with minority opinions is negligible, and another one where the is a high degree of polarization. These two classes can be predicted through the emotional content of the discussions.*

Hypothesis 4 *The distributions of number of* `likes`, `dislikes`, *and emotional comments per item show regularities across different participatory media.*

In Chapter 4, we analyze distributions of collective evaluations in four online social media, we explore the evidences of the digital traces of the burst of the filter bubble, and we additionally investigate the role of emotional expressions in leading to extreme `likes` and `dislikes`, in other words to opinion polarization.

1.3.4 Models of Online Human Appraisal

While numerous studies have been performed on positive evaluations, studies on the dynamics of *negative collective evaluations* however are scarcer. Some recent works illustrate that social influence effects are present in movie ratings on `imdb.com` (Lorenz, 2009), and that controversiality expressed through movie ratings evolves with time (Amendola *et al.*, 2015). Additionally, herding effects have been observed in random manipulations of votes on `Reddit` (Weninger *et al.*, 2015), which shows that the way users vote depends on the votes of other users. Further research on `Reddit` (Van Mieghem, 2011) showed a non trivial dependency between `likes` and `dislikes` at the collective level.

Our study in Chapter 4 will show that not only positive evaluations follow the log-normal distribution, which has been observed in earlier works, but also negative evaluations, *e.g.* `dislikes`, are well fit to the log-normal distribution. Furthermore, the burst of the filter bubble results in the non-linear coupling between the evaluations in local and global regimes. In light of these two results, we believe that it is possible to investigate underlying processes that produce the observed distribution. Namely, the log-normal distribution of collective evaluations is the sign of the existence of *multiplicative growth processes* of social interaction.

The multiplicative growth models have already been successfully applied to recovering the log-normal distribution in the number of visitors to a website (Adamic, 2001; Huberman and Adamic, 1999), in the number of edits of `Wikipedia` pages (Wilkinson and Huberman, 2007), in the number of story reads on `Digg` (Van Mieghem *et al.*, 2011), in the number of views on `YouTube` (Szabo and Huberman, 2010), in the number of `Twitter` trending topics (Asur *et al.*, 2011) and in the number of online petitions (Yasseri *et al.*, 2013).

In Chapter 5, we build the statistical model based on the existing models of multiplicative growth in order to reproduce the log-normal distribution of collective evaluations. Additionally, we account for the non-linear coupling in the relationship between `likes` and `dislikes` and reproduce the local and global regimes of collective attention and the threshold of the burst of the filter bubble.

Thesis overview Summing up the presented research questions and reference to the relevant chapters, in Figure 1.4 we show the flow diagram of the thesis, which depicts the connection between the chapters, and may help reader to be guided through the thesis.

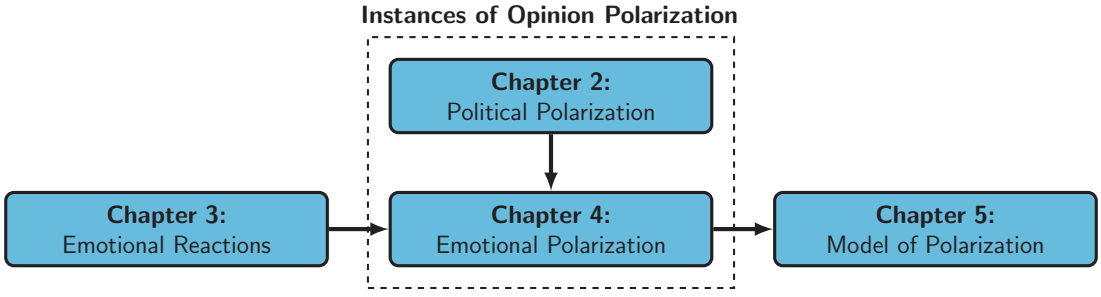


Figure 1.4: The flow diagram of the thesis.

1.4 Opinion Polarization in Online Communities

One of the basic ideas of studying opinion polarization through online channels is the abundance and granularity of online data, which comes in various forms, like user social network profiles, digital traces of conversation, *e.g.* fora, blogs and posts, and online metrics of user activity and interaction, such as `likes`, `dislikes`, views and `shares`. Such high resolution data on human activity allows us to get an in-depth understanding of mechanisms behind opinion polarization. Previous research that leveraged on the power of online resources has investigated polarization from the network perspective in political domain, *e.g.* democratic and republican party blogs (Adamic and Glance, 2005), follower and mention links on `Twitter` (Conover *et al.*, 2011a), as well as in non-political domains like friendship networks (Guerra *et al.*, 2013), and cultural expression (García and Tanase, 2013).

Another reason, why studying polarization online is attractive, is anonymity (Sunstein, 2002). In anonymous settings, participation of users increases (Blau and Caspi, 2010; Konnikova, 2013), as users have a will to stand out individually and they do not fear the risk of rejection. Such settings might dampen the amplification of *the spiral of silence* (Noelle-Neumann, 1974). This theory explains the emergence of the single public dominant opinion in a situation when the minority opinion holders are not ready to speak up due to the fear of separation or exclusion from the majority opinion holders. Contrary to the spiral of silence, online settings might motivate to more salient expression of different opinions and eventually lead to opinion polarization. However, the questions of the extent to which opinion polarization is present on the Internet and that of the role of online media in collective opinion formation remains polarizing even for the scholars (Schulz and Roessler, 2012).

One group of scholars does not observe an increasing opinion polarization in online settings. Wu and Huberman (2008) show that contrary to the common phenomenon of increasing group polarization observed offline, on the Internet groups rather move towards moderate opinions. Gentzkow and Shapiro (2010) continue and agree that ideological segregation on the Internet has not increased over time, and is much lower than the segregation in the real world. Scholars who are in favor of decreasing online polarization agree, that, thanks to the Internet, individuals get access to the diverse news resources that do not necessarily reflect their initial opinion stance. This exposure to different opinions softens their initial position, and on an aggregate level decreases the online opinion polarization.

Another group of scholars, however, claim that online polarization should be higher than offline for the following reason. Apart from the anonymity effect discussed earlier, too much of exposure to the heterogeneous information on the Internet leads to users to becoming enclosed in a so-called information or filter bubble. Users either control their information environment themselves, *e.g.* subscription, choosing whom to follow, or they are controlled

by the technological filters, *e.g.* recommender systems. The selective exposure theory, or the confirmation bias, or the cognitive dissonance theory (Festinger, 1957) might explain why users opt out to filter information on the Internet themselves. Users deliberately reach out for those news that confirm their initial position. This is due to the unpleasant and threatening self-arousal state that they experience, when they encounter information that is in conflict with their initial opinions. To reduce the caused dissonance, individuals tend to prefer information that supports their pre-existing views. While selective exposure to conforming information is done by the individuals, the technological filters like search engines and social networks latently sort "relevant" information that conforms to user's opinion. This is achieved by deducing the user's alignment through the history of their online activity, *e.g.* search terms or web page visits. In both cases, communication becomes enclosed within each group and interaction between groups diminishes, which eventually enforces polarization of opinion space.

In the absence of consensus among scholars on the extent of the presence of opinion polarization on the Internet, the role of online media in opinion formation and opinion diversity remains an open research question that requires further empirical work.

Between consensus and conflict Lastly, scholars unanimously agree that the notion of *conflict* goes unseparated from the notion of polarization. Conflict is projected onto various social system dimensions, be it racial conflict, ethnic conflict, or political conflict. Esteban and Schneider (2008) discuss that highly polarized state of a system in whichever context increases the risk of conflict, including armed violence. Hetherington (2009); Levendusky (2013) indicate negative consequences of polarization for the mass public and for the public policy, such as introducing uncertainty in policy decision making. In politics, polarization may harm the electoral process by further disconnecting voters and elites.

If we view polarization as negative and try to understand the mechanisms of a process behind it in order to avoid it, then we should mention the opposite of a highly polarized state which is the state of *social consensus*. Benefits of reaching consensus are numerous, especially in the problem solving situations and in decision making scenarios (Hefte, 2015). However, in opinion formation there is a danger that consesus might be alleged, for instance as a product of propaganda, which is described as the process of manufacturing the consent (Herman and Chomsky, 2010). One of the negative characteristics of propaganda is the impression of the emergence of a single public dominant opinion which leads to suppressing the voices of minority due to their fear of social rejection (Noelle-Neumann, 1974). One of the counter-balancing forces against the "spiral of silence" can be the polarization of opinions via the most natural form of human communication, namely conversation. Moreover, the usage of the Internet, proliferation of blogs and social networks, and abundance of online discussion platforms can facilitate the untwisting of the spiral of silence.

The concept of dualism in society addressed by Simmel (1906) states that in order to reach group unity in society, conflict is necessary. He continues that no society exists without the notion of conflict and that it is not only an inevitable part of society, but it is also beneficial. The idea of duality can be projected onto opinion formation: as long as we strive for opinion consensus – in order to reach it, we also need to hold on to opinion diversity. The basis of any democratic society is polarization of opinions. Therefore, it is important for a social system to be polarized to some extent to achieve consensus in problem solving, and it is important to be polarized to some extent to not approach conflict.

Chapter 2

Political Polarization in Online Communities

Summary

Political polarization is traditionally analyzed through the ideological stances of groups and parties, but it also has a behavioral component that manifests in the interactions between individuals. We present an empirical analysis of the digital traces of politicians in politnetz.ch, a Swiss online platform focused on political activity, in which politicians interact by creating support links, comments, and likes. We analyze network polarization as the level of intra-party cohesion with respect to inter-party connectivity, finding that supports show a very strongly polarized structure with respect to party alignment. The analysis of this multiplex network shows that each layer of interaction contains relevant information, where comment groups follow topics related to Swiss politics. Our analysis reveals that polarization in the layer of likes evolves in time, increasing close to the federal elections of 2011. Furthermore, we analyze the internal social network of each party through metrics related to hierarchical structures, information efficiency, and social resilience. Our results suggest that the online social structure of a party is related to its ideology, and reveal that the degree of connectivity across two parties increases when they are close in the ideological space of a multi-party system.

Based on paper "Ideological and temporal components of network polarization in online political participatory media" by David Garcia, Adiya Abisheva, Simon Schweighofer, Uwe Serdült, and Frank Schweitzer, *Policy & Internet* 7(1), pp. 46–79, (2015). A.A. contributed to designing the research questions. A.A. produced the majority of the statistical analyses and the plots. A.A. was involved in writing the manuscript.

2.1 Introduction

In this Chapter, we investigate one of the instances of opinion polarization, namely the political polarization. We have presented a number of research questions and a hypothesis associated with political polarization in Section 1.3.1. Among those questions, in this Chapter we will look at the interaction aspect of opinion polarization and differences between the left- and right-aligned political communities, see Hypothesis 1, Section 1.3.1.

Online political participatory media, like `opencongress.org`[1] and `politnetz.ch`,[2] serve as a digital representation of a political system where voters and politicians can discuss in an online medium. These political participatory media serve as crowdsourcing platforms for the proposal and discussion of policies, leaving digital traces that allow unprecedented quantitative analyses of political interaction. In this study, we present our analysis of `politnetz.ch`, a Swiss platform that allowed us to obtain data about online political participation. On `politnetz.ch`, politicians and citizens freely discuss political topics, and also weave a social network around them by expressing their `support` for politicians or by `liking` each others' contributions. Though `politnetz.ch` is of course subject to many limitations and distortions, it can be seen as a fairly faithful online representation of Swiss politics. The plurality of Swiss politics allows us to study polarization along party, as well as along ideological lines. Additionally, the real-time quality of `politnetz.ch` data enables us to analyze polarization with a very high temporal resolution. While traditional research mainly focuses on changes in polarization within years or decades (McCarty *et al.*, 2006), we can detect them within the scope of days and weeks. This makes it possible to quantify the influence of day-to-day political events, such as elections and referendums (which are of particular relevance in Swiss politics (Serdült, 2014)). Therefore, we analyze the network topology of `politnetz.ch`, its evolution over time, and additionally the topologies the social networks within each party. As a consequence, we characterize the role of elections and party ideology in network polarization, both at the party and at the global level. The fact that most politicians in `politnetz.ch` clearly state their party membership and make their online interaction with other politicians via `likes`, `supports`, `comments` publicly available, allows us to test the following hypotheses: i) political polarization is present in layers with a positive connotation, *i.e.* `supports` and `likes`, ii) polarization, if present, is not solely grounded upon political party alignment, but also depends on politically relevant events, such as elections, and on the distance between parties in ideological space.

Online participatory media offer different possibilities of interaction between individuals. Online actions, such as creating a social link or commenting on a post, can be used in different context. Nevertheless, interdependencies might exist between these interaction types. For example, if users press the like button only to posts they positively comment, the liking action would not contribute additional information to the communication process.

[1] `http://www.opencongress.org`
[2] `http://www.politnetz.ch`

In this work, one of our goals is to quantify how much information about one interaction type is contained in another, assessing the added value of including all these interaction types in the analysis of online communities. In our analysis, we explore the three main interaction types between politicians in `politnetz.ch`: `supports`, `likes`, and `comments`, to measure the differences of politicians' behavior in each interaction context. For instance, do politicians only `like` posts of politicians they `support`? Do the patterns of `support` among politicians project onto `comments` and `likes`? In our analysis, we aim to discover groups of politicians determined by the networks of `supports`, `likes` and `comments`, and not necessarily by their party affiliation. It might be the case that two politicians that are the members of the same party, might not support each other, and conversely, it can be the case that two politicians that are *not* the members of the same party *do* support each other. How often do such scenarios occur? Does the party affiliation of a politician define whom they `support`, `like`, and `comment`, producing network polarization?

Finally, we ask the question of whether the social structure of parties is related to their ideology. Previous research in the US political system showed differences in online network topology between the right- and left-aligned groups (Conover *et al.*, 2011a). This opens questions whether i) there are differences in the party structures of the political systems with more than two parties, and whether ii) the pattern in the US for right- and left-subcommunities holds in a multi-party system. We study Switzerland as an example of a country with a multi-party political system, with the presence of major and minor parties and additional ideological dimensions beyond left and right.

2.2 Dataset Description

`Politnetz.ch` is an online platform that enhances communication among politicians and voters in Switzerland. Profiles in `politnetz.ch` are registered either for voters or politicians, where politician profiles have an additional section called "Political information". This section includes the party a politician belongs to, selected from among valid parties in Switzerland.[3] Registration as a politician on `politnetz.ch` is not verified by the platform, due to the fact that politicians are not motivated to misrepresent themselves. This becomes evident in their profile information: two thirds of the politician profiles include links to their homepages, while none of the voter profiles has an external website link.

With respect to the amount of politicians in Switzerland, `politnetz.ch` contains a large set of politician accounts, with a certain bias towards German-speaking regions. The amount of voter accounts is more limited with respect to the size of the electorate, and

[3] "Political information" is exclusive to politician profiles and includes 4 fields: party affiliation, political career, duration as a party member at the federal or local level, and involvement if any in Swiss "Verein", *e.g.* NPOs, NGOs, trade unions, or other organizations. The set of parties to choose from is limited to 65 parties and politically relevant organizations in Switzerland. Politicians are not able to insert other party names.

Party	Description	N	Colour
SP	Social Democratic Party	609	
SVP	Swiss People's Party, incl. EDU	484	
FDP	The Liberals	473	
Christian	CSP, CVP, EVP	423	
Grüne	Green Party	298	
GLP	Green Liberal Party	286	
BDP	Conservative Democratic Party	151	
Piraten	Pirate Party	93	
AL	Alternative Left, incl. PdA	28	
	Unaligned	596	
	Independent	47	
	NGO/Union	101	
	No affiliation	448	

Table 2.1: Number of politicians in each of the 9 parties and in the unaligned category. Politicians labeled as unaligned can be further divided into three subgroups: independent politicians, NGO/NPO/trade union representatives, and politicians with no party affiliation. The fourth column shows colour codes of each of the 9 parties, which we use throughout this article.

voters do not have any field for party affiliation in their profile. As mentioned above, we focus exclusively on politicians' accounts, excluding voters accounts from all analyses.

Politicians have three major means of social interaction in `politnetz.ch`: i) they can `support` other politicians, ii) they can write posts and `comments` on the posts of other politicians, and iii) they can explicitly `like` posts, which are publicly displayed in their profile. Our `politnetz.ch` dataset includes the full track of interaction between politicians for more than two years, including 3441 politicians, which created 16699 `support` links, 45627 `comments`, and 10839 `likes` to posts.

In our dataset, more than 80% of the politicians declare themselves as members of an existing party. We simplify the party affiliation data by merging local and youth versions of the same party, creating table in which each politician is mapped to one of the 9 parties, or left unaligned. The latter tag – "unaligned" politician – is an umbrella label for three distinct categories of political affiliation: 1.3% of the politicians stated they are independent politicians,[4] 2.9% chose membership to non-governmental and nonprofit organizations, and the remaining 13% of the politicians who did not provide affiliation information to any party or organization, which in general are politicians active at the local level without alignment to any party at the federal level. In Table 2.1 we show the absolute count of politicians in the 9 parties and in the unaligned category.

[4]In German *Parteilos*, `http://www.parteifrei.ch`

2.3 Methods

Multiplex network analysis The digital traces of politicians in `politnetz.ch` allow us to study interaction in three **network layers**, one composed of `supports`, a second one of `likes`, and a third one of `comments`. Every node belongs to each layer, and represents a politician with an account in `politnetz.ch`. A politician p_1 has a directed link to another politician p_2 in the `supports` layer if p_1 has p_2 in its list of supported politicians. The `likes` and `comments` layers are also directed, but in addition, links have weights equivalent to respectively the amount of `likes` and `comments` that p_1 gave to the posts and comments of p_2. These three layers compose a **multiplex network**, also known as a multimodal, multirelational or multivariate network (Menichetti *et al.*, 2013), as depicted in Figure 2.1. The paradigmatic example of a multiplex network is a social network with different types of social relationships (friendship, business, or family) (Szell *et al.*, 2010). Examples of previously analyzed multiplex networks are air transportation networks, in which airports are connected through different airlines (Cardillo *et al.*, 2012), and online videogames where players can fight, trade, or communicate with each other (Szell *et al.*, 2010). The concept of a multiplex network can be applied to social media in a political context, as for example politicians can follow, retweet, and mention others in Twitter (Aragón *et al.*, 2013; Lietz *et al.*, 2014).

The network layers are visualized in Figure 2.2, where nodes are colored according to their party alignment. Following the Fruchterman-Reingold layout algorithm (Fruchterman and Reingold, 1991), which locates connected nodes closer to each other, an initial observation of Figure 2.2 motivates our research question: politicians seem to be polarized along party lines when creating `support` links, while this pattern is not so clear for `likes` and `comments`.

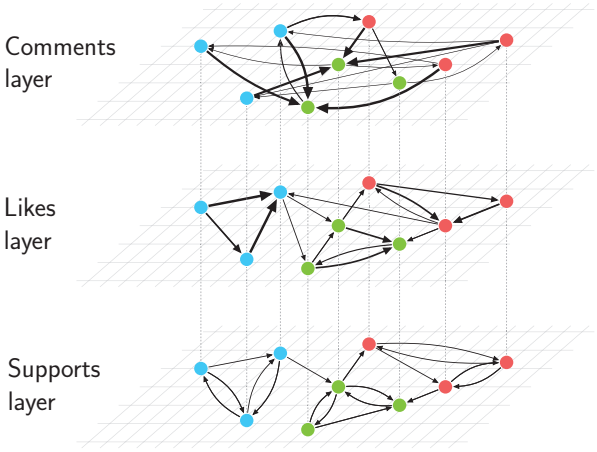


Figure 2.1: Three layers of the *multiplex* network in `politnetz.ch`. The layer of `supports` is directed and unweighted, and the layers of `likes` and `comments` are directed and weighted. Node colors illustrate an party affiliation, and link widths are proportional to their weights.

(a) `Supports` layer, $|V| = 2,740$, $|E| = 16,657$

(b) `Likes` layer, $|V| = 1,147$, $|E| = 6,386$

(c) `Comments` layer, $|V| = 1,091$, $|E| = 6,800$

Figure 2.2: Visualization of network layers of `supports`, `likes` and `comments` excluding unaligned politicians. Colours of the nodes are labelled according to the parties self-reported by politicians. Party colors are reported in Table 2.1. The networks are drawn using the Fruchterman-Reingold layout algorithm (Fruchterman and Reingold, 1991).

In this multiplex network scenario, if a layer is ignored in the analysis, relevant information about the interaction between nodes might be lost. By measuring the interdependence of interaction types among politicians, it is possible to address questions on the relationships across layers. These relationships can exist at the link (microscopic) level, where links between the same pairs of nodes co-occur in two different layers, or at the group (mesoscopic) level, where the group to which each node belongs is similar across layers.

To measure the overlap between layers at the link level, we calculate the **Partial Jaccard similarity coefficient** between the sets of links in two layers. Given that the links of some layers are weighted, we calculate this coefficient over binary versions of the weighted layers, *i.e.* in which all weights are taken as 1. The Jaccard coefficient, see Appendix A.1, measures similarity in sets, and is defined as the size of the intersection divided by the size of the union of both sets. It shows the tendency of nodes connected in one layer to also connect in another layer.

The second metric we apply to both at the link and the group levels is the **Normalized Mutual Information** (NMI) among layers, see Appendix A.1. The NMI of X with respect to Y, NMI$(X|Y)$, is the mutual information (MI), see Appendix A.1, between X and Y divided over the entropy of X. If the MI tells how much shared information there is between X and Y, and the entropy of X quantifies information content of X itself, then the NMI with respect to Y measures to which extent Y contributes to information content of X. In other words, it gives a fraction of information in X that is attributed to Y, hence the NMI is also known as the *uncertainty coefficient* and has values between 0 and 1. With respect to X, if the NMI $\to 1$, then knowing layer Y can predict most of

the links in layer X; conversely, if the NMI $\to 0$, then there is no dependence between the layers. To compute the NMI, we first calculate information content of each matrix via the Shannon entropies (Appendix A.1) with the Miller-Madow correction (Appendix A.1). Then, we compute the empirical mutual information between the layers, and finally obtain the normalized mutual information with respect to each layer. At the link level, the NMI of layer X with respect to layer Y quantifies to which extent links in the network layer Y tell us about links in the network layer X. At the group level, the NMI is computed over group labels, which allows the additional comparison with the *ground truth* of party affiliation, *i.e.* how much information about party affiliation is contained in the group structure of certain layer of interaction.

Network polarization Given a partition of politicians into groups, for example given their party affiliation, their network polarization can be computed through modularity metrics. We apply the **Q-modularity** metric (Newman, 2006), see Appendix A.2, to each layer, measuring the tendency of politicians to link to other politicians in the same group and to avoid politicians of other groups. The Q-modularity of a certain partition of politicians in a layer has a value between -1 and 1, where 1 implies that all links are within groups, and -1 that all links are across groups. This metric is specially interesting, since it measures a comparison between the empirical network and the ensemble of random networks with the same in-degree and out-degree sequences, allowing us to quantify the significance of our polarization estimate against a null model.

The natural partition of politicians is given by their party labels, from which we obtain 9 groups and the *unaligned* group as explained in Table 2.1. For each layer, we compute the network polarization along party lines through the Q-modularity of party labels Q_{party}. In addition, politicians can be partitioned in other groups rather than parties, which could have higher modularity than Q_{party}. Finding such a partition with the maximal modularity is a computationally expensive problem, also known as community detection problem. We approximately solve it by applying the state-of-the-art community detection algorithms from the open source software `Radatools` (Gómez, 2011). Following is the list of the algorithms that are implemented in the software, abbreviations of the algorithms are given in brackets: extremal optimization (e) (Duch and Arenas, 2005), spectral optimization (s) (Newman, 2006), fast algorithm (f) (Newman, 2003), fine-tuning by reposition (r). The algorithms (s) and (e) that have a stochastic behaviour must be executed with several repetitions, *e.g.* (esrfr-30) means perform 30 repetitions of the algorithm (e), 30 repetitions of (s), 1 times (r), 1 times (f) and finally 1 times (r), refer to the Gómez (2011) for more details. We run the following combinations of the algorithms for each layer: e-1, esrfr-30, r-1, f-1, s-10, rfr-1, rsrfr-30, and store the partition that gives the highest Q-modularity from among all the runs. This optimal partition, which is computed independently from the party alignment, determines the Q_{comp} polarization score. We repeated this analysis

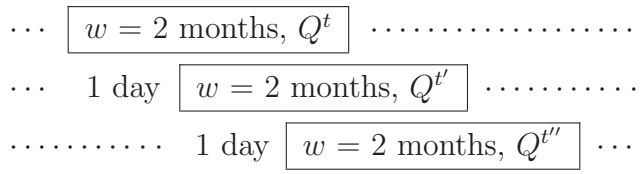


Figure 2.3: Time-series segmentation approach of computing modularity score of a network with a sliding window. At each shift of the window, modularity is computed on the network with the links present within the current time window. The size of the window is set at 2 months, the step of the window shift is 1 day.

for each layer, which resulted in three values of the Q_{comp}: one for `supports`, one for `likes`, and another one for `comments`. This way, for each layer we have another network polarization value Q_{comp}, computationally found from the empirical data in that layer instead of from party alignment.

To understand the origins of polarization, we perform temporal and topical analysis. First, we analyze the time series of network polarization over a sliding time window of two months, which takes into account only the links created within that window, see Figure 2.3. This method is known as a *rolling window* technique. The network data is divided along the time axis in the overlapping intervals of a fixed size. Each interval overlaps with the next one on a fixed subinterval, where the difference in their starting dates is a constant *step* size. This technique smooths out the data through the aggregation of links within the window, and preserves a granularity defined by the step size. For our analysis, we chose a window of 2 months and a 1-day step, in order to aggregate sufficient data in each window and to preserve daily resolution in the time series. In every time interval, we compute the Q-modularity on the network with links that have timestamps within the specified interval. We illustrate the method with an example and Figure 2.3. At the timestamp of the first link, t, we obtain the first-window modularity score for the interval $[t, t + 2 \text{ months}]$, then we move the window one step forward to the starting time $t' = t + 1$ day, recompute the metric on the network with the links present within the new window, to obtain the modularity score at the time t'. We continue sliding the time window, computing the modularity until reaching the timestamp of the last link. Hence, with this approach, we record the evolution of the network polarization around the dates in each time window. This time series analysis of network polarization allows us to empirically test if polarization changes around politically relevant events, such as elections.

To explore the origins of polarization, we also perform topical analysis of the comments layer. It includes contextual information in the text of the comments between politicians of each group. To identify the topics discussed in each group of the comments layer, we compute the Pointwise Mutual Information (PMI) of each word in each group, see Appendix A.1. The PMI of a word compares its relative frequency within the group, $p(x|y)$,

with its frequency in all comments from all politicians, $p(x)$. To control for statistical significance, we performed χ^2 tests on the ratio of frequencies, filtering out words with the PMI significance below the 99% confidence level. The list of words with significantly high PMI of each comment group can be referred to as the discussion topics of each group of politicians which can shed light on the reasons of polarization among politicians.

Social structure and ideological position of parties Beyond network polarization, the multiplex network among politicians contains information about the social structure of parties and the interaction of politicians across parties. To complete our picture of Swiss online political activity, we analyze the intra-party and inter-party structures present in the network.

First, we apply metrics from social network analysis to compare the network topologies inside each party, similarly to previous works on the US (Conover *et al.*, 2012) and Spain (Aragón *et al.*, 2013). With three types of social interaction, we first have to select the layer that captures the social structure of a political party in online participatory media. The `supports` layer carries a positive connotation that does not change or evolve in time. We choose to analyze `support` links, as leaders of the party or politicians with authority will accumulate more `support` links, but the amount of `likes` and `comments` they receive depends on their activity in `politnetz.ch`. For each party, we extract its internal network of supports, capturing a snapshot of its online social structure in terms of leadership. We then calculate three network metrics to estimate three structural properties of each network:

- **Hierarchical structure: in-degree centralization.** The basic idea of the network centralization is to calculate the deviation of the in-degree of each node from the *most central* node, which has a special position with respect to the rest in terms of influence (Freeman, 1978), see Appendix A.2. This way, in-degree centralization is computed as an average difference between the in-degree of the politician with the most supports within the party and the rest. A party with an in-degree centralization of 1 would look like a star in which the central node attracts all supports, representing a network with the strongest hierarchical structure. A party with in-degree centralization of 0 has support links distributed in a way such that every node has exactly the same amount of supporters, showing the most egalitarian and least hierarchical structure.
- **Information efficiency: average path length.** This social network metric measures the efficiency of information transport in a network: shorter path length indicates an easily traversable network, in which it takes fewer steps to reach any other node. The average path length is defined as the sum of the shortest paths between all pairs of nodes in a network normalized over the number of all possible links in a network with the same number of nodes, see Appendix A.2. We compute the average path length

between all pairs of connected nodes, in order to have a measure of the information efficiency of their social structure.

- **Social resilience: maximum k-core.** The ability of a social group to withstand external stresses is known as *social resilience*. Social networks can display different levels of social resilience from the point of view of cascades of nodes leaving the network, having a resilient structure if such cascades have small impact. Under the assumption of rationality, the social resilience of a network can be measured through the k-core decomposition (García *et al.*, 2013a), indicating how many nodes will remain under adverse conditions. This method assigns a k-core value to each node by means of a pruning mechanism, see Appendix A.2. In essence, the k-core captures cohesive regions of a network (Seidman, 1983), which are subsets characterized with high connectedness, formally defined as the maximal subnetwork in which all nodes have a degree at least k. Applying the k-core decomposition on the subnetwork of each party, we aim at discovering such a resilient core of political leaders, estimating the social resilience of a party as the maximum k-core number of is social network.

Second, we measure the level of inter-party connectivity by means of the **demodularity** score, see Appendix A.2, which measures the opposite of Q-modularity: the tendency of a party to connect to another, as compared to a random ensemble of networks. The demodularity score measures to which extent parties interact with other parties, or how strongly politicians of one party preferentially attach to politicians of *another* party. This way, we compute a score of demodularity from each party to each other party.

Party positions in ideological space We quantify the ideological position of Swiss parties along the dimensions *Left-Right* and *Conservative-Liberal* stance. This is necessary to capture the multi-party system of Switzerland, in which the position of parties cannot be simply mapped to a Left-Right dimension. We use the party scores of external surveys provided by Hermann and Städler (2014). The authors of the study give the following interpretation for both types of ideological dimensions: the *Left-Right* dimension expresses the understanding of the concept of the state by each party. Left-wing politics sees the role of the state in promotion of the economic and social well-being of citizens with equally distributed welfare; while right-wing politics understands the primary role of the state as maintaining order and security. Additionally, the difference between the left- and right-politics comes with the difference in priorities set by each side. While the left-wing politics is concerned with the issues on the environment protection and asylum granting policies; the right-wing politics is committed to strengthening of security forces, and to the competitiveness of the economy. The *Conservative-Liberal* dimension encompasses the concepts of openness and willingness for political changes. It covers the stance of parties on economics, social and political issues, for instance, the position of a party on questions ranging from globalisation to abortion. In the social area, political issues such as

abortion, partnership law, etc. are covered; in the economic area, questions of structural change, competition and attitude towards globalisation, including reduction of subsidies, free advertising, free trade, etc., are discussed. And finally, in the state policy field, debates are between centralisation and internationalization, such as Schengen and international peace-keeping missions, versus the preservation of the federal system. These values are consistent with other sources of party positioning data in Switzerland (Germann *et al.*, 2014), stemming from Voting Advice Applications such as `preferencematcher.org`[5] and `smartvote.ch`.[6]

2.4 Polarization in a Multiplex Network

2.4.1 Layer Similarity

Before measuring polarization in the three layers of the `politnetz.ch` multiplex network, we need to verify if each layer contains additional information, or if one can be predicted based on another. Our first step is to measure the similarity between layers at the link level, quantifying the tendency of pairs of nodes to connect in more than one layer. For this analysis, we regard information in a network layer as expressed via presence or absence of the interaction links between politicians, specifically we extracted the binary, directed adjacency matrices of `supports`, `likes`, and `comments`.

We computed link overlaps between layers, as the ratio of links of layer $\mathcal{X}$ that co-ocurr with links in $\mathcal{Y}$, among all links in $\mathcal{Y}$. We measure this overlap through the partial Jaccard coefficient of variables X and Y, which take value 1 if certain link is present in layer $\mathcal{X}$ and in layer $\mathcal{Y}$ respectively, and 0 otherwise. To test statistically the significance of these metrics, we apply the jackknife bootstrapping method,[7] creating subsets of the network obtained by leaving one node out. Overlaps across layers are reported in Table 2.2, revealing that the maximum overlap in the data is between `comments` and `likes`, where 25.45% of the `comments` links have an associated link in the `likes` layer. While significantly higher than zero, these values are relatively low, with more than 70% of the links not overlapping across layers.

In addition, we computed the Normalized Mutual Information (NMI) between every pair of the layers which estimates how much information of one layer is contained in the links of

[5] `http://www.preferencematcher.org`

[6] `http://smartvote.ch`

[7] *Bootstrapping* is one type of the resampling methods to assess accuracy (the bias and variance) of estimates or statistics (*e.g.* the mean or other metric in study). *Jackknife* resampling is referred to obtaining a new sample called *bootstrap* from the existing data by leaving one observation out. With N observations we obtain N bootstrap samples. For each sample, we compute an estimate, in our case the Jaccard coefficient and the NMI. With the set of N statistics obtained from the bootstrap samples, we are able to calculate the standard deviation of our estimate.

layer $\mathcal{X}$ layer $\mathcal{Y}$	likes supports	supports likes	comments supports	supports comments	comments likes	likes comments
overlap	18.13%	6.96%	7.13%	2.91%	23.9%	25.45%
NMI	8.5%	3.7%	2.5%	1.2%	15.2%	16%

Table 2.2: Link overlap and NMI across layers. Each measure was computed over jackknife bootstrap estimates on each node, giving values of $2\sigma < 10^{-3}$.

the other one. The NMI of layer $\mathcal{X}$ with respect to $\mathcal{X}$ tells us which fraction of information of the layer $\mathcal{X}$ is attributed to knowing the layer $\mathcal{Y}$. Table 2.2 shows the statistics for the NMI between the three layers. All values are significantly larger than 0, and reveal weak correlations between the layers. 8.5% of the information in the `likes` layer is contained in the `supports` layer, and less than 4% of the information in the `supports` layer is contained in the `likes` layer. `Supports` give 2.5% of the information in the `comments` binary network, and only 1.2% is contained in the opposite direction. The `likes` and `comments` layers share normalized information of about 16%, showing that there is a bit of information shared across layers, but that they greatly differ in most of their variance, in particular for the `supports` layer. These low levels of overlap and NMI indicate that each layer contains independent information content that does not trivially simplify within a collapsed version of the network.

2.4.2 Network Polarization

The visualization of the three layers of the network in Figure 2.2 suggests the existence of network polarization in the `supports` layer, where politicians of the same party appear close to each other. We quantify the level of network polarization among politicians in each layer, given two types of partitions: i) by their party affiliation, producing the Q_{party} modularity score, and ii) by the groups found through computational methods on the empirical data of each layer, resulting in the modularity score Q_{comp}. Naturally, low modularity scores ($-0.5 \leq Q < 0.3$) imply that no polarization exists among politicians; high modularity scores ($0.3 \leq Q < 1$) will indicate the existence of polarization. Similar values of Q_{party} and Q_{comp} indicate that the maximal partition in a layer is close to the ground truth of party affiliation, while different values suggest that groups in a layer are not created due to polarization along party lines. Across network layers, we hypothesize that Q_{party} is strong in the layers with the positive semantics such as the links of `supports` or `likes`. Furthermore, we investigate the role of the unaligned politicians by measuring the network modularity including and excluding unaligned politicians. We aim at testing whether these nodes act as cross-border nodes between parties, therefore decreasing polarization when included in the analysis. For each measure of polarization, we test its statistical significance by applying the jackknife bootstrapping test on the networks – by recomputing the modularity score on the bootstraps – each time with one node left out.

	Supports			Likes			Comments		
	$\langle Q \rangle$	2σ	$\lvert C \rvert$	$\langle Q \rangle$	2σ	$\lvert C \rvert$	$\langle Q \rangle$	2σ	$\lvert C \rvert$
Q_{party}(incl.)	0.677	$6.8 \cdot 10^{-4}$	10	0.303	$11 \cdot 10^{-4}$	10	-0.009	$7.6 \cdot 10^{-4}$	10
Q_{comp}(incl.)	0.746	$6.1 \cdot 10^{-4}$	13	0.460	$21 \cdot 10^{-4}$	12	0.341	$26 \cdot 10^{-4}$	16
Q_{party}(excl.)	0.743	$6.1 \cdot 10^{-4}$	9	0.377	$13 \cdot 10^{-4}$	9	-0.009	$13 \cdot 10^{-4}$	9
Q_{comp}(excl.)	0.745	$6.1 \cdot 10^{-4}$	10	0.472	$25 \cdot 10^{-4}$	17	0.336	$41 \cdot 10^{-4}$	16

Table 2.3: Modularity score and number of groups found by the community detection algorithms and given politicians' parties labels for the layers including (first two rows) and excluding the unaligned politicians. Standard deviations are calculated through the jackknife bootstrapping on nodes. We report the mean of modularity for bootstrap estimates, $\langle Q \rangle$.

From the results in Table 2.3, we observe that the modularity score slightly differs when unaligned politicians are ignored, having lower polarization when they are present. This indicates that the unaligned group is not cohesive, as it does not represent an explicit party, and unaligned politicians connect across parties. For this reason, we remove unaligned politicians from our subsequent analysis of polarization, as their absence of affiliation does not signal their belonging to an additional group.

Among the three types of interactions, the `supports` layer shows the highest modularity score 0.74 for both Q_{party} and Q_{comp}, showing that, when making a `support` link, politicians act as partisans. The `likes` layer is also polarized, but the modularity score 0.47 is lower than in the `supports` layer. Hence, liking a post is still a signal of an adherence to a party, however cross-party `like` links are more frequent in comparison to cross-party `support` links. Finally, the `comments` layer divided algorithmically hints on a modular structure of the network resulting in $Q_{\text{comp}} = 0.34$, however, such partition is not attributed to party membership of politicians, with $Q_{\text{party}} = -0.009$. This suggests that `comments` group politicians around discussion topics, motivating our investigation of the origins of polarization in the layer of `comments` in Section 2.5.

2.4.3 Group Similarity across Layers

The similar values of Q_{party} and Q_{comp} for the supports layer suggest that the partition of politicians into parties might be very similar to the results of community detection algorithms, while the different values for comments suggest the opposite. To empirically test this hypothesis, we compare group and party labels among politicians in each layer and across layers through the NMI at the group level. Within each layer, the NMI tells whether group labels discovered algorithmically can be predicted via party labels and vice versa. To do this, we follow a similar methodology as in Section 2.4.1. We compute the entropies of a layer based on detected groups and based on party labels, then we calculate

X Y	parties supports	parties likes	parties comments	supports likes	supports comments	comments likes
NMI($Y\|X$)	90.04%	3.99%	3.29%	3.83%	3.4%	11.77%
NMI($X\|Y$)	84.51%	3.87%	4.4%	4.12%	4.5%	8.56%

Table 2.4: Normalized mutual information of the group labels computationally found for each layer, and the party label. Groups of `supports` are very similar to parties, but the rest has low mutual information.

the mutual information between the different partitions, and finally obtain the NMI scores with respect to algorithmic and party partition. In this application of the NMI, random variables are the group labels of each node in each layer and party affiliation.

The results in Table 2.4 show that `supports` groups and party labels mostly match, $84-90\%$, contrary to `likes` and `comments`. Across layers, only `comments` and `likes` show a weak similarity in group labels $8-11\%$ and nearly no similarity to the `supports` layer. This result allows us to confirm that polarization in `supports` is due to party alignment of politicians, which does not hold for `likes` and `comments`. The high modularity score in the `likes` layer and the low information overlap with the `comments` layer tells us that the `like` signal has a dichotomous role as a social link. In certain situations, a `like` to a post signals a party affiliation; in other scenarios, it also shows politicians favour posts not only based on a party membership but due to the content of the post.

2.5 Origins of Network Polarization

2.5.1 Topic Analysis of Comment Groups

The above results show the existence of a partition of politicians in the `comment` layer that conveys certain polarization ($Q_{\text{comp}} = 0.336$), but which does not match to their party alignment. This suggests that `comments` happen more often across parties, partitioning politicians into groups by some other property besides party affiliation. To understand the reasons for such partition, we investigate the content of the `comments` between the politicians of each group. 10 of the 15 groups are very small – they contain reduced sets of politicians that exchange very few `comments` and are isolated from the rest. The 5 largest groups, which cover most of the politicians, have sufficient `comments` to allow us an analysis of their words. For each group, we computed a vector of word frequencies, ignoring German stopwords.[8] To measure the extent to which a word is characteristic for a group of politicians, we compute the *Pointwise Mutual Information* (PMI) of the frequency of the word in the group, compared to the frequency of the word in the set of all `comments`, discussed in Section 2.3. We quantify the significance of the PMI through a

[8] `http://solariz.de/649/deutsche-stopwords.htm`

	C1	C2	C3	C4	C5
1	ZauggGraf	SA	Noser	Fatah	Murmeln
2	Standardsprache	Hilton	Mängel	Abbas	wolf
3	GfSt	Dragovic	Noten	PA	unsre
4	Fitze	Botschaftsasyl	PK	pal	Lei
5	Digitalpolitik	Hollande	Raucher	Hamas	1291
6	Berufsbildungsfonds	Spahr	Weiterbildung	Libanon	Dokument
7	Bahnpolizei	Jungsozialisten	Asylsuchenden	Gaza	Wolf
8	Kamera	P21	WidmerSchlumpf	Jordanien	Atommüll
9	Hannes	Affentranger	Pensionskassen	Westbank	einwenig
10	Jeanneret	fur	Liberalisierung	palästinensischer	Gripen
Words	356	207	28	129	30
Comm.	10,145	11,312	452	899	656
Users	436	273	64	61	51

Table 2.5: Top words for each of the 5 largest `comments` groups, ordered by PMI. Number of words with PMI significant at the 99% confidence level, and amounts of users and comments in the group. Word colors classify them as follows: purple for Swiss Politician names, red for economic terms, blue for terms related to immigration, gray for words related to the conflict in the Middle East, green for terms related to security issues, and black for the rest. The top words significantly differ across groups, indicating that the partition in `comments` is aligned on the topics of discussion.

ratio test, only selecting words with $p < 0.01$, producing the word lists reported in Table 2.5.

All five groups have words with significant PMI, showing that they can be differentiated from other groups with respect to the words that politicians used in their `comments`. This difference highlights the topics discussed in every group, showing that the group structure in the `comments` layer is driven by topical interests. While some straightforward topics can be observed, especially for C4, we refrain from interpreting the terms of Table 2.5 within the Swiss politics context. These results show that modularity in the `comments` layer is topic-driven, and not party-driven, grouping politicians along their interests and competences.

2.5.2 The Temporal Component of Polarization

Our approach to polarization allows its measurement through behavioral traces, which lets us create real-time estimates with high resolution. The `politnetz.ch` dataset provides timing information for the creation of each `support`, `like`, and `comment`, giving us the opportunity to study the evolution of polarization through time. Including a time component in our analysis has the potential to reveal periods with higher and lower polarization, allowing us to detect potential sources that create polarization. We construct a set of time series of polarization along party lines $Q_{party}(t)$ in all three layers with a sliding window of two months.

Figure 2.4 shows the time series of network polarization along party lines for the three layers of the network. The `comments` layer shows negligible levels of polarization, fluctuating

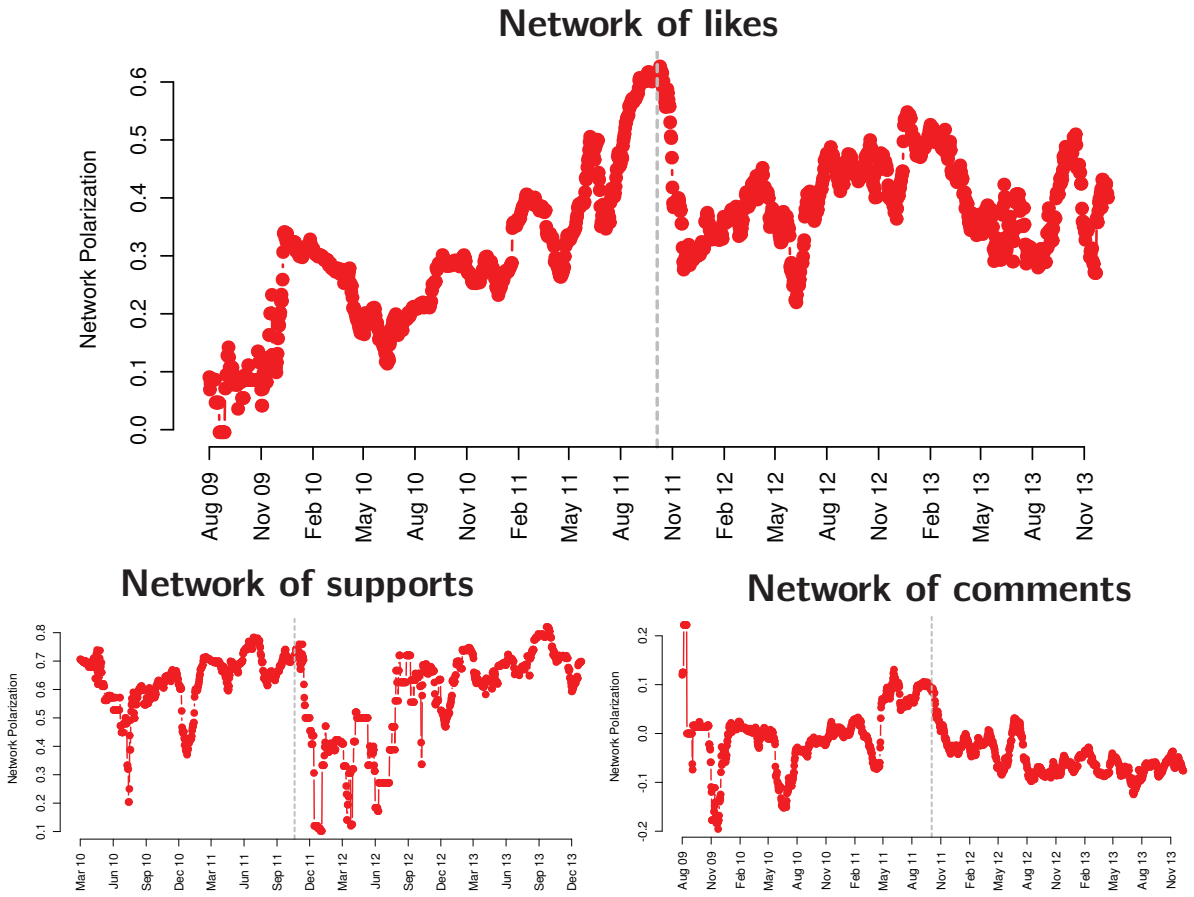


Figure 2.4: Time series of Q-modularity by party labels of the layers of `supports` (left), `likes` (top) and `comments` (right). The grey dashed line denotes Swiss federal parliament election in 2011. Two trends can be observed: polarization among politicians peaks at pre-election time in the network of `likes`, and post-election time is characterized by lower levels of polarization in `likes` and `comments`.

around 0 for the whole time period. Polarization in `likes` shows an increasing pattern up to late 2011, reaching a value above 0.6 shortly before the Swiss federal parliament election in 2011. Right after the election, polarization in `likes` strongly corrects to levels slightly above 0.2. Similarly, polarization in `supports` has relatively stable values around 0.6 before the federal election, dropping to values below 0.3 right afterwards. This analysis shows that network polarization, as portrayed in the online activity of politicians, is not a stable property of a political system, revealing that politically relevant events bias politician behavior in two ways: polarization reaches maxima during campaign periods, and polarization levels relax quickly after elections.

The pattern in `likes` suggests that politicians avoid awarding a `like` to a politician of another party when elections are close, but in other periods they display a less polarized pattern that allows `likes` across parties. On the other hand, polarization in `supports` stays generally high in the whole period of analysis but after the election, which suggests that some election winners might concentrate support, lowering party polarization along

coalition structures. These observations, while sensitive to the size and activity in the network, appear in contrast with the control scenario of `comments`, in which no artificial pattern of party polarization appears in the whole period.

2.6 Party Structures

In this section, we focus in two aspects of online political activity that are not captured by polarization metrics, *i.e.* intra-party structures and inter-party connectivity. In particular, we want to answer two questions: Do parties with different ideologies create different social structures in online communities? And do parties with similar positions in political space connect more to each other, despite the general pattern of polarization?

2.6.1 Intra-party Structures

Previous research on the online political activity of users in the US discovered that right-leaning users showed higher online social cohesion than left-leaning users (Conover *et al.*, 2012). In the following, we extend the quantification of each party both in its social structure and position in ideological space. With respect to the latter, we locate the ideology of each party in the two-dimensional space of Left-Right and Conservative-Liberal dimensions. To ensure the statistical relevance of our metrics, we restrict our analysis to the 6 parties with more than 200 politicians each, which cover the two-dimensional spectrum of Swiss politics.

We analyze the social structure of each party based on the `supports` subnetwork among the politicians of the party, capturing the asymmetric relationships that lead to prestige and popularity. On each of these party subnetworks, we quantify three metrics related to relevant properties of online social networks: hierarchical structures through in-degree centralization, information efficiency through average path length, and social resilience through maximum k-core numbers. While these three metrics are not independent from each other, they capture three components of online political activity that potentially differentiate parties: how popular their leaders are in comparison to a more egalitarian structure, how efficient their social structure is for transmitting information, and how big is the core of densely connected politicians who would support each other under adverse conditions.

Figure 2.5 shows the value of the three social network metrics versus the position of each party in both dimensions. The two parties in the farthest right part of the spectrum, SVP and FDP, show higher average path lengths and lower maximum k-core numbers than left-wing parties such as SP and Grüne. This points to a difference in online activity in Switzerland dependent on the the political position of parties: right parties have created

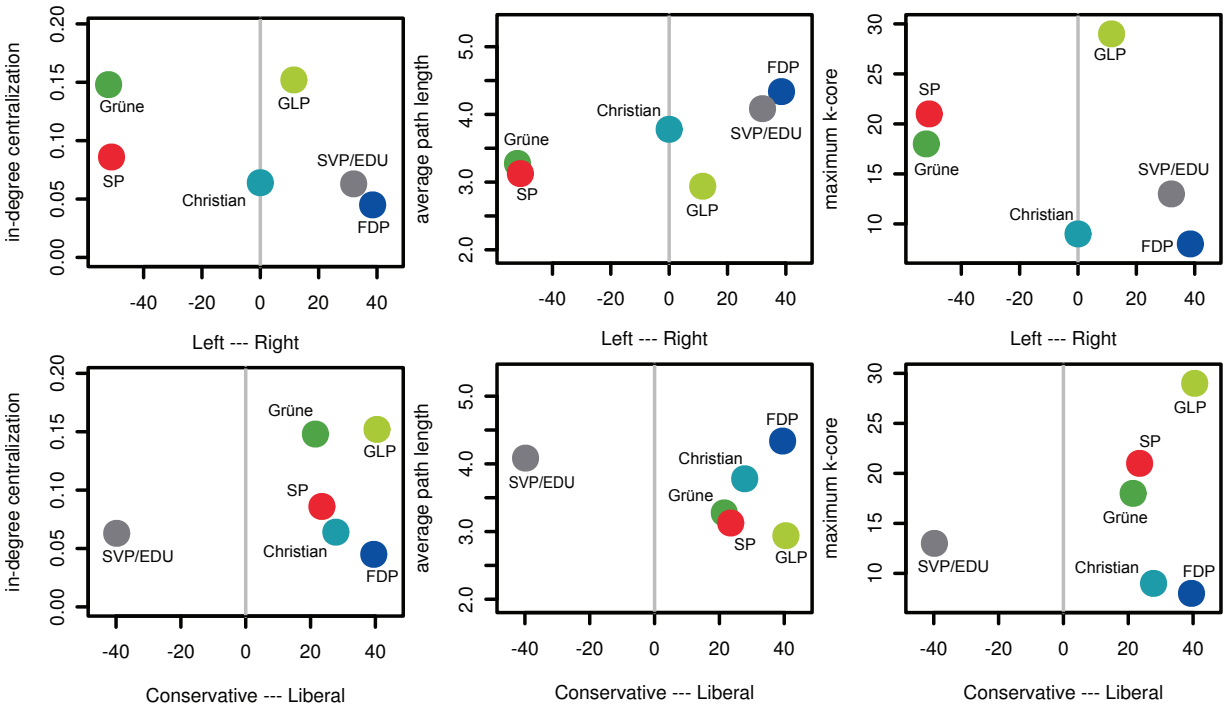


Figure 2.5: Social network metrics versus their position in ideological space, *Left-Right* position (top), *Conservative-Liberal* position (bottom). The `supports` subnetworks of parties with federal representation and more than 200 politicians each in `politnetz.ch`.

online social networks with lower information efficiency and lower social resilience. This poses a contrast with previous findings for US politics, in which politically aligned communities displayed the opposite pattern. This leads to the conclusion that the position of an online community in the left-right political spectrum does not universally define the properties of its social structure, and that particularities of each political system create different patterns.

There is no clear pattern of in-degree centralization in the *Left-Right* or *Liberal-Conservative* dimensions, but green parties (GLP and Grüne) show a significantly higher in-degree centralization than the rest. This result can be explained by the structure of the Swiss government, which is composed of seven politicians from a coalition of various parties. At the time of the study, this coalition includes politicians from the other four parties, excluding both GLP and Grüne. This would give an incentive to these parties to highlight a relevant member in order to gain a seat in the seven-member government, and thus creating higher in-degree centralization in online media.

2.6.2 Inter-party Connectivity

Our second question tackles the activity across parties with respect to their political position, hypothesizing that parties closer in ideological space will be more likely to connect

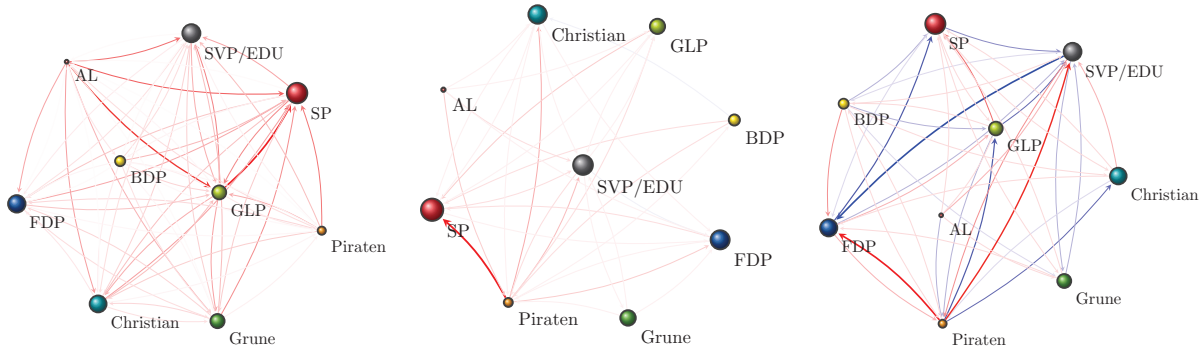


Figure 2.6: Networks of demodularity across parties for `supports` (left), `likes` (centre), and `comments` (right). Link width and color intensity is proportional to demodularity score between parties, colored blue for positive values and red for negative ones. Node size is proportional to the amount of politicians of each party. Demodularity scores are strongly negative for supports, weakly negative for likes, and both positive and negative for comments.

to each other. To do so, we first require a measure to estimate the tendency of one party to connect to another under the presence of polarization. To complement our analysis of Q-modularity, we use demodularity between pairs of parties, comparing their tendency to connect with what could be expected from a random network. Negative demodularity indicates that a party consistently avoids interaction with another party, contrary to positive demodularity, which indicates that interactions are more frequent than expected at random.

Figure 2.6 shows a visualization of the network of politicians aggregated by parties, for each layer of interaction. It can be observed that demodularity scores have strong negative values in the `supports` layer, where all links have a negative weight, see the values in Table 2.6. This is in line with the strong polarization among party lines, which makes `support` links to stay within parties. In addition, the overall level of outgoing negative demodularity shows a certain level of heterogeneity, as some parties have a stronger tendency to not support any politician from another party. In the `likes` layer, demodularity scores are less negative, indicating that there are weaker incentives to avoid giving a `like` to a politician of another party, compared to `supports`. Positive scores are only present in the `comments` layer, which shows that politicians are more likely to comment posts of other parties, creating debates with opponents.

To relate the demodularity with the position in ideological space of parties, we measured the Euclidean distance between pairs of parties in a two-dimensional ideological space. Our hypothesis is that in layers of interaction with positive connotation (`supports` and `likes`), parties that are further from each other have weaker cross-party interactions, in contrast to the `comments` layer, where links can be used to express disagreement and thus political distance increases with cross-party interaction. To test these hypotheses, we

Supports	AL	BDP	Christian	FDP	GLP	Grune	Piraten	SP	SVP
AL	-	-0.012	-0.046	-0.043	-0.099	-0.036	-0.016	-0.077	-0.07
BDP	-0.002	-	-0.01	-0.006	-0.021	-0.014	-0.004	-0.034	-0.016
Christian	-0.002	-0.003	-	-0.012	-0.031	-0.014	-0.006	-0.046	-0.021
FDP	-0.002	-0.003	-0.013	-	-0.028	-0.016	-0.005	-0.042	-0.014
GLP	-0.005	-0.009	-0.039	-0.033	-	-0.045	-0.011	-0.125	-0.06
Grune	-0.003	-0.005	-0.019	-0.02	-0.044	-	-0.008	-0.055	-0.032
Piraten	-0.002	-0.005	-0.023	-0.019	-0.039	-0.024	-	-0.066	-0.031
SP	-0.003	-0.007	-0.03	-0.027	-0.058	-0.024	-0.01	-	-0.043
SVP	-0.003	-0.003	-0.015	-0.01	-0.038	-0.022	-0.007	-0.057	-

Likes	AL	BDP	Christian	FDP	GLP	Grune	Piraten	SP	SVP
AL	-	0	-0.001	-0.001	-0.001	0	-0.003	0	-0.002
BDP	0	-	0.001	0	0	0	-0.002	-0.002	0
Christian	0	0	-	-0.001	0	0	-0.001	-0.001	-0.001
FDP	0	0	0	-	0	0	-0.001	-0.002	0
GLP	0	0	0	-0.001	-	0	0	-0.003	-0.001
Grune	0	0	0	-0.001	0	-	-0.001	0	-0.001
Piraten	0	-0.001	-0.005	-0.003	-0.002	-0.003	-	-0.016	-0.004
SP	0	0	-0.001	-0.001	-0.001	0	-0.003	-	-0.002
SVP	0	0	0	0.001	-0.001	-0.001	-0.002	-0.003	-

Comments	AL	BDP	Christian	FDP	GLP	Grune	Piraten	SP	SVP
AL	-	0	0	0	-0.001	-0.001	0.002	-0.001	-0.002
BDP	0	-	-0.001	-0.003	0.002	0.001	-0.001	0.002	-0.001
Christian	0	0	-	-0.001	0	0	0.001	-0.001	-0.002
FDP	0	0	-0.002	-	-0.002	0	-0.002	0.005	0.002
GLP	0	-0.001	0	-0.005	-	0	0.002	-0.004	0.004
Grune	0	0	0	-0.001	0	-	0	-0.001	0.002
Piraten	0	-0.001	0.004	-0.007	0.005	-0.003	-	0	-0.006
SP	0	0	-0.001	0.001	-0.001	-0.001	0	-	0.003
SVP	0	0.001	0	0.007	0	0.002	-0.003	0	-

Table 2.6: Demodularity scores of the network layers. `Supports` layer: Scores in the range $(-0.1, -0.05]$ are highlighted in light red colour, and scores in the range $(-\infty, -0.1]$ are displayed in red. `Comments` layer: Positive scores in the range $(0, 0.005)$ are highlighted in light blue colour, and scores in the range $[0.005, \infty)$ are displayed in blue.

	Supports	Likes	Comments
r_{pearson}	-0.14 ($p = 0.37$)	**-0.45** ($p < 3 \cdot 10^{-3}$)	**0.45** ($p < 2.7 \cdot 10^{-3}$)

Table 2.7: Pearson correlation coefficients of the demodularity scores between parties and their pairwise Euclidean distance.

compute the Pearson correlation coefficient of the demodularity score of the pairs of parties versus the political Euclidean distance between them. This correlation coefficient allows us to empirically test the existence of a linear relationship between ideological distance and cross-party interaction in each layer.

Table 2.7 shows the results for the three correlation coefficients. First, the `supports` layer shows no significant correlation, as data is not sufficient to reject the null hypothesis that both variables are not related. This points to the high level of polarization in `supports`, which makes links across parties so scarce that demodularity scores are equally negative for parties close and far in ideological space. The layer of `likes` shows a significant negative correlation, indicating that demodularity is higher between parties with closer ideologies. On the other hand, the correlation in the `comments` layer is positive, showing that politicians are more likely to participate in online debates with politicians of parties that hold opposite views. This highlights the meaning of `comments`, which are mainly used to argue with politicians of other parties as opposed to `likes`, which are used to agree with politicians with similar views, even though they might not belong to the same party.

2.7 Discussion

Our work explores behavioral aspects of political polarization, measuring network polarization over the digital traces left by politicians in `politnetz.ch`. Our approach is centered around the construction of a multiplex network with politicians as nodes and three layers of directed links: one with `support` links, a second one with link weights as the amount of `comments` a politician made to another politician, and a third one with weights counting the amount of times a politician `liked` the posts of another. We studied network polarization in the three layers as the level of intra-party cohesion with respect to inter-party cohesion, measuring network polarization as the modularity with respect to party labels, using Newman's Q-modularity metric. This methodology allows us to investigate polarization at a scale and resolution not achievable by traditional opinion survey methods, including the time evolution of polarization on a daily basis. Furthermore, we provide a quantitative analysis of the ways in which politicians utilize participatory media, the content of discussion groups between politicians of different parties, and the conditions that increase and decrease online network polarization.

We compared the information shared across the three layers, and found that each layer contains a significant amount of link and community information that is not included in any other layer. The layers of `supports` and `likes` revealed significant patterns of polarization with respect to party alignment, unlike the `comments` layer, which has negligible polarization. This is particularly interesting with respect to opinion dynamics models, which frequently assume that the presence of opinion polarization implies a bias in the underlying communication network. While polarization is clearly present on politnetz with respect to like and support links, it does not seem to have decisive influence on overal communication. We applied community detection algorithms at all three layers, and compared the computationally found groups with the parties of the Swiss system. At the `comments` layer, the community detection algorithms reveal that a partitioning

of politicians conveys higher modularity than party alignment, suggesting that groups in `comments` are not party-driven. This is confirmed by a content analysis of the `comments` in each group, where the most informative terms show that each group discusses different topics. At the `supports` layer, the partition of politicians into parties is very similar to the maximal partition found by algorithms, suggesting that party alignment is nearly the most polarizing partition of politicians. Further analysis will be necessary to test this observation, measuring how the party alignment of a politician might be predicted by its social context. In addition, our work focuses on data from politicians, constituting an analysis of elite polarization. Future works can include datasets from the electorate at large, measuring mass polarization from other digital traces such as blogs (Adamic and Glance, 2005) and `Twitter` (Conover *et al.*, 2011a; Lietz *et al.*, 2014).

We computed the time series of network polarization of each layer, revealing how the polarization in `likes` increased significantly around the federal elections of 2011, compared to moments without electoral campaigns. Polarization in `supports` and `likes` showed a sharp decrease after the elections, possibly as an effect of coalition-building processes, and polarization in `comments` was close to 0 for the whole study period. The evolution of polarization in `likes` and `supports` reveals a relation between online activity and political events, in which social interaction becomes more influenced by party membership when elections are close. Our approach to the time evolution of polarization can be applied to test the influence of other political events, for example how referendums might increase polarization along the opinions about the issue being voted.

The `comments` layer showed no party-alignment polarization, but the computational detection of modular structures reveals a different partition of politicians. Our analysis of the content of the `comments` in these partitions reveals that they follow different topics, and that modularity in `comments` can be attributed to similar interests and competences among politicians. This is emphasized when analyzing demodularity across parties and ideological distance: demodularity in `comments` increases with ideological distance, showing that parties that are further from each other tend to create debates in which politicians of different views `comment` on each other's contributions. Furthermore, demodularity in `likes` decreases with distance between parties, showing that the missing polarization in `likes` is due to politicians acknowledging the contributions of people from other parties with similar ideologies. At the level of individual politicians, this kind of data analysis will allow testing the relation between individual reputation and contribution to polarization, quantifying if strong biases in online behavior are tied to intra-party leadership and election results.

We analyzed the social networks of politicians in `politnetz.ch` to explore the relation between ideology and social structures in online interaction. We found that green parties (GLP and Grüne) have a higher in-degree centralization than the rest, indicating that their internal structure is more unequal with respect to popularity. Two possible expla-

nations for this are the current configuration of the Swiss government, which excludes these two parties, or a hypothetical relation between environmental politics and party organization. Left-aligned parties have lower average path length and higher maximum k-core numbers than right-aligned ones, showing that left parties create social networks with higher information spreading capabilities than right parties. This result is in contrast with previous findings (Conover *et al.*, 2011a) which found the opposite pattern for the networks of US `Twitter` users. In addition, demodularity metrics across parties indicate that their connectivity in terms of `likes` increases with closeness in political space, while the connectivity in terms of `comments` increases with distance. These findings lead to a set of hypotheses to test in other multi-party systems, in order to understand the conditions that link online social network structures and the political position of parties. The digital traces left by politicians in Canada (Gruzd and Roy, 2014), Germany (Lietz *et al.*, 2014), and Spain (Aragón *et al.*, 2013) are first potential candidates for this kind of analyses.

Our results highlight that the strategies for campaigning and mobilization in online media differ with respect to the political position of a party, and that polarization in online behavior is heavily influenced by party membership and upcoming elections. Our findings show that the previously found relations between social network structures and ideology are not universal, calling for new theories that apply for multi-party systems. We showed that the analysis of network polarization through digital traces conveys an alternative approach to traditional survey methods, capturing elements of the time dependence and the phenomena that influence polarization. The multi-party nature of Switzerland and the crowdsourced online activity of its politicians allowed us to test the relation between the ideological distance and online interaction between parties, showing us a very clear picture of online interaction. First, `supports` are strongly biased towards politicians of the same party. Second, `comments` happen across parties of different political views, clustering politicians into groups with similar interests. And third, `likes` cross party lines towards politicians of similar opinions, but this effect is attenuated when elections are close, creating a highly polarized state that relaxes after elections are over.

Chapter 3

Emotional Reactions in Online Communities

Summary

We test the predictions of one of the most influential psychological theories of emotions, Cacioppos evaluative space model, (Cacioppo and Berntson, 1994; Cacioppo and Gardner, 1999; Cacioppo *et al.*, 1997), in a real-life social setting. We do this by analyzing a large number of messages and replies from three popular online communities (16-24 millions each). We examine a) the bivariate valence (*i.e.* the negativity and positivity) of the original messages and the replies, using the state-of-the-art sentiment detection tool, b) whether a given message was replied to at all, and c) the interevent time between original messages and replies. As predicted, the negative content of a message has a higher influence on the likelihood of the message being answered than its positive content. Surprisingly, the highest likelihood for response is observed for messages with both high positive *and* high negative valence. Contrary to our expectations, however, we do not observe significant differences between the interevent time distributions of negative and positive messages. Thus, the time scale of online replies seems to be largely determined by content-independent factors. Thanks to the scale and granularity of this approach, we reach tails of longitudinal and unlikely behavior that are nearly impossible to elicit in experimental setups or to sample in survey studies. In this way, we analyze the complete picture of reaction tendencies to emotional expressions, including naturally evoked social sharing of both positive and negative emotions.

Based on a working paper "Response patterns to emotional expression in online interaction" by Adiya Abisheva, David Garcia, and Frank Schweitzer, (2016). A.A. is the main responsible for the design and implementation of the statistical analyses, plots and the manuscript.

3.1 Introduction

In this Chapter, we present our findings on the theory of the negativity bias and the theory of social sharing of emotions, discussed thoroughly in Section 1.2. As part of the research questions, presented in Section 1.3.2, we also test the Cacioppo's hypothesis that positive and negative stimuli result in different human behavioral reactions, see Hypothesis 2, Section 1.3.2.

In human communication, emotions play a twofold role: emotional language can modify the impact of a message on the receiver, who in turn might reply with emotional language. In addition to influencing the emotional content of a reply, the emotions in an original message might also influence how fast this message receives reply and whether it receives a reply at all. Based on the principles of the evaluative space model (Section 1.2), we can make the following predictions: Due to the negativity bias (Section 1.2) negative messages should evoke stronger and faster reactions than positive ones, which in turn should have a stronger effect than neutral messages. The prediction for ambivalence is more difficult; on simple behavioral dimensions (*e.g.* approach vs. avoidance), ambivalence should lead to "stalemate", *i.e.* neither behavior is expressed. In human communciation however, a reply, just like the original message, can contain both positive and negative emotions. Thus, we would predict that replies to messages with mixed emotional content should have high levels of mixed emotional content themselves. However, no clear predictions can be made on the speed and likelihood of response on the basis of evaluative space theory.

We test these hypotheses using the state-of-the-art sentiment analysis tool `SentiStrength` (Abbasi *et al.*, 2014; Thelwall *et al.*, 2010a), which enables us to quantify the emotions expressed through text on separate positivity and negativity scales, both for original messages and replies. In addition, we measure the likelihood of a message receiving at least one reply, the number of replies, and the waiting time between an original message and a reply.

3.2 Dataset Description

Data used in this research is a result of our 1-4 year crawl of three popular public online communities.

`YouTube` (`http://www.youtube.com`) is a video-sharing website on which registered users can upload and view videos, post comments on videos, and reply to other users' comments. Our daily crawl, which was launched in June 2011, collected the comments of the top 100 videos in four categories: most discussed, most popular, top favorites and top rated videos of the day. By September 2014, the number of collected comments amounted to 128.6 millions, out of which 23.8 millions were reply comments to other comments.

`4chan` (`http://www.4chan.org`) is an imageboard website where users create and participate in the discussion anonymously. The site does not require registration. Conversations

	Dataset	# Text messages	# Responses
4chan	all	43,642,476	16,370,788
	Miscellaneous /b/	25,147,940	8,577,131
	Rest boards	18,494,536	7,793,644
YouTube	all	128,629,381	23,897,370
	Entertainment	20,232,452	3,293,755
	Gaming	23,883,666	4,468,468
	Music	13,145,421	2,046,590
	News & Politics	12,240,567	4,200,138
Reddit	all	51,378,747	24,908,544
	Funny subreddit	6,928,832	3,155,326
	Rest subbreddits	44,449,912	21,504,712

Table 3.1: Statistics of the text messages and responses in English language. Number of text messages and responses in total and per category across three datasets.

between users appear in one of the many thematic boards, such as politics, technology, science fiction, adult board, or the majority of discussions happen in the miscellaneous board called /b/ with no content restriction. Our daily crawl collected 43.6 millions user posts between 2012 and 2013, out of which 16.3 millions were the reply posts of users to posts by other users.

Reddit (http://www.reddit.com) is a message board website where registered users submit text posts, vote up and down for entries and comment to posts, and therefore creating discussions. Similar to 4chan, the topics of discussions are organized around so called subreddits. Analogous to random board /b/ on 4chan, subreddit *Funny* on Reddit contains discussions on miscellaneous topics. Our daily Reddit crawl collected 51.3 millions user posts from 2012 to 2014, out of which 23.9 millions were reply posts to other users' text submissions. Detailed descriptive statistics on the number of posts in each discussion board is shown in Table 3.1. Furthermore, original crawl in the dataset contained messages of different languages given the public nature of the online platforms and their exposure to the large audience speaking different languages. We applied a language classification analysis (Nakatani, 2010–current) for each message to select only messages in *English* language.

3.3 Response Time and Emotional Expression

3.3.1 Response Pairs and Waiting Time

We study a collection of text responses across three social network datasets, YouTube, 4chan and Reddit with 23.8 millions, 16.3 millions and 23.9 millions response pairs respectively. Each dataset covers various topics as diverse as music, politics, random, controversial, computer related, which ensures that our analysis spans a wide range of categories

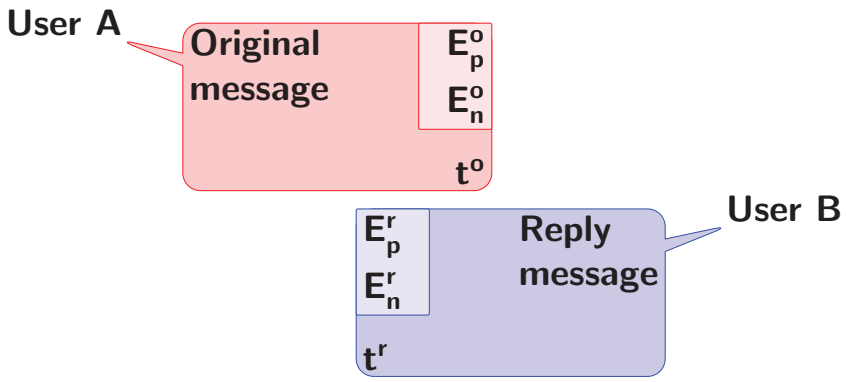


Figure 3.1: Communication activity in the online community. Example of a *reactive interaction*: User B interacts with user A by creating at time t^r a reply message to the message of user A, which was written at time t^o. Both messages express positive E_p and negative E_n emotions.

(Garas *et al.*, 2012). The three online platforms are characterized by the public nature of their messages, the potentially large audience to whom the messages are broadcasted, and the persistence of the messages in the history of the platforms. In contrast to online chats, user interaction is not necessarily synchronous. Such online platforms, together with blogs and forums, are commonly known as slow persistent interactive communication channels (Garas *et al.*, 2012).

Each response pair item in our dataset consists of an original message and a reply message (Figure 3.1). At time t^o, user A generates a text message (an original message), in which positive E^o_p and negative emotions E^o_n are expressed. Later, at time $t^r > t^o$, user B, upon reading the message by user A, *might* create a reply message, which also carries an emotional content E^r_p and E^r_n. In our study, we want to understand the effect of the emotions of the original message E^o and that of the reply message E^r on the interaction between the user A and the user B. One way to measure the effect of emotions is by computing the speed of the reply message dependent on the emotions expressed in the original message. Response or waiting time is defined as the interval between receiving and replying to a message (Wu *et al.*, 2010), $\Delta t = t^r - t^o$. For each response pair, we compute Δt conditional to the emotional valence of the original message and assess the statistical differences between the distributions of the response times.

To study this, for each response pair (Figure 3.1) we measure the waiting time Δt – the time interval between the creation of a message and receiving a reply (Wu *et al.*, 2010). In communication theory literature, such a response pair exchange is known as a *reactive* interaction (Rafaeli and Sudweeks, 1997), which is a two-way communication where a message is a reaction to a preceding one. It has been shown that reactive messages are the most frequent type of interaction between users in discussion groups in early online forums, such as Usenet, Bitnet Listservs, and Compuserve SIGs, as compared to interactive messages defined as the sequences of responses on the same topic of discussion (Rafaeli

and Sudweeks, 1997). This is one of the reasons of our focus on collective response pair items. Another reason is that we adopt an *ergodic* approach (Crane *et al.*, 2010), which "assumes that sampling collective responses of many individuals in time is equivalent to sampling many realizations of the same stochastic process" (Crane *et al.*, 2010). Response pairs in our study are generated by different individuals, as compared to previous works where interevent or waiting time analysis was done on the messages exchanged between two individuals only. Collective nature of responses from different users allows us to study their *collective temporal behaviour* (Garas *et al.*, 2012).

The collective response activity of users can be studied by analyzing the waiting time distribution $P(\Delta t)$ (Barabasi, 2005; Garas *et al.*, 2012; Malmgren *et al.*, 2008; Oliveira and Barabási, 2005; Wu *et al.*, 2010). In our work we go one step further and analyze the waiting time distribution conditional to emotional expression E of the original or the reply message $P(\Delta t|E)$. This way we characterize the temporal interaction between users dependent on the emotions expressed in the input (original message) or output (reply message) stimuli.

Classification of emotions To detect the emotions, we perform a sentiment analysis of each text message, the original and the reply, which is described in detail in the Introduction. This automatic classification returns both a positive E_p and a negative E_n valence for each message, *i.e.* two discrete values in the range of $[+1, +5]$ and $[-5, -1]$ respectively. A combination of two valence $E_{np} = (E_n, E_p)$ characterizes a two-dimensional emotional charge of a single text, which agrees with the theory of the evaluative bivalence (Norman *et al.*, 2011) presented in the Introduction. Such combination produces 25 possible discrete values of E_{np}: $[(-1, +1), (-1, +2), \ldots, (-5, +4), (-5, +5)]$. The two-dimensional classification allows us to distinguish between the different classes of emotions: positive valence E_p, negative valence E_n, bivalence E_{np} and emotional *arousal* E_a. Arousal represents the level of activation of emotion and is computed as an absolute difference between positive and negative sentiment scores $E_a = |E_p - E_n|$. Arousal value is a discrete number that falls in the interval of $[2, 10]$.

3.3.2 Results

Figure 3.2 shows the empirical accumulative distribution of the waiting time $(1 - F(\Delta t))$ conditional to positive and negative emotional valence of the original message and that of the reply message. Visually, we observe nearly identical curves across the different classes of emotions in each dataset: across negative emotional spectrum, across positive emotional spectrum, and inter-valence class (positive and negative emotions). Similarity in the time distribution indicates that the speed of reaction does not depend on the valence of emotional stimuli. The waiting time probability of original messages expressing extreme negative emotions of -5 or -4 is nearly identical to the waiting time probability of messages eliciting weak negative or no emotions, *e.g.* -2 or -1. There is also no

difference between the waiting time probability of messages expressing strong positive and those of expressing weak positive emotions. Similarly, we do not observe any temporal difference dependent on the expressed emotions among the reply messages, see insets of Figure 3.2.

To confirm the observed similarity of waiting time sets, we performed the pairwise Kolmogorov -Smirnov test (K-S test) of waiting distribution time sets of different emotions, *e.g.* the test between the sets of Δt of positive emotion +1 and +2, which we abbreviate as K-S($|(\Delta t|E_p^r = +1)|$,$|(\Delta t|E_p^r = +2)|$). The results of all pairwise K-S tests across all datasets are shown in Tables 3.2– 3.7. An important finding is that all tests output the p-values that are below 0.05 or close to 0. With this result, we have to reject the null hypothesis, and conclude that the waiting times are not sampled from the same distribution. However, we argue that given the large size of the compared data, it is almost impossible not to reject the null hypothesis. According to Lin *et al.* (2013), applying small-sample statistical tests, such as K-S test, to large-sample datasets ($\geq$ 10000 observations) almost always produces p-value close to zero, leading to falsely rejecting the null hypothesis. Differences that are small and non-notable in large size datasets are often found to be statistically significant, but this statistical significance does not tell anything about the *practical significance of such a small difference*. The implication of this statement is that when applying K-S test on large datasets like in our studies, relying solely on the p-value is not accurate, and therefore we must also consider the obtained D. Non zero D indeed reports that two datasets are statistically different. However, in case of large datasets the small difference can be considered negligible and not practically significant. The $D < 0.1$ is considered to be a small difference. In our pairwise K-S tests, the only difference that is larger than 0.1 is obtained between the waiting times of emotions +1 and +4 on `4chan`, $D = 0.16$. The rest 119 tests yield much smaller difference, $D \ll 0.1$. In light of these results and the controversy of p-value in K-S tests applied on large sample datasets, we suggest that there are not enough evidences to reject the null hypothesis that the waiting time distributions of different emotions are sampled from the same distribution.

To complete our analysis, we model a log-transformed interevent time a) as a linear function of the emotional expression of the original message and its message length (measured in the number of words), see Figure 3.3, and b) as a linear function of the emotions of the reply message and its message length, see Figure 3.3. All predictors in our models are normalized to [0..1] for comparison purposes. As expected, we obtain a positive correlation between the message length and the reply time: longer messages take longer time to reply. On emotional dimension, messages expressing negative emotions receive replies quicker than those expressing positivity. This might serve as an additional confirmation to the negativity bias. However, the difference between the coefficients of negativity and positivity is very small. Among all predictors, message length is a better predictor for the reply time as compared to the expressed emotions. The strength of the coefficients of

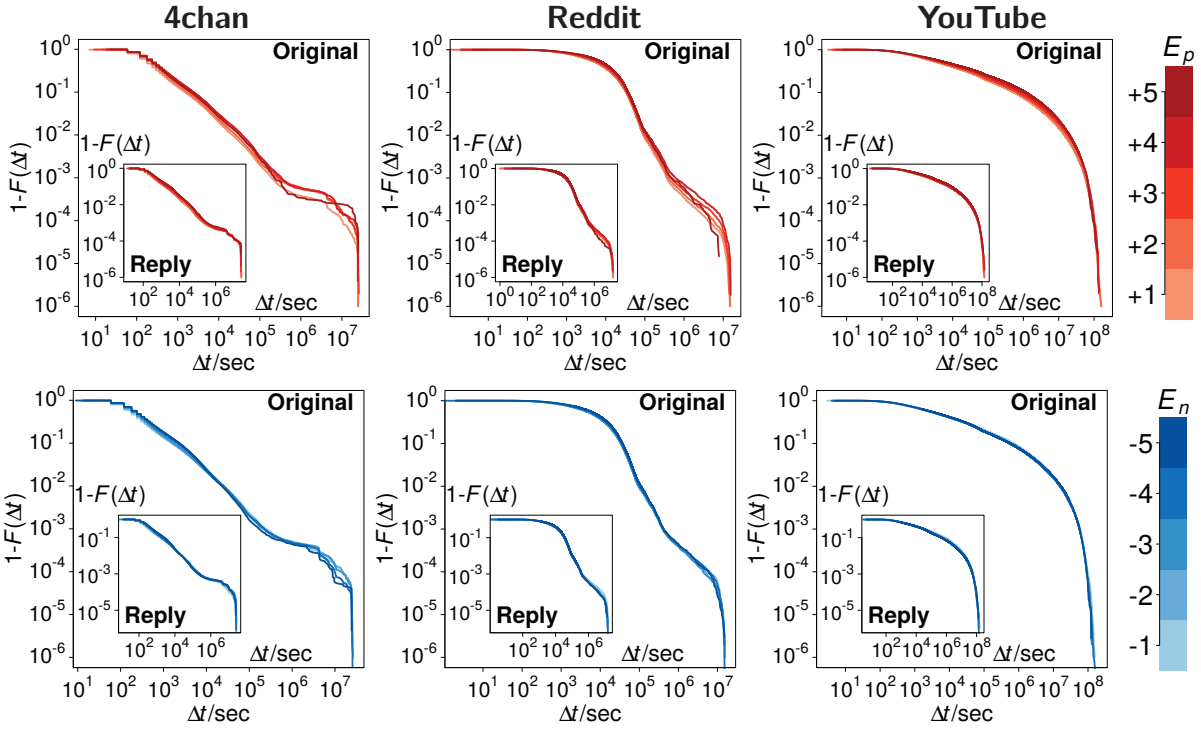


Figure 3.2: Communication activity in the online community. Empirical accumulative distribution of the waiting time of the user activity conditional to emotional valence expressed in original messages and in reply messages (inset) across three communities `4chan`, `Reddit` and `YouTube`: to positive emotions (red curves) and to negative emotions (blue curves). The activity is expressed as the time interval Δt between the timestamps of reply and original message. The time is measured in seconds.

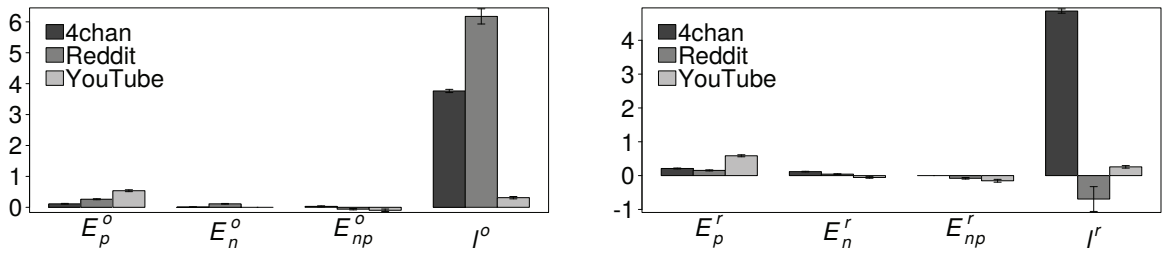


Figure 3.3: Barplots of regression coefficients and their confidence intervals obtained for modelling waiting time Δt as a function of emotional expression and length of: original message (left), $\ln(\Delta t) \sim E_p^o + E_n^o + E_{np}^o + \ell^o$; and of reply message (right), $\ln(\Delta t) \sim E_p^r + E_n^r + E_{np}^r + \ell^r$. Normalized predictor variables are: positive emotion E_p, negative emotion E_n, co-activation of both positive *and* negative emotions E_{np} and the length of a message measures in the number of words ℓ.

expressed emotions is approximately 18 times lower than the strength of the coefficient of the message length. This might be an explanation why we observed no difference between the distribution of waiting times dependent on the expressed emotions. We conclude based on the results of the model and the analysis of the distributions that the reply time to a message does not depend on the emotionality expressed in the message.

K-S test of waiting time distributions of different emotions: original message

	+1	+2	+3	+4	+5	E_p^o
-5		$D = 0.058$ $p = 0$	$D = 0.095$ $p = 0$	$D = 0.12$ $p = 0$	$D = 0.094$ $p = 0$	+1
-4	$D = 0.01$ $p = 0$		$D = 0.038$ $p = 0$	$D = 0.064$ $p = 0$	$D = 0.038$ $p = 0$	+2
-3	$D = 0.042$ $p = 0$	$D = 0.036$ $p = 0$		$D = 0.027$ $p = 0$	$D = 0.004$ $p = 0.51$	+3
-2	$D = 0.064$ $p = 0$	$D = 0.059$ $p = 0$	$D = 0.023$ $p = 0$		$D = 0.027$ $p = 0$	+4
-1	$D = 0.11$ $p = 0$	$D = 0.1$ $p = 0$	$D = 0.071$ $p = 0$	$D = 0.051$ $p = 0$		+5
E_n^o	-5	-4	-3	-2	-1	

Table 3.2: Pairwise K-S test comparing distributions of interevent time conditional to positive emotions E_p^o (red) and negative emotions E_n^o (blue) of original message on `4chan`.

	+1	+2	+3	+4	+5	E_p^o
-5		$D = 0.017$ $p = 0$	$D = 0.042$ $p = 0$	$D = 0.079$ $p = 0$	$D = 0.095$ $p = 0$	+1
-4	$D = 0.019$ $p = 0$		$D = 0.025$ $p = 0$	$D = 0.063$ $p = 0$	$D = 0.079$ $p = 0$	+2
-3	$D = 0.047$ $p = 0$	$D = 0.028$ $p = 0$		$D = 0.038$ $p = 0$	$D = 0.054$ $p = 0$	+3
-2	$D = 0.056$ $p = 0$	$D = 0.038$ $p = 0$	$D = 0.01$ $p = 0$		$D = 0.017$ $p = 2.3\times10^{-15}$	+4
-1	$D = 0.058$ $p = 0$	$D = 0.04$ $p = 0$	$D = 0.017$ $p = 0$	$D = 0.009$ $p = 0$		+5
E_n^o	-5	-4	-3	-2	-1	

Table 3.3: Pairwise K-S test comparing distributions of interevent time conditional to positive emotions E_p^o (red) and negative emotions E_n^o (blue) of original message on `Reddit`.

	+1	+2	+3	+4	+5	E_p^o
-5		$D = 0.024$ $p = 0$	$D = 0.036$ $p = 0$	$D = 0.058$ $p = 0$	$D = 0.067$ $p = 0$	+1
-4	$D = 0.002$ $p = 0.18$		$D = 0.012$ $p = 0$	$D = 0.037$ $p = 0$	$D = 0.05$ $p = 0$	+2
-3	$D = 0.009$ $p = 0$	$D = 0.008$ $p = 0$		$D = 0.027$ $p = 0$	$D = 0.042$ $p = 0$	+3
-2	$D = 0.018$ $p = 0$	$D = 0.016$ $p = 0$	$D = 0.009$ $p = 0$		$D = 0.016$ $p = 1.2\times10^{-11}$	+4
-1	$D = 0.026$ $p = 0$	$D = 0.027$ $p = 0$	$D = 0.021$ $p = 0$	$D = 0.014$ $p = 0$		+5
E_n^o	-5	-4	-3	-2	-1	

Table 3.4: Pairwise K-S test comparing distributions of interevent time conditional to positive emotions E_p^o (red) and negative emotions E_n^o (blue) of original message on `YouTube`.

K-S test of waiting time distributions of different emotions: reply message

	+1	+2	+3	+4	+5	E_p^r
-5		$D = 0.085$ $p = 0$	$D = 0.13$ $p = 0$	$D = 0.16$ $p = 0$	$D = 0.14$ $p = 0$	+1
-4	$D = 0.025$ $p = 0$		$D = 0.051$ $p = 0$	$D = 0.082$ $p = 0$	$D = 0.072$ $p = 0$	+2
-3	$D = 0.052$ $p = 0$	$D = 0.04$ $p = 0$		$D = 0.033$ $p = 0$	$D = 0.042$ $p = 0$	+3
-2	$D = 0.087$ $p = 0$	$D = 0.077$ $p = 0$	$D = 0.038$ $p = 0$		$D = 0.022$ $p = 0$	+4
-1	$D = 0.15$ $p = 0$	$D = 0.15$ $p = 0$	$D = 0.11$ $p = 0$	$D = 0.079$ $p = 0$		+5
E_n^r	-5	-4	-3	-2	-1	

Table 3.5: Pairwise K-S test comparing distributions of interevent time Δt conditional to positive emotions E_p^r (red) and negative emotions E_n^r (blue) of reply message on `4chan`.

	+1	+2	+3	+4	+5	E_p^r
-5		$D = 0.006$ $p = 0$	$D = 0.012$ $p = 0$	$D = 0.034$ $p = 0$	$D = 0.073$ $p = 0$	+1
-4	$D = 0.019$ $p = 0$		$D = 0.009$ $p = 0$	$D = 0.033$ $p = 0$	$D = 0.074$ $p = 0$	+2
-3	$D = 0.034$ $p = 0$	$D = 0.015$ $p = 0$		$D = 0.03$ $p = 0$	$D = 0.071$ $p = 0$	+3
-2	$D = 0.034$ $p = 0$	$D = 0.017$ $p = 0$	$D = 0.007$ $p = 0$		$D = 0.041$ $p = 0$	+4
-1	$D = 0.024$ $p = 0$	$D = 0.02$ $p = 0$	$D = 0.019$ $p = 0$	$D = 0.019$ $p = 0$		+5
E_n^r	-5	-4	-3	-2	-1	

Table 3.6: Pairwise K-S test comparing distributions of interevent time Δt conditional to positive emotions E_p^o (red) and negative emotions E_n^o (blue) of reply message on `Reddit`.

	+1	+2	+3	+4	+5	E_p^r
-5		$D = 0.028$ $p = 0$	$D = 0.043$ $p = 0$	$D = 0.065$ $p = 0$	$D = 0.065$ $p = 0$	+1
-4	$D = 0.01$ $p = 0$		$D = 0.015$ $p = 0$	$D = 0.038$ $p = 0$	$D = 0.039$ $p = 0$	+2
-3	$D = 0.012$ $p = 0$	$D = 0.009$ $p = 0$		$D = 0.025$ $p = 0$	$D = 0.035$ $p = 0$	+3
-2	$D = 0.018$ $p = 0$	$D = 0.019$ $p = 0$	$D = 0.01$ $p = 0$		$D = 0.012$ $p = 1.7 \times 10^{-5}$	+4
-1	$D = 0.034$ $p = 0$	$D = 0.039$ $p = 0$	$D = 0.031$ $p = 0$	$D = 0.021$ $p = 0$		+5
E_n^r	-5	-4	-3	-2	-1	

Table 3.7: Pairwise K-S test comparing distributions of interevent time Δt conditional to positive emotions E_p^o (red) and negative emotions E_n^o (blue) of reply message on `YouTube`.

Origins of the waiting time distributions To understand the process that generates the waiting time distributions *independent of sentiment* classes expressed in the messages, $P(\Delta t)$, we fit three known parametric distributions related to complex growth phenomena (Mitzenmacher, 2004): power law, log-normal and exponential distribution. The implementation of the fit is provided in the `powerlaw` python package (Alstott *et al.*, 2014). We compare the likelihood of each distribution using the log-likelihood ratio $R = \ln(\frac{L_1}{L_2})$ between the two candidate distributions and its significance value p. Positive ratios indicate the likelihood of data in the first distribution, and negative ratios for the second one. Instead of testing the hypothesis of the data following a certain distribution, this comparative test answers the question of which parametric distribution provides the best fit available, following the principle of Maximum Likelihood estimation (Alstott *et al.*, 2014). Practical reason for computing log-likelihoods instead of bootstraping the goodness of fit is the computational feasibility especially for a large-scale data as in our research. In Table 3.8, we display the comparison of the likelihoods of three distributions.

The log-likelihood comparison of three statistical distributions and the K-S statistics suggest that the log-normal distribution might be the "winning" fit to the waiting time distributions. However, we can easily observe in the cumulative density plots in Figure 3.4 that the log-normal fit (red dashed line) is far from accurate. Furthermore, we visually observe the "hump" starting at 3.5 days ($10^{5.5}$ sec) on `4chan` and `Reddit`, which suggests an evidence for the bimodality pattern (Wu *et al.*, 2010). This finding implies that the bursts of conversations on these platforms lasts for approximately 3.5 days, but this statement needs a further statistical confirmation.

In conclusion, we were not able to fit the distributions of waiting times to any of the existing statistical distributions. However, we observed the bimodality pattern on the two (semi-)anonymous communities `4chan` and `Reddit`, but not on `YouTube`. This requires a

	`4chan`		`Reddit`		`YouTube`	
	R	p-value	R	p-value	R	p-value
$\ln(\frac{L_{LN}}{L_{PL}})$	6750806	0	30291784.9	0	13513406	0
$\ln(\frac{L_{LN}}{L_{SEXP}})$	909715745	0	435399.4	0	4679804	0
$\ln(\frac{L_{LN}}{L_{EXP}})$	213225539	0	9779085.8	0	1321722526	0
$\ln(\frac{L_{PL}}{L_{SEXP}})$	902964939	0	-29856385.5	0	-8833602	0
$\ln(\frac{L_{PL}}{L_{EXP}})$	206474733	0	-20512699.1	0	1308209120	0
$\ln(\frac{L_{SEXP}}{L_{EXP}})$	-696490206	0	9343686.4	0	1317042722	0

Table 3.8: Comparison of fits of waiting times Δt to the families of distributions. The result of the comparison is the loglikelihood ratio R between the likelihoods of the two candidate distributions. Positive ratio indicates the likelihood L of the first distribution (nominator) given data, and negative ratio indicates the likelihood of the second distribution (denominator). p-value shows the significance of the normalised loglikelihood ratio.

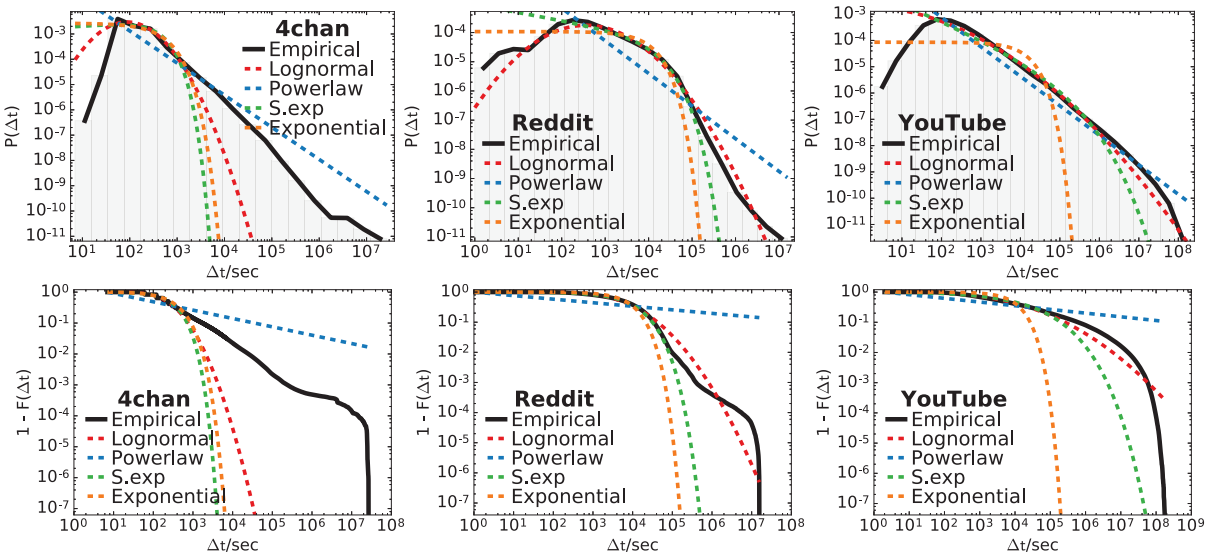


Figure 3.4: Probability density plots and cumulative density plots of the waiting times Δt across three communities and fits of the waiting times to the families of distributions.

further statistical validation, *e.g.* fitting data before and after the "hump" to the known distributions. Contrary to our expectation, the results of the temporal analysis indicate that on the Internet users do not assign higher priority to the response to negative stimuli, when the priority is measured as the time to reply to a message.

3.4 Response Likelihood and Emotional Expression

Internet user activity can be characterized using a temporal metric, such as interevent and waiting time distribution. In literature, however, there are other measures of Internet user behaviour which omit the time scale. For example, Chmiel *et al.* (2011b) expressed user activity as the total number of posts written by a user in a discussion board during the observation period. Pfitzner *et al.* (2012) has introduced a measure of information sharing behaviour and has defined it as a retweet likelihood conditional to the emotion of an observed tweet. We adopt this approach to characterize the priority that a user assigns to replying to online posts of different sentiment classes.

3.4.1 Evaluation Metrics

We introduce several metrics which quantify the deviations from the baseline property of the message in the presence of some message attribute. We measure the two baseline properties of a message, namely the probability of expressed emotions in a message $P(E)$ and the probability of a message to receiving a reply $P(O)$.

$P(E)$, the baseline probability of observing expressed emotion E in a message, is computed as a fraction of the total number of all messages,[1] which exhibited emotion E over the number of all messages in the dataset.

$P(O)$, the baseline probability of an original message O to receiving a reply, is calculated empirically as a fraction of the number of messages that received at least one reply to the number of all messages in each dataset.

To differentiate between the different types of messages, we introduce a notation scheme for the variables. An *original* message, or a message that receives a reply, is denoted by a subscript $_O$, while a *reply* message, which is a reaction to some original message, is denoted by a subscript $_R$.

Likelihood of emotions We introduce our first metric, namely the likelihood of emotions expressed in *messages receiving a reply*, also known as original messages, and in *reply messages*, denoted as $\alpha_O(E)$ and $\alpha_R(E)$ respectively. We compute $\alpha_O(E)$ as a conditional probability of emotional expression in an original message that receives at least one reply $P(E|O)$ normalized over the baseline probability of emotions $P(E)$. To compare effects, we take the natural logarithm of the ratio:

$$\alpha_O(E) = \ln\left(\frac{P(E|O)}{P(E)}\right). \tag{3.1}$$

Similarly to $\alpha_O(E)$, the likelihood of emotions expressed in *reply messages* $\alpha_R(E)$ is computed as a conditional probability of emotional expression in a reply message $P(E|R)$ normalized over the baseline probability of emotions $P(E)$:

$$\alpha_R(E) = \ln\left(\frac{P(E|R)}{P(E)}\right). \tag{3.2}$$

We compute the conditional probabilities from the empirical data. $P(E|O)$ is calculated as a fraction of the number of messages that receive a reply *and* express emotion E over the total number of replies in the dataset.[2] Similarly, $P(E|R)$ is computed as a fraction of the number of replies exhibiting emotion E over the total number of replies.

The emotional likelihood $\alpha(E)$ can take positive, negative values and zero and indicates how the emotional expression in replied and in reply messages deviates from the baseline emotional expression. We provide the interpretation of each case on the example of the

[1] All messages in the dataset include reply messages R, messages that receive replies O as well as those that never receive replies.

[2] An original or stimulus message might receive zero, one or more than one replies. For original messages that receive at least one reply, we assign a weight to the emotions expressed in those messages according to the number of received replies.

emotional likelihood of original messages. Positive values of $\alpha_O(E)$ indicate that emotional expression E is more likely to be observed in messages receiving a reply than in all messages in the dataset, while negative values indicate that this emotional expression is less likely to be elicited in the replied messages. A value of 0 points that there are no differences in the emotional expression of messages receiving a reply and the baseline emotional expression.

Likelihood to receiving replies We introduce our second metric, namely the likelihood of an original message to receiving a reply $\alpha_E(O)$, which measures the effect of emotions elicited in a message on the probability of such a message to receiving a reply. We compute $\alpha_E(O)$ as a conditional probability of an original message to receiving a reply given it's expressed emotion $P(O|E)$ normalized over the baseline probability of an original message to receiving a reply $P(O)$:

$$\alpha_E(O) = \ln\left(\frac{P(O|E)}{P(O)}\right). \tag{3.3}$$

The conditional probability $P(O|E)$ is computed from the empirical data, and is a fraction of the number of messages that receive a reply *and* express emotion E over the total number of all messages that express emotion E.

When the conditional and the baseline probability are equal, resulting in $\alpha_E(O)$ to be 0, this implies that emotions E expressed in the messages do not affect the probability of the message to receiving a reply. When emotions have a positive effect for an original message to receive replies, $\alpha_E(O)$ takes on positive values. On the contrary, $\alpha_E(O)$ is negative when emotions have a negative effect on obtaining responses.

To sum up, the likelihood of emotions $\alpha(E)$ expresses the differences between the emotional expression in the baseline emotions and that of in messages with replies and reply messages. And, the likelihood of reply to original messages $\alpha(O)$ quantifies the effect of emotions expressed in original messages to receiving a reply.

3.4.2 Results

Taking into account the metrics described earlier, we compute the normalised log-likelihood of emotional expression in messages that receive a reply (Eq. 3.1) and in reply messages (Eq. 3.2) for the two classes of emotions: bivalence E_{np} (Figure 3.5) and arousal E_a (Figure 3.6A).

Our discussion of results starts with the findings on the baseline emotional expression of *valence* in all messages. The *baseline* probability $P(E_{np})$ shown in Figure 3.5A reveals that the majority of messages elicit emotions, but out of the 25 discrete emotional values that we detect, the dominant emotion expressed in the messages across three datasets is $(-1, +1)$. Such neutral emotional expressions account to 35%, 30% and 24% on `4chan`, `Reddit` and `YouTube`. The probabilities of strictly positive $P(E_p|E_n = -1)$ and strictly

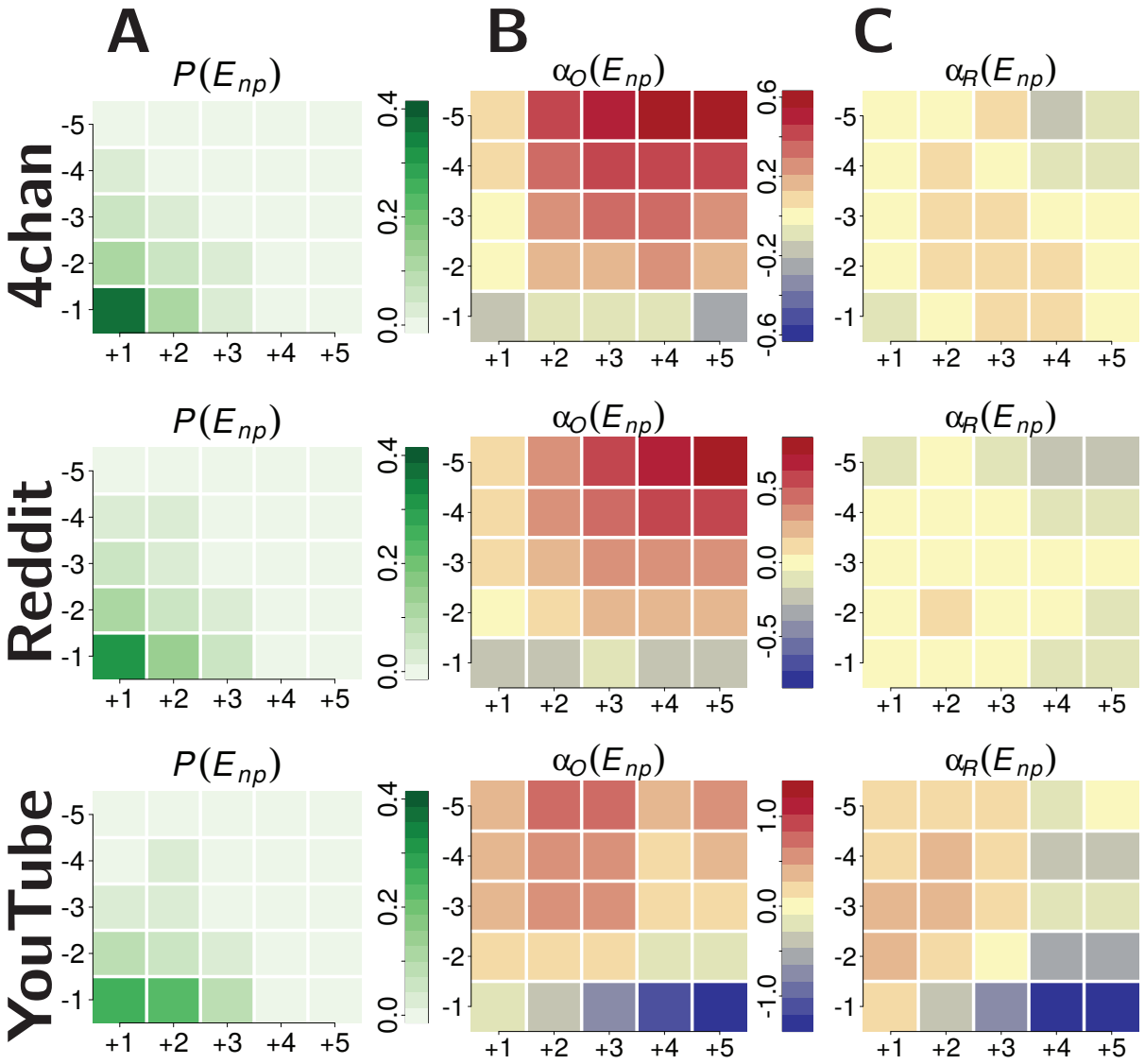


Figure 3.5: Emotional pattern of communication activity in the online community. A) The baseline probability of the co-activation of both positive and negative emotions in all messages across three datasets $P(E_{np})$. B) The normalized log-likelihood of the co-activation of both positive and negative emotions E_{np} in the original message $\alpha_O(E_{np})$ and C) in the reply message $\alpha_R(E_{np})$ across responses on `4chan`, `Reddit` and `YouTube`.

negative $P(E_n|E_p = +1)$ messages differ across datasets. On `YouTube`, the number of strictly positive messages is twice of the number of strictly negative messages (0.35 vs. 0.16), which is an evidence for the presence of the *positivity offset*. On `Reddit`, the fractions of strictly positive and negative messages do not differ (0.20 vs. 0.22), however on `4chan` negative messages are slightly prevalent than positive (0.18 vs. 0.24). The different statistics of positive and negative messages can be possibly explained by the difference in the functional usage of these online platforms or the privacy concerns of users, see Section 3.5. From this analysis, we conclude that across three datasets neither negative nor positive emotional expressions prevail, with just slight positivity offset observed on `YouTube`.

Compared to the baseline emotional expression, emotions expressed in the messages that receive a reply and in the reply messages show significant differences. In Figure 3.5B, the heatmaps of the likelihood of emotions $\alpha_O(E_{np})$ in *messages that receive a reply* show the concentration of extreme positive values (red and dark red colour) in the top-right corner and the top row across three datasets. The top-right values correspond to the co-activation of both extreme positive and extreme negative emotions in the original message, such as $(-5, +5)$ or $(-5, +4)$, and the top row values map to a presence of extreme negative emotions, such as -5 or -4, regardless of the expression of positive emotions. This finding demonstrates a *distinct emotional pattern* of messages that elicit response in users, and sheds light on which type of messages users have a tendency to reply to. A) These messages express strong negative sentiments, which confirms the theory of the *negativity bias*. B) These messages are also emotionally ambivalent, *i.e.* they elicit simultaneously both strong positive and strong negative emotions. And C) these messages have a lower likelihood of containing positive emotions, this effect is especially pronounced on `YouTube`. Additionally, we confirmed this emotional pattern in messages that receive replies across different topics and categories in three datasets, see Appendix B.2.

The effect of the likelihood of emotions in *messages receiving a reply* is even more pronounced when computing the likelihood of *emotional arousal* $\alpha_O(E_a)$ (Figure 3.6A, red lines). We observe an increasing difference in the conditional probability of extreme emotional arousal values in messages receiving replies as compared to the baseline probability of arousal. Across all datasets, the emotional arousal of 8 is 1.5 ($= \exp^{0.4}$) times more likely to be observed in a message that receives a reply $P(E_a = 8|M^r)$ than in a randomly selected message $P(E_a = 8)$. On `Reddit`, for instance, the emotional arousal of 10 is 2.2 ($= \exp^{0.8}$) times more likely to be expressed in a message that has a reply than in any randomly drawn message from the `Reddit` sample. A clear arousal threshold of 4 divides the low-arousal regime from the high-arousal regime. At $E_a > 4$, the probability of emotional arousal among messages receiving replies is higher than of the baseline arousal probability. This finding is in line with the findings in the likelihood of retweets by Pfitzner *et al.* (2012). The mechanism of retweets and replies is different, but the emotional pattern of posts that are either retweeted or that are replied shows similarity. Our finding on emotional arousal and valence together with the consistent results from the different categories of datasets suggest that users react to the messages of a distinct emotional pattern.

Incidentally, among the *reply messages*, we do not observe a clear emotional pattern: neither in the expressed valence $\alpha_R(E_{np})$ (Figure 3.5C), nor in the expressed emotional arousal $\alpha_R(E_a)$ (Figure 3.6A, blue lines). To complete the analysis, we *model the emotion of the reply message* as a linear function of the emotions expressed in the original message, see Figures 3.7A–C). The model reveals that the values of the coefficients of the expressed negative emotion are greater than of the expressed positive emotion across all datasets. This results in stronger negative replies rather than in stronger positive replies and serves as an additional evidence for the presence of the negativity bias online.

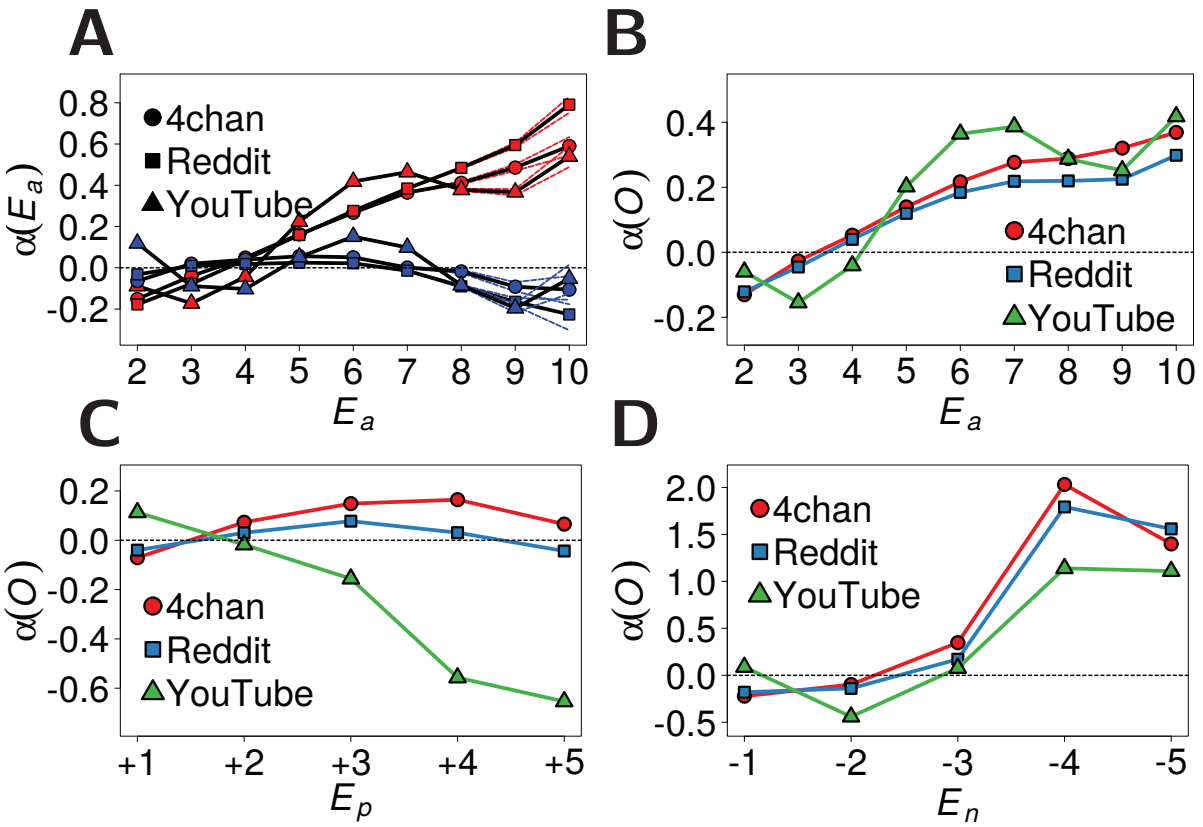


Figure 3.6: Emotional pattern and the influence of emotions in the online community. A) The normalized log-likelihood of emotional arousal E_a in the **original message that receives a reply** $\alpha_O(E_a)$ (red) and in the **reply message** $\alpha_R(E_a)$ (blue) across response pairs on `4chan`, `Reddit` and `YouTube`. Dashed lines show confidence intervals. Dashed black line indicates the baseline probability $P(E_a)$ of expressed emotional arousal E_a in a message. B) The normalized log-likelihood of original (or stimuli) messages to receiving a reply $\alpha(O)$ conditional to emotional arousal E_a, C) conditional to positive emotions E_p, and D) conditional to negative emotions E_n. Dashed black line indicates the baseline probability of an original message to receiving a reply $P(O)$. Positive values indicate positive effect of expressed emotions on receiving a reply, negative – negative effect, 0 – no effect.

In the last paragraph, we discuss the results of our second introduced metric in Equation 3.3, which measures the effect of emotional arousal and valence expressed in a message on a probability of this message to receiving a reply, namely $\alpha_{E_a}(O)$ and $\alpha_{E_{np}}(O)$. First, we observe that arousal has a positive effect on a message to receiving a reply, see monotonously increasing graphs in Figure 3.6B. The effect of negative emotions is even more pronounced (Figure 3.6C). For example, on `4chan` and `Reddit` an original message expressing extreme negative emotions of -4 is 7.3 ($= \exp^{2.0}$) times more likely to receive a reply, and on `YouTube` such extreme negative emotions in text give a three ($= \exp^{1.1}$) times factor increase to a message to receiving a reply. On the other hand, an expression of positive emotions does not have an effect on receiving a reply, see `4chan` and `Reddit` (Figure 3.6D). On `YouTube`, for instance, extreme positivity in text even reduces by halve the chance of a message to receiving a reply. To complete analysis, we model the probability

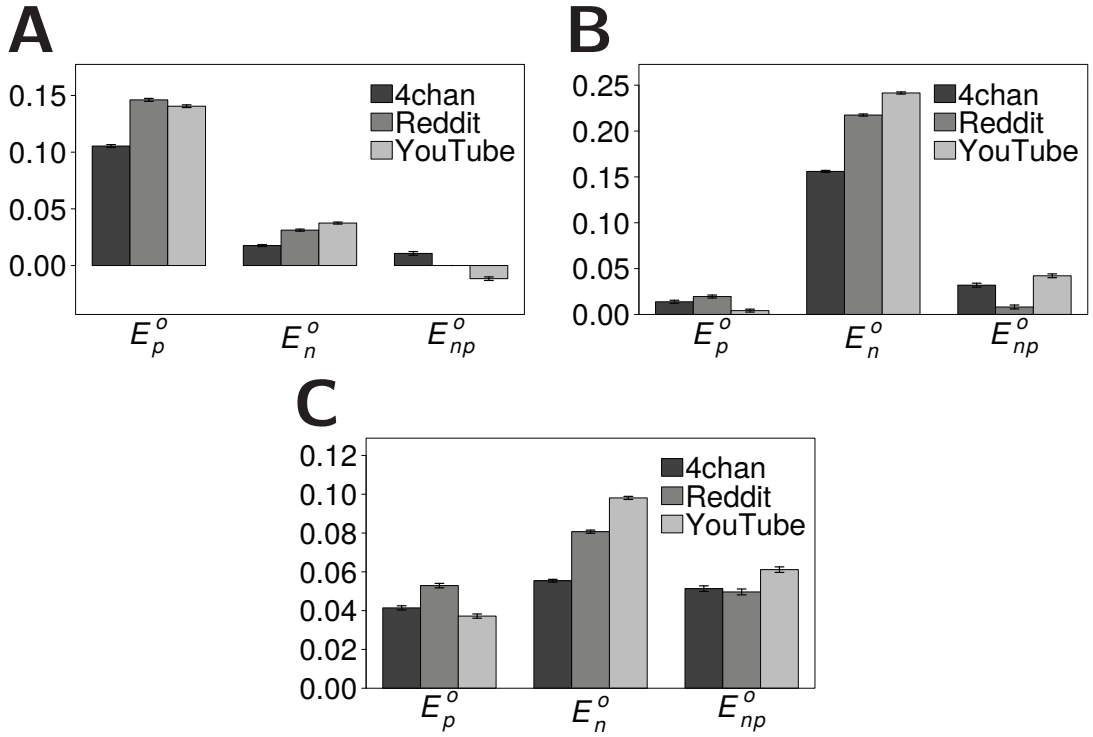


Figure 3.7: Barplots of regression coefficients and their confidence intervals obtained for modelling: A) positive emotion of reply message, $E_p^r \sim E_p^o + E_n^o + E_{np}^o$, B) negative emotion of reply message, $E_n^r \sim E_p^o + E_n^o + E_{np}^o$, C) co-activation of both positive and negative emotion of reply message, $E_{np}^r \sim E_p^o + E_n^o + E_{np}^o$. Normalized predictor variables are emotional expression of a replied (original) message: positive emotion E_p^o, negative emotion E_n^o and co-activation of both positive *and* negative emotions E_{np}^o.

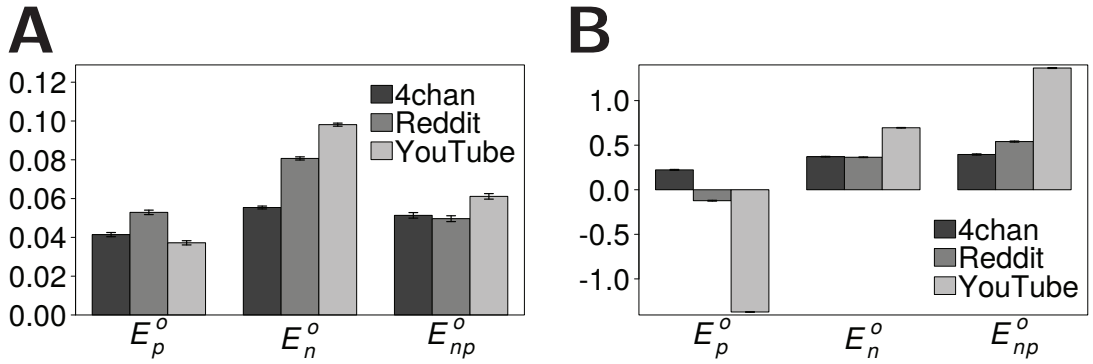


Figure 3.8: Barplots of regression coefficients and their confidence intervals obtained for modelling: A) probability for a message to have a reply, $P(R) \sim E_p^o + E_n^o + E_{np}^o$, B) number of responses, $|R| \sim E_p^o + E_n^o + E_{np}^o$. Normalized predictor variables are emotional expression of a replied (original) message: positive emotion E_p^o, negative emotion E_n^o and co-activation of both positive *and* negative emotions E_{np}^o.

of a message to receiving a reply and the number of replies it receives as a function of emotions it expresses, see Figure 3.8D and Figure 3.8E. Through the model we show that the coefficients of expressed negativity and of ambivalent emotions are greater than that of expressed positivity. These results serve as an additional evidence for the presence of the negativity bias.

Our findings are in line with the previous observations of emotional interactions and emotional influence in persistent type of online communication such as blogs and fora. Early findings have shown that negative emotional posts drive communication among users and extend the lifetime of the online discussion in forum (Chmiel *et al.*, 2011b; Mitrović *et al.*, 2010). Our results suggest that not only the negatively charged posts make people to respond to them, but emotionally ambivalent,[3] *e.g.* polarized, divergent and maybe controversial, messages promote discussions among users. These results are an additional contribution into the research community that studies the role of emotions on human conversations online.

3.5 Discussion

We started with the question of online *social sharing of emotions* (Section 1.2). Our analysis used large-scale data of collective responses (16-24M) of individuals across three popular online communities `4chan`, `Reddit` and `YouTube`. Following Garas *et al.* (2012), we described the communication process as two-dimensional: the activity and the emotional expressions of users, and tackled our question by measuring the activity of users as a function of their emotional expression. If the activity of users is expressed as a time interval Δt between a message and a reply to the message, our findings show no differences in response time to messages expressing positive versus negative emotions. This was evidence against the *negativity bias* (Cacioppo and Berntson, 1994; Cacioppo and Gardner, 1999; Cacioppo *et al.*, 1997; Miller, 1961), the theory that negative stimuli elicit faster responses, discussed thoroughly in Section 1.2. The picture differs when measuring the activity of users as a likelihood to reply, following the Bayes' approach by Pfitzner *et al.* (2012). Here, priority to reply to a message is characterized not by the time scale but solely by the reaction: to reply to a message or ignore it. Our results show that messages that elicit replies are more emotionally ambivalent and exhibit stronger negative emotions, which supports the hypothesis of a presence of a negativity bias in online interaction. Furthermore, episodes that express emotional arousal above a certain threshold have a higher chance of receiving a response. This threshold has been observed in the retweeting behaviour of users (Pfitzner *et al.*, 2012), but our study confirms it for a different mechanism, namely replying. Combined with the first finding on temporal interaction, the second result on response likelihood shows that what makes us react online is more salient than when we react, which highlights the difference between emotional interaction in offline face-to-face interaction and in online computer mediated communication.

[3]See Appendix B.1 for a sample of posts which were classified by the emotion detection tool `SentiStrength` as ambivalent messages. These are the messages expressing mixed emotions with both positivity and negativity.

Our first finding indicates that in online communication the Internet medium changes the time scale of emotional interaction such that rapid emotional responses are not observable. This can be due to reduced spontaneity in the computer-mediated communication (CMC) as compared to face-to-face (Derks *et al.*, 2008). In CMC, users have more time to think and reflect upon a message and then they also need time to type the message. This process creates a time lag between the initial emotional reaction and its written expression. While online environment and CMC control the speed of the emotional expression and inhibit the emotional impulses of the user, this does not imply that online emotional interactions are poor. The second finding shows that Internet users *do* respond to the emotion eliciting episode. Rather, the response to such an emotion eliciting episode does not happen instantaneously and might materialize even years later. Users' attention is attracted not only by the novelty of the story but also by its emotional charge. This finding has a practical implication for the designers of online platforms, who might be interested in measuring the emotional level of discussions and in promoting emotionally rich content.

One of our findings shows that emotionality expressed in replies depend on the emotions in the stimuli messages. Users not only react to emotional messages but they react differently. For instance, on YouTube, messages with negative emotions receive negative replies. This agrees with previous study showing that negative emotions drive discussions and boost user activity (Chmiel *et al.*, 2011b). Furthermore, emotional reactions in online discussions have been shown to be of predictive power for presidential approval rates (Gonzalez-Bailon *et al.*, 2010). Given the increasing amount of time that individuals spend online and participate in discussion forums, understanding emotional interactions among Internet users might help mitigate social conflicts (Khatib *et al.*, 2012), which may have important consequences for democratic processes (Chmiel *et al.*, 2011b).

Analysis of online posts must also discuss the effect of anonymity on the Internet (Chmiel *et al.*, 2011b). Our findings revealed a slight *positivity offset* only among `YouTube` postings, but not on `4chan` or `Reddit`. This can be explained by the difference in the functional usage of these online platforms. On `YouTube`, video owners receive substantial amount of laudatory comments, *e.g.* "Thanks for a video", "Great video". Because it is primarily a video-hosting website, comments and thus also discussions are a secondary function of `YouTube`. `4chan` and `Reddit`, however, were initially created for sharing stories and for discussions among users, therefore, positive comments might not be as prevalent as compared to `YouTube`. Privacy concerns of the users can be another explanation for *positivity offset* observed on `YouTube`. It is the most de-anonymized platform, whereas `4chan` is the most anonymous of the three. It has been shown that in anonymous communication, people tend to show unrestrained behaviour, be more critical and express more negative emotions than when they can be easily recognized (Derks *et al.*, 2008; Kushin and Kitchener, 2009; Siegel *et al.*, 1986). Anonymous communication on `4chan` and `Reddit` might

lead to users expressing themselves without including positive words, which as a result shifts the positivity offset to a neutral one.

Finally, our results are in line with psychological tests, where physiological reactions in individuals were elicited due to the emotional content of online posts (Kappas *et al.*, 2010, 2011; Küster *et al.*, 2011; Theunis *et al.*, 2010). Combined with the earlier results in psychology, our findings might serve as an additional microscopic rule for modeling interpersonal emotional communication using agent-based modeling (Garas *et al.*, 2012; Schweitzer and García, 2010).

Chapter 4

Emotional Polarization in Online Communities

Summary

We analyze online collective evaluation processes through positive and negative votes in various social media. We find two modes of collective evaluations that stem from the existence of filter bubbles. Above a threshold of collective attention, negativity grows faster with positivity, as a sign of the burst of a filter bubble when information reaches beyond the local social context of a user. We analyze how collectively evaluated content can reach large social contexts and create polarization, showing that emotions expressed through text play a key role in collective evaluation processes.

Based on the extended abstract "When the Filter Bubble Bursts: Collective Evaluation Dynamics in Online Communities", by Adiya Abisheva, David Garcia, and Frank Schweitzer, *Proceedings of the 8th ACM Conference on Web Science*, pp. 307–308 (2016). In this chapter we include the extended version of the published work. A.A. produced the majority of the statistical analyses and the plots, and gave the major input in writing the manuscript.

4.1 Introduction

In this Chapter, we investigate another instance of opinion polarization, namely the emotional polarization. As part of the research questions, we test the Hypotheses 4 and 5, see Section 1.3.3, on the statistical regularities of collective online evaluations and the differences between the popular discussions online. We start by presenting an elucidating example of an online item, that became both popular and polarizing. This example illustrates further the concepts of local and global collective attention, see Section 1.3.3.

Rebecca Black, an amateur teenage singer, posted a music video[1] on `YouTube` on February 10, 2011. The song originally circulated mostly among the Facebook friends of its 13-year old singer and was loved and positively commented. Rebecca Black's song received the "all the usual friends things" (Larsen, 2011) and was enough to please her, but it suddenly went viral *in the wrong direction*. From initial 4,000 views on `YouTube` her song skyrocketed to 13 millions views. This sudden popularity brought mostly negative attention, up to the point of becoming officially the most disliked `YouTube` video,[2] and by June 15, 2011 the song received 3.2 millions `dislikes` in `YouTube` against less than half a million `likes`. From *local* fame her song soared to the heights of *global* shame.

The anecdotal example of Rebecca Black's song is paradigmatic of some aspects of the collective dynamics of evaluations in online media. A video can become relatively popular within a small community and receive initial positive evaluations, but when larger audiences are reached, negativity rises faster than in early moments. Figure 4.1 shows this phenomenon through an example of the relative daily volume of `likes` and `dislikes` of a `YouTube` video. Initially, the video is positively evaluated, but the volume of `likes` decreases quickly. While initial `dislikes` also decrease, they start rising after the fourth day, reaching a peak at the ninth day.

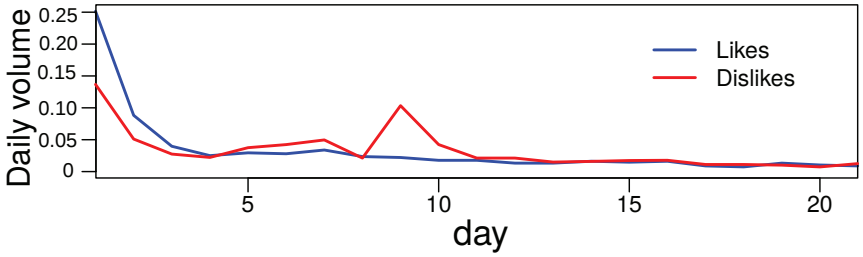


Figure 4.1: Example of evaluation dynamics in Youtube. Normalized daily volume of `likes` and `dislikes` for a video in our `YouTube` dataset. `Likes` appear soon after the video is uploaded, while `dislikes` tend to appear later.

[1]The original video was deleted and reuploaded again at: `https://www.youtube.com/watch?v=kfVsfOSbJY0`

[2]`http://knowyourmeme.com/memes/rebecca-black-friday`

The early viewers of a YouTube video are prone to like it, either due to a social connection with the uploader, or given the similarity of the video with their past liked content. This is a consequence of the purpose of social filtering mechanisms and recommender systems, which is to personalize content selection such that users find content that they consider relevant and of good quality. In contrast, the video can also spread through other media towards more general users, and eventually reach a global audience with users more critical or negative towards the video. Beyond YouTube videos, this phenomenon can be seen as another aspect of the filter bubble (Pariser, 2011): The reinforcement of opinions caused by filtering mechanisms creates an initial pocket of positivity, but *when the filter bubble bursts*, collective negativity can backlash.

In this chapter we set out to understand collective evaluation processes in various social media through likes and dislikes, as manifestations of opinions towards the evaluated content. We test the duality of collective evaluations in the local versus global behavior illustrated above, looking for the existence of a threshold of positivity after which negative evaluations rise faster and polarization emerges.

4.2 Dataset Description

The data used in this research is the result of our crawl of four publicly accessible online communities.

YouTube (http://www.youtube.com) is a video sharing website on which registered users can upload and view videos, as well as post comments and rate videos with likes and dislikes. Our crawl[3] was launched in June 2011 to daily collect a combination of top videos in various categories and to iteratively explore the channels of general users (Abisheva *et al.*, 2014), including 6.3 millions videos by February 2015.

Reddit (http://www.reddit.com) is a message board in which registered users submit posts with links and text, and vote up and down for posts to appear on a frontpage. Conversations between users appear in one of the many thematic boards, called subreddits, covering diverse topics from politics to science fiction and adult content. From 2012 to 2014 our daily Reddit crawl[4] collected 338,000 submissions from 1,972 subreddits. While the user interface of Reddit provides fuzzed amounts of votes, it is possible to construct the total amount of up and downvotes to a submission based on the JSON fields of reddit score and like ratio. This way, we count with the text and the final amount of up and downvotes for each submission in our dataset.

Imgur (http://www.imgur.org) is an image hosting and sharing website where registered users upload, rate, and discuss uploaded images. Image sharing traffic of Imgur has a large

[3]YouTube Data API Java wrapper (https://developers.google.com/youtube/v3/)
[4]PRAW (https://pypi.python.org/pypi/praw)

Dataset	Number of items, N			Number of `likes`	Number of `dislikes`
	N_{crawled}	$N_{\text{year}\geqslant 1}$	$N_{\text{L,D}\geqslant 1}$		
`Urban Dict.` definitions	220,270	213,512	208,441	61,100,699	26,508,869
`YouTube` video descriptions	6,279,461	3,864,480	2,750,554	763,291,676	41,214,035
`Reddit` submissions	338,845	174,444	142,662	5,078,242	947,519
`Imgur` image titles	201,181	147,752	125,230	54,786,629	1,931,918

Table 4.1: Number of items in each dataset. N_{crawled} counts the number of crawled items, and $N_{\text{year}\geqslant 1}$ the number of items in English *and* that existed for more than a year. $N_{\text{L,D}\geqslant 1}$ counts items that received at least 1 `like` *and* 1 `dislike`.

presence in `Reddit` such that every 6th successful `Reddit` post has a link to an image on `Imgur` (Olson, 2015). Our daily crawl[5] collected 200,000 images and their user activity statistics between December 2015 and January 2016.

Finally, `Urban Dictionary` (`http://www.urbandictionary.com`) is an online crowdsourcing platform consisting of non-standard lexicon of slang words and idioms. Registered users can submit new terms and provide definitions, and all users of the website, registered and anonymous, can vote up and down for the best definitions. Between April and May 2013 our python-based crawl collected 220,000 definitions and their votes.

All platforms provide functionality for users to evaluate uploaded content positively and negatively by clicking an upvote/`like` or downvote/`dislike` button respectively. For simplicity, from now on we refer to evaluated videos, submissions, images and definitions as *items* and we denote as `likes` and `dislikes` to positive and negative evaluations, including up and down votes respectively.

Sentiment analysis To quantify emotional expression, we applied sentiment analysis to headers or titles of each item, leaving for a future research the analysis of longer descriptions, transcripts, and comments. We applied sentiment analysis techniques to video descriptions in `YouTube`, image titles in `Imgur`, submission headers in `Reddit` and term definitions in `Urban Dictionary`. Headers and titles are a good proxy of the emotional tone of a discussion, in line with earlier research on forum-like conversations (Gonzalez-Bailon *et al.*, 2010).

We measured emotional content of items by applying two complementary sentiment analysis methods introduced in the Introduction. First, we apply a lexicon of affective norms of valence **V**, arousal **A** and dominance **D** of nearly 14,000 English words (Warriner *et al.*, 2013). In line with previous findings (Warriner *et al.*, 2013), the scores of valence and dominance in our dataset are highly correlated, in comparison with the weaker correlation between valence and arousal as explained more in detail in Section 4.6. This motivates our focus to only valence and arousal as suggested by the theory of core affect (Russell and Barrett, 1999).

[5]PyImgur Python API wrapper (`https://github.com/Damgaard/PyImgur`). Seed images were selected from `Imgur`'s gallery sitemap (`http://imgur.com/gallery/sitemap.xml`)

Second, we apply the `SentiStrength` classifier (Thelwall *et al.*, 2012, 2010a), a state-of-the-art lexicon-based method (Abbasi *et al.*, 2014; Kucuktunc *et al.*, 2012) described in detail in the Introduction. The final sentiment score is composed of a positive **P** and a negative **N** score for each text as two discrete values in the range of $[+1, +5]$ and $[-5, -1]$ respectively. In our analysis, we normalize all emotions variables to [0..1] mapping **P** from $[+1, +5]$ to $[0, 1]$ and reversing and rescaling **N** from $[-1, -5]$ to $[0, 1]$.

To ensure a valid measurement of sentiment and collective evaluations, we apply two filters to our datasets. First, since both sentiment analysis techniques are designed only for English texts, we apply language classification (Nakatani, 2010–current) and filter out all non-English texts. Second, we remove all items with less than a `like` and a `dislike`, and that existed for less than a year in all platforms, to ensure that positive and negative evaluations are stable. Detailed statistics on the number of posts in each dataset are shown in Table 4.1, showing that they are still sufficient for large scale analyses.

4.3 Statistical Analysis Methods

Distribution fits We apply a Maximum Likelihood criterion to fit the distributions of likes and dislikes (Alstott *et al.*, 2014), to confirm early findings of the fits of the popularity distribution to the log-normal distribution (Asur *et al.*, 2011; Van Mieghem *et al.*, 2011). We use the `powerlaw` python package to fit four statistical distributions related to complex growth phenomena (Mitzenmacher, 2004): power law, log-normal, truncated power law and exponential distributions. We compare the likelihood of each distribution using the log-likelihood ratio $R = \ln(\frac{L_1}{L_2})$ between the two candidate distributions and its significance value p. Positive ratios indicate evidence for the first distribution, and negative ratios for the second one. Instead of testing the hypothesis of the data following a certain distribution, this comparative test answers the question of which parametric distribution provides the best fit available, following the principle of Maximum Likelihood estimation (Alstott *et al.*, 2014). To finally assess the quality of the best fit, we measure the Kolmogorov-Smirnov distance between the best fitting distribution and the emprical data.

Dual regime analysis We test the existence of a dual local versus global regime in collective evaluations by analyzing the non-linear properties of the relationship between the amounts of likes and dislikes for each item.

We use an extension of a traditional linear modelling, multivariate adaptive regression splines (`MARS`) (Friedman, 1991, 1993) implemented in the **R** programming language package *earth*. `MARS` fits a continuous piecewise regression function with *knots* that join locally linear pieces. In our analysis, we are interested to test a dual pattern in the relationship

between the number of likes L and the number of dislikes D, therefore we set the number of knots to one and fit a model of the form

$$\mathrm{D(L)} = I + \alpha_1 * max(0, \mathrm{L} - L_c) + \alpha_2 * max(0, L_c - \mathrm{L}). \tag{4.1}$$

The values of likes above L_c *and* the values of dislikes above $\mathrm{D}(L_c)$ correspond to observations in the global regime, after the bubble bursts, and the values in which any is below map to the local regime.

To evaluate the quality of the `MARS` model, we compare it to the Ordinary Least Squares (OLS) regression using the Generalized Cross-Validation prediction error (GCV):

$$GCV = \frac{RSS}{N * (1 - \frac{\mathrm{ENP}}{N})^2}, \tag{4.2}$$

where N is the number of observations, RSS is the residual sum of squares, and ENP is the effective number of parameters to avoid overfitting (Friedman, 1993). We use the implementation provided by the package *boot* in **R** as well as the coefficient of determination R^2 of both OLS and `MARS` fits.

Emotion and polarization analysis Having identified the two regimes and their thresholds in the relationship between the number of dislikes and the number of likes, we can mark items either in the global or the local regime as a binary class. We test how emotions influence the chances of items reaching the global regime through two logistic regression models, one for each sentiment analysis technique. Similarly, we combine the values of likes and dislikes through their geometric mean to measure polarization, as manifested by simultaneous large amounts of positive and negative evaluations. We regress this measure of polarization through two linear models depending on the emotions expressed on the items. Prior to modelling, we examine the normalized emotional dimensions for multicollinearity by computing the Spearman's rank correlation coefficients, to avoid singularities. We assess the quality of fits in comparison to null models, by measuring the χ^2 statistic of model likelihood ratio tests implemented in the `lmtest` **R** package.

4.4 Stylized Facts of Evaluation Distributions

Figure 4.2 shows the probability density functions of the distributions of the amount of `likes` and `dislikes` for items in each of the four datasets. To understand the process that generates these distributions, we fit a set of parametric distributions that provide insights into how `likes` and `dislikes` are given to items. Following the categorization of Mitzenmacher (2004), generative mechanisms produce stylized size distributions that can be traced back to the properties of growth processes. If the appearance of `likes` and `dislikes` follows an uncorrelated process and new evaluations are independent of previous ones, `likes` and `dislikes` should follow *exponential* distributions. On the other hand, the

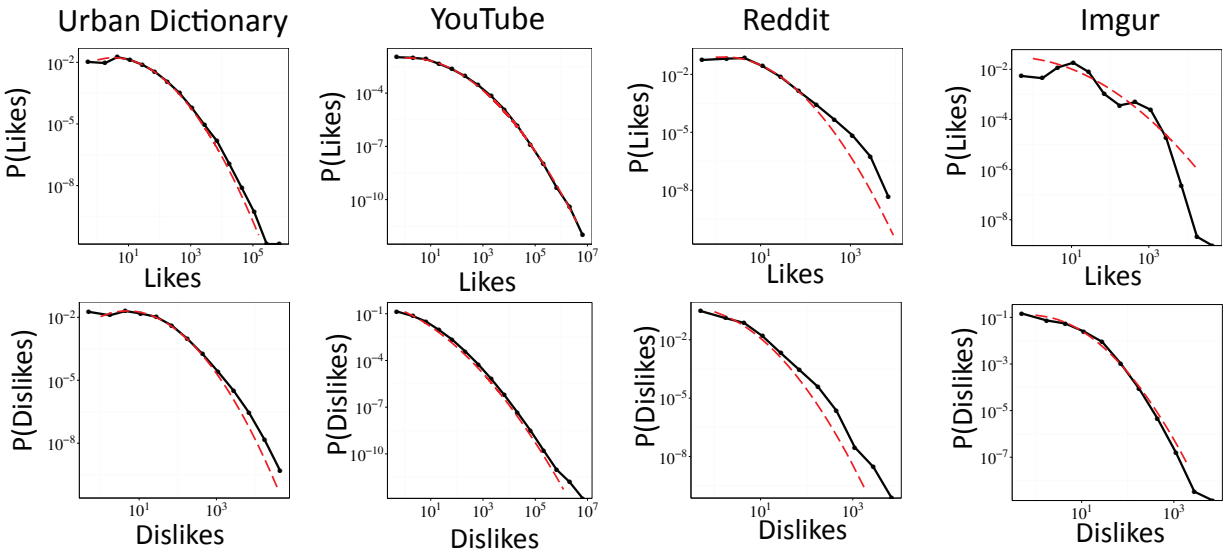


Figure 4.2: Probability density function of collective evaluations. Probability density function of the number of `likes` (top) and the number of `dislikes` (bottom) with exponential binning and fits to log-normal distribution $\ln\mathcal{N}(\mu,\sigma)$ (red dashed lines). For all datasets, the results of the log-likelihood pairwise comparisons of the four distributions (see text) identified the log-normal distribution as the best fit.

	Urban Dictionary		YouTube		Reddit		Imgur	
	P(likes)	P(dislikes)	P(likes)	P(dislikes)	P(likes)	P(dislikes)	P(likes)	P(dislikes)
μ	4.092	3.657	5.492	1.405	2.197	0.492	4.668	1.821
σ	1.705	1.435	2.28	2.528	1.332	1.35	2.46	1.447
D	0.008	0.008	0.009	0.002	0.029	0.01	0.098	0.031
$\ln\left(\frac{L_{\mathrm{LN}}}{L_{\mathrm{PL}}}\right)$	123115.6***	129462.1***	174835.6***	42309.5***	45355.3***	13898.9***	65314.8***	34103.6***
$\ln\left(\frac{L_{\mathrm{LN}}}{L_{\mathrm{TPL}}}\right)$	55538.6***	61135.1***	252646.1***	42592.3***	17182.5***	10084***	20108.5***	3662.4***
$\ln\left(\frac{L_{\mathrm{LN}}}{L_{\mathrm{EXP}}}\right)$	260098.8***	126865***	953075***	1270237.6***	149405***	70261.5***	75817.7***	26681.4***

Table 4.2: Log-normal fit of parameters of collective evaluations and comparison to other distributions. Estimated parameters of the fitted log-normal distribution $\ln\mathcal{N}(\mu,\sigma)$ and Kolmogorov-Smirnov distances D. The bottom row shows the log-likelihood ratios of pairwise comparison between the log-normal distribution fit (numerator) and the other three distributions: power law, truncated power law and exponential. All three ratios are positive, large and significant ($p < 0.05$) which confirms that among the four candidate distributions the log-normal distribution is the best fit.

presence of `likes` and `dislikes` can motivate further evaluations through social effects, creating multiplicative growth (also known as preferential attachment in the context of networks). In the presence of multiplicative growth, if items have similar lifespans, `likes` and `dislikes` follow *log-normal* distributions. On the other hand if multiplicative growth is combined with heterogeneous lifespans, `likes` and `dislikes` follow a *power law* distribution. This power law can be corrected by adding an exponential cutoff if finite size effects limit the growth of `likes` and `dislikes`, a case in which the distributions would be better fitted by a *truncated power law*.

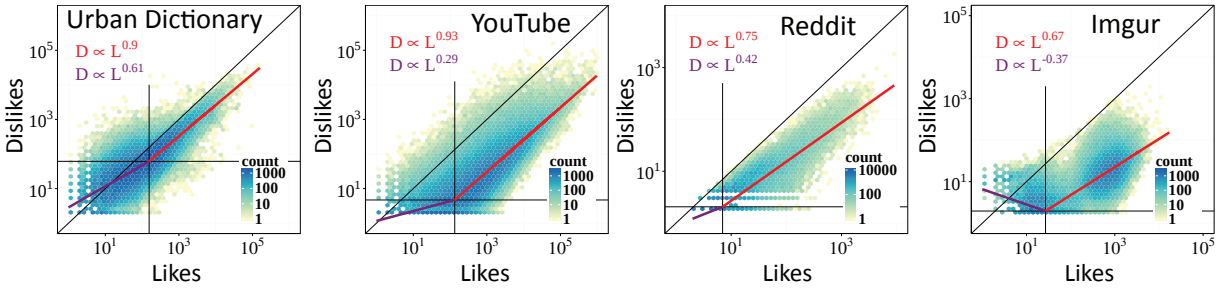


Figure 4.3: Relationship between the number of `dislikes` and `likes`. Two-dimensional joint distributions with 50 bins, bin colors indicate the count of observations within the bin. Purple and red lines show the local and global regimes of the non-linear relationship between the number of `dislikes` and the number of `likes`. Threshold estimates are located at L_c, estimated as $L_c = 155$ in `Urban Dictionary`; $L_c = 131$ in `YouTube`; $L_c = 7$ in `Reddit`; and $L_c = 27$ in `Imgur`.

For all datasets, the results of pairwise comparisons of the four proposed distributions identified the *log-normal* distribution as the best fit, with significant and positive log-likelihood ratios as shown in Table 4.2 along with the best fitting parameter estimates. The dashed lines in Figure 4.2 show the fitted distributions, revealing the quality of the fit. The cases of `YouTube` and `Urban Dictionary` provide very good fits with extremely low Kolmogorov-Smirnov D statistics. The fits are not so good at the tails of `Reddit` and `Imgur`, but the the Kolmogorov-Smirnov D statistic provide good values below 0.05 and the *log-normal* distribution clearly outperforms all others. The worst fit is for the number of `likes` in `Imgur`, for which Figure 4.2 suggests a bimodal pattern. Identifying the possible mechanisms that can produce such bimodality goes beyond the scope of this research. We can conclude that the amounts of `likes` and `dislikes` display a general heavy tailed behavior of *log-normal* distributions, lending evidence for the production of evaluations following socially coupled growth processes with homogeneous life spans.

4.5 The Dual Pattern of Collective Evaluations

We explore the existence of a dual relationship between `likes` and `dislikes` through non-linear `MARS` fits, testing if the relationship can be divided in a local and a global regime. We restrict the number of model terms to have a single knot, measuring if a dual model outperforms a linear pattern. Figure 4.3 shows the results of `MARS` fits between the logarithms of `likes` and `dislikes`. Vertical and horizontal lines mark the `likes` cutoff value L_c and its corresponding value of `dislikes` in the fit $\mathrm{D}(L_c)$. These cutoff values divide the system in a local versus a global regime, with the fitted functions of the form $D \propto L^{\lambda}$ and $D \propto L^{\gamma}$ respectively.

model	Urban Dict.	YouTube	Reddit	Imgur
R^2 (lm)	0.646	0.634	0.727	0.505
R^2 (MARS)	0.654	0.683	0.741	0.597
GCV (lm)	0.785	1.283	0.301	0.804
GCV (MARS)	0.767	1.111	0.286	0.654

Table 4.3: The goodness of fit of the dual and that of the linear model. Comparison of the linear and the MARS models of the relationship between the number of dislikes and the number of likes. Top row shows the coefficient of determination R^2 (higher is better). Bottom row shows the generalized 10-fold cross-validation prediction error (GCV) (lower is better). The dual model outperforms in R^2 and in GCV compared to the linear model.

In all datasets, the exponent of the global regime is larger than exponent of the local one, for example in YouTube $\gamma = 0.93 > \lambda = 0.29$. While both exponents are below 1 and indicate sublinear scaling, the much higher value of the second one shows that, beyond a threshold value of likes, the dislikes given to items grow faster than below the threshold as a sign of the burst of a filter bubble. The presence of scaling in Reddit votes was previously reported in a smaller data subsample (Van Mieghem, 2011), concluding the existence of superlinear scaling of dislikes with likes. Our analysis shows that the relationship between likes and dislikes in Reddit is better approximated by a dual regime model, in line with the results of the other three datasets.

We evaluate the goodness of the dual model against a single regime model in Table 4.3. The dual model outperforms in R^2 and GCV to the single regime model, lending strong evidence to the existence of two regimes. We further tested if additional knots could improve the fits, and found that a dual regime is the optimal model for Urban Dictionary, YouTube, and Reddit, and only a 4 knot model could improve the Imgur fit by a marginal GCV of less than 0.01.

4.6 Analysis of Emotions

4.6.1 Emotions in the Global Regime

Figure 4.4 illustrates the rank correlations between emotional dimensions. In all datasets valence and dominance are *highly correlated* with $\rho \geqslant 0.7^{***}$, and therefore we discard the dominance variable from regression analysis as it is difficult to distinguish from valence. Valence and positivity **P** have a minor positive significant correlation $\rho \in [0.2, 0.3]$, and valence and negativity **N** have a slightly negative correlation $\rho \approx -0.3$, illustrating the relation of emotion variables accross both valence/arousal and positive/negative models.

We fit two regression models in which the probability of the event of an item reaching the global regime G depends on the emotions expressed in the evaluated item. The first model

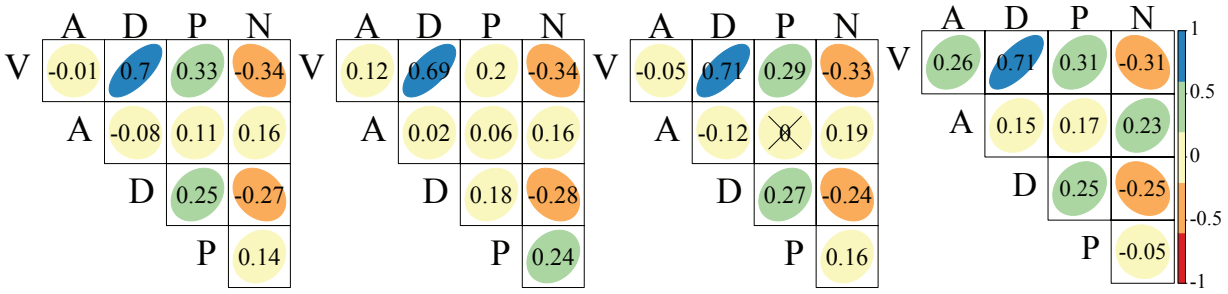


Figure 4.4: Correlations of emotions. Spearman's correlation matrix of emotional dimensions, in an order from left to right: A) `Urban Dictionary`, B) `YouTube`, C) `Reddit`, D) `Imgur`. Significance level $p < 0.05$. Insignificant correlations are crossed out. Dominance is highly correlated with valence, and therefore the dominance variable is discarded from the further analysis.

	`Urban Dict.`	`YouTube`	`Reddit`	`Imgur`
Intercept	-2.071^{***}	-0.305^{***}	-0.111^{***}	0.228^{***}
V	0.976^{***}	0.618^{***}	-0.262^{***}	-0.209^{***}
A	0.584^{***}	-0.049^{**}	$-0.006(n)$	0.300^{***}
χ^2	547.3^{***}	2791.4^{***}	44.1^{***}	35.7^{***}
	`Urban Dict.`	`YouTube`	`Reddit`	`Imgur`
Intercept	-1.369^{***}	-0.115^{***}	-0.259^{***}	0.261^{***}
P	1.019^{***}	0.581^{***}	-0.166^{***}	-0.296^{***}
N	0.170^{***}	0.218^{***}	$-0.006(n)$	0.191^{***}
χ^2	2552.6^{***}	17150.9^{***}	41.9^{***}	120.3^{***}

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, (n) not significant.

Table 4.4: The role of emotions in the global regime. Logistic regression models, $logit(G) \sim V + A$ and $logit(G) \sim P + N$, results for probability of an item to be in a global regime. The effect of arousal and valence is heterogeneous, and depends on the dataset.

uses **V** and **A** as explanatory variables, and focuses on the role of emotions as quantified through their pleasant/unpleasant and active/calm dimensions. The second model takes **P** and **N** as predictors, and measures significance of positive and negative sentiments in bringing an item to global regime. Table 4.4 reports the results of logistic regression of the form $logit(G) \sim V + A$ and $logit(G) \sim P + N$ respectively.

The role of arousal is heterogeneous, having a significant positive effect in `Imgur` and `Urban Dictionary`, but a weak negative effect in `YouTube` and a non-significant one in `Reddit`. The effect of valence is also mixed, in `Urban Dictionary` and `YouTube` the chances of reaching the global regime grow with valence, while in `Reddit` and `Imgur` is the opposite case. The second model sheds more light to this: the pattern is the same for positive sentiment, but negative sentiment increases the chance of reaching the global regime in all datasets but `Reddit`, where the effect is not significant.

4.6.2 Emotions as Predictor of Polarization

Since the distributions of `likes` and `dislikes` are approximately log-normal, we can treat the logarithms of `likes` $\ln(L)$ and `dislikes` $\ln(D)$ as centrally distributed around their means $\langle \ln(D) \rangle$ and $\langle \ln(L) \rangle$. We standardize the logarithmic counts $\ln(D)$ and $\ln(L)$ as:

$$Z_L = \frac{\ln(L) - \langle \ln(L) \rangle}{sd(\ln(L))} \quad Z_D = \frac{\ln(D) - \langle \ln(D) \rangle}{sd(\ln(D))}, \tag{4.3}$$

where $sd(\ln(L))$ and $sd(\ln(D))$ are the standard deviations. Then, we compute a measure of polarization as the geometric mean of both values: $Pol = \sqrt{Z_L * Z_D}$. This measure captures the principle that polarization is high under simultaneous large amounts of positive and negative evaluations, and that polarization is low when only one of the values is dominant.

To understand which kind of emotional content creates polarization, we fit two regression models as in Section 4.6.1, one of polarization as a function of valence and arousal in the evaluated item, and another as a function of positive and negative sentiment scores. The results of the fits are shown in Table 4.5. In line with the theory that links arousal to more extreme opinions (Paulhus and Lim, 1994; Reisenzein, 1983; Zillmann, 1971), we find a general pattern in three datasets where arousal leads to higher levels of polarization. While there is no significant effect in `Reddit`, all the other datasets show that items that contain words that transmit higher arousal also create a stronger polarized response.

This also manifests in the model using positive and negative scores, where negative content predicts higher polarization in the same three cases as for arousal. The results of these two metrics are consistent with the hypothesis that the expression of activating and negative feelings, such as anger or outrage, tend to create more polarized responses, in line with the

	`Urban Dict.`	`YouTube`	`Reddit`	`Imgur`
Int.	2.0508^{***}	1.4543^{***}	1.3511^{***}	1.7623^{***}
V	0.3132^{***}	0.2980^{***}	-0.1954^{***}	-0.1908^{***}
A	0.2662^{***}	0.1005^{***}	$-0.0327(n)$	0.2399^{***}
χ^2	480.64^{***}	3420.4^{***}	97.315^{***}	88.669^{***}
	`Urban Dict.`	`YouTube`	`Reddit`	`Imgur`
Int.	2.2744^{***}	1.5896^{***}	1.2220^{***}	1.7625^{***}
P	0.4889^{***}	0.3059^{***}	-0.1107^{***}	-0.1484^{***}
N	0.1194^{***}	0.1698^{***}	$0.0077(n)$	0.1672^{***}
χ^2	3271.2^{***}	19830.0^{***}	63.073^{***}	170.81^{***}

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, (n) not significant

Table 4.5: The role of emotions in the polarization. Linear regression models, $Pol \sim V + A$ and $Pol \sim P + N$, results for polarization of the evaluation of an item as a function of emotions expressed on its text. Arousal and negativity drive polarization in all datasets except `Reddit`. The effect of valence and positivity is dataset-dependent.

theoretical argument that poses emotions as mechanisms to speed up evaluation processes at the expense of more extreme reactions.

Valence in evaluated items creates different responses. Two communities, `Imgur` and `Reddit`, show a negative relation of polarization with valence and positive sentiment. The other two, `Urban Dictionary` and `YouTube`, show the opposite, where polarization increases with valence. This suggests a context dependent interpretation of positive expression, which does not necessarily motivate positive empathy but can also fuel polarized responses. The positive and negative scores model works better than the valence and arousal model in all cases but `Reddit`, where the valence and arousal model was more explanatory for polarization, as evidenced by χ^2 tests comparing both models.

4.7 Discussion

Our study of emotions focuses on understanding the role of emotions expressed in the text of items with relation to the chances that the items reach the global regime and produce polarized evaluations. While we used two established and validated sentiment analysis methods based on metrics from psychology, future advanced techniques can reveal new patterns and potentially falsify the conclusions of our analysis with current techniques. Furthermore, deeper analyses on individual data can correlate the expression of individual emotions in the comments of a user and the evaluations given by the user, bridging closer this way the measurement of emotional states and evaluations and providing a better understanding of interpersonal emotions.

Following an observational approach to collective evaluations has the advantage of having high ecological validity, but lacks the level of control that can be induced in experimental scenarios. We can deduce insights on the factual properties of collective evaluations, such as the dual regime between likes and `dislikes`, but testing the conditions that produce them requires a controlled set up. Our motivation and explanation for the dual regime stems from the phenomenon of filter bubbles (Pariser, 2011), but to fully understand how these filters affect our behavior we need to experiment on how individual evaluations respond to filtering mechanisms. While these experiments can be carried out it in typical psychological settings and surveys, large platforms like `Facebook` can also experiment with the behavior of their users in this respect (under the appropriate ethical considerations). A complete understanding of online evaluations can only be achieved when our results are complemented by experimental approaches.

The use of observational data has the advantage of taking a *natural exposure* approach: we analyze the evaluations of what people actually see, rather than the *forced exposure* to content in experiments (McPhee, 1963). In contrast, using digital traces of evaluations contains a selection bias by which some users might be responsible for much larger

amounts of `likes` and `dislikes` than other users. While this selection bias is natural at the collective level, inferring conclusions about the behavior of individuals needs to consider corrections and use richer datasets (Cuddeback *et al.*, 2004), or apply agent-based modelling approaches to connect the micro and macro levels (Schweitzer and García, 2010).

We explain the dual pattern between `likes` and `dislikes` as the result of filter bubbles, but other possible explanations might also be plausible. Some unkown deleting mechanism might downsample videos with a lot of `dislikes` in the local regime, or some external factor like audience size might explain the values of the thresholds. The results of our statistical analyses of distributions of `likes` and `dislikes` fit to hypothetical mechanisms of multiplicative growth, in line with previous findings on popularity metrics rather than evaluations (Asur *et al.*, 2011; Van Mieghem *et al.*, 2011). Our in depth statistics also provide a clear view on the limits of our results, for example in the worse fits of log-normal distributions in `Imgur`. Future research can conjecture on the possible alternative explanations of our findings, in particular with respect to which filtering mechanisms are in place. Our results do not allow us to distinguish social filtering, based on friends and follower links, from recommender systems, which are based on previous evaluations of a user. Further research with information on individual behavior can shed light on these different processes, for example measuring evaluation tendencies to content produced by friends versus strangers, or across assortative and disassortative links with respect to opinions.

Our analysis of the relation between `likes` and `dislikes` is based on the amounts given to items after a long time has passed. This way, we evaluate items after they do not attract lots of attention and their counts are stable. In a figurative way, we study the *fossils* of broken filter bubbles, but we do not study them in a live setting. To fully understand the dynamics of collective evaluations, we need data with temporal resolution on the counts of `likes` and `dislikes`. In general, such data is not publicly available on the sites, which requires a much more powerful crawling approach to monitor items on a frequent basis, or access to proprietary data.

Our contributions Our analysis of collective evaluations across various online media shows statistical regularities in the distributions of evaluations and their relationships. Our contribution is threefold: First we report that the distributions of the amounts of likes and dislikes per item are well fitted by log-normal distributions, a result that gives insights into the properties of the process that creates evaluations. Second, we test the existence of a dual pattern in the relation between likes and dislikes, finding robust evidence of the existence of a local and a global regime that is consistent with our hypotheses about the burst of filter bubbles. Third, we found evidence for the role of emotions in the creation of polarization and the access to the global regime, lending support for psychology theories about the role of affect, in particular arousal, in the polarization of opinions.

Our results have implications for the design of online platforms and filtering mechanisms. Recommender systems and filtering mechanisms allow users to discover content of relevance and quality, but can have unintended consequences in the large scale. Our results suggest that the increasing polarization levels of discussions might be created by these filtering mechanisms, and that users are at risk of receiving a negative backlash to their content when it goes beyond their local social context. Such abrupt behavior with respect to negative evaluations can have important consequences to user motivation and engagement, which might only be visible on the long run.

Our findings shed light on fundamental polarization processes, in particular with respect to the role of emotions. Increasing levels of polarization pose a risk of social conflict and hinder collaboration and common goods, but a healthy society needs certain level of disagreement to be able to deliberate, discuss, and take decisions about important topics. Calibrating the design of web and social media offers this way the chance to find a balance between stagnation and polarization, leading to productive interaction in our current online society.

Chapter 5

Multiplicative Growth Model of Collective Evaluations

Summary

We develop a data-driven statistical model of coupled collective evaluations governed by the law of proportionate effect and the decay in collective attention. The motivation comes from the previous findings on the statistical regularities of collective evaluations, namely the log-normal distribution, and the dual modes in the relationship between positive and negative online evaluations. From the time-stamped data on the number of `likes` and that of `dislikes` on `YouTube`, we learn that a) the growth rate of evaluations decays sharply since the initial upload of videos, and that b) the new amount of `likes` and the new amount of `dislikes` depend on the opposite signal. We incorporate these empirical observations of coupling of positive and negative signals and that of the decay in collective attention in the proposed model. The model reproduces well the statistical parameters of the empirical distribution of collective evaluations as well as certain characteristics of the burst of the filter bubble. However, the complete trajectories of the dynamics of collective evaluations cannot be recovered. We propose that model limitations can be later overcome by an incorporation of the emotional dynamics of agents.

A.A. contributed to designing the research questions. A.A. is the main responsible for the design and implementation of the statistical analyses, plots and the manuscript.

5.1 Introduction

In this Chapter, we present the model of human appraisals, to which we also refer as the model of collective evaluations. In Section 1.3.4, we have shown that previous research on collective evaluations has been mostly focused on positive evaluations. In this Chapter, we close this research gap by using the empirical findings on positive and negative evaluations from Chapter 4.

Plethora of research has been performed on popularity in online communities measured through the number of `views`, the number of story reads, the number of trending topics, or the number of upvotes or `likes`, see Chapter 1. Remarkable characteristics of the distribution of the online popularity have been revealed, namely the distribution follows the log-normal distribution. In Chapter 4, we have further confirmed this evidence on four large online datasets of different purposes, *e.g.* image-sharing, video-sharing, crowdsourcing, and news and discussion. Additionally, we have not only observed the log-normal distribution among positive evaluations, but also negative evaluations, *e.g.* `dislikes` or downvotes, are log-normally distributed with different parameters. Building on previous studies and our findings of statistical regularities in online evaluations, in this chapter we study in depth one of the known processes that generate the log-normal distribution, namely the multiplicative growth. Furthermore, we have observed that on platforms with both, liking and disliking capability, the distribution of both evaluations has different log-normal parameters and the resulting relationship between positive and negative evaluations becomes non-linear. Two modes of collective attention can be clearly distinguished – local popularity, where `likes` are greater than `dislikes`, and global attention where growth of `dislikes` is faster than that of `likes`. In this chapter, we build on the knowledge of the multiplicative growth model, also known as the law of proportionate effect, in order to model the growth of collective evaluations and, additionally, we propose an extension of this model that incorporates the coupling between positive and negative evaluations.

5.2 Background

5.2.1 The Law of Proportionate Effect

Log-normal distribution A log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed (Mitzenmacher, 2004). The random variable $\mathbf{X}$ has the log-normal distribution if $\mathbf{Y} = \ln(\mathbf{X})$ has the normal distribution. Equivalently, if $\mathbf{Y}$ has the normal distribution, then $\mathbf{X} = e^{\mathbf{Y}}$ has the log-normal distribution. From such expression, it is clear that log-normally distributed variable can only take positive real values.

Probability density function, $f(x)$	$\frac{1}{\sqrt{2\pi}x\sigma}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$
Cumulative density function, $F(x)$	$\Phi\left(\frac{\ln x-\mu}{\sigma}\right)$
Mean, $\mathbf{E}[X]$	$e^{\mu+\frac{\sigma^2}{2}}$
Variance, $\mathbf{VAR}[X]$	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$

Table 5.1: Key characteristics of the two-parameter log-normal distribution. Note: in CDF, Φ denotes the standard normal distribution function.

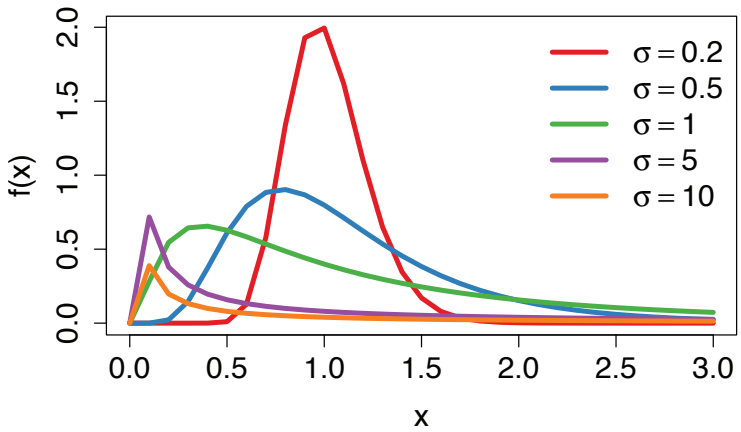


Figure 5.1: Examples of log-normal density functions with identical $\mu = 0$ but differing parameter σ.

In Table 5.1, we briefly list the key characteristics of the two-parameter form of the log-normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma \in (0, \infty)$. General form of the log-normal distribution takes, however, three parameters including the shift θ. Throughout the paper we use the following notation of a log-normal distribution with two parameters

$$\mathbf{X} \sim \ln \mathcal{N}(\mu, \sigma), \tag{5.1}$$

which says that random variable $\mathbf{X}$ is log-normally distributed with μ and σ. Given the properties of log-normally distributed variable, $\ln(\mathbf{X})$ is normally distributed with mean μ and standard deviation σ. In some sources the parameters of log-normal distribution are referred to as *location* μ and *scale* σ. Log-normal distribution with location 0 and scale 1 is known as the *standard log-normal distribution*. Examples of log-normal density functions with varying parameters are shown in Figure 5.1.

Genesis Generation of the log-normal distribution is governed by plethora of processes (Crow and Shimizu, 1988). One of such processes is the *law of proportionate effect* proposed by Gibrat (1930, 1931). Initially, Kapteyn (1903) considered the following equation

$$N(t) - N(t-1) = g(t) \cdot f(N(t-1)), \tag{5.2}$$

where $N(0)$ is the initial variable, $N(t)$ is the t-th step of the variable N and $\{g(t)\}_{1\leq t\leq T}$ is a set of mutually independent and identically distributed random variables, a.k.a as the *growth rate*, with mean $\mathbf{E}[g]$ and standard deviation $\mathbf{SD}[g]$, and t is a natural positive number, $t \in \mathbb{N}^+$. Moreover, $\{g(t)\}_{1\leq t\leq T}$ is statistically independent of the set $\{N(t)\}$. The special case where $f(N) = N$ reduces (5.2) to

$$N(t) - N(t-1) = g(t) \cdot N(t-1), \tag{5.3}$$

which tells that the value of the variable N at step t is some fraction $g(t)$ of the variable N at previous step $t-1$. The process that determines the sequence $\{N(t)\}$ given $N(0)$ is said to obey the law of proportionate effect, a.k.a the law of proportionate growth, (Gibrat, 1930, 1931).

The generation of the log-normal distribution is explained briefly as follows. Iterations of (5.3) lead to

$$N(T) = N(0) \prod_{t=1}^{T} (1 + g(t)). \tag{5.4}$$

Assuming that initial value is positive, $N(0) > 0$, we take the natural logarithm of (5.4) and multiply and divide the result by the number of steps T

$$\ln(N(T)) = \ln(N(0)) + T \cdot \frac{\sum_{t=1}^{T} \ln(1 + g(t))}{T}. \tag{5.5}$$

Let's introduce $X(t) = \ln(1 + g(t))$. By the Central Limit Theorem, if $S_T = \frac{\sum_{t=1}^{T} X(t)}{T}$ is a sample average of random variable X at time T, then as T gets larger, the distribution of S_T gets closer to the normal distribution with mean $\mathbf{E}[X]$ and standard deviation $\frac{\mathbf{SD}[X]}{\sqrt{T}}$. By the properties of a normal variable, the linear transformation of a normal variable results also in a normal variable. Given that $\ln(N(0))$ and T are constants, then $\ln(N(0)) + T \cdot S_t$ is also a normal variable with mean $\ln(N(0)) + T \cdot \mathbf{E}[X]$ and standard deviation $T \cdot \frac{\mathbf{SD}[X]}{\sqrt{T}}$. Substituting back the $X(t) = \ln(1 + g(t))$ into S_T, we show that in (5.5) the $\ln N(T)$ follows a normal distribution with mean and standard deviation as follows

$$\begin{aligned} \mathbf{E}[\ln(N(T))] &= \ln(N(0)) + T \cdot \mathbf{E}[\ln(1 + g)], \\ \mathbf{SD}[\ln(N(T))] &= \sqrt{T} \cdot \mathbf{SD}[\ln(1 + g)]. \end{aligned} \tag{5.6}$$

This implies that $N(T)$ follows the *log-normal distribution*, whose location parameter is defined as $\mu_{LN} = \mathbf{E}[\ln(N(T))]$ and scale parameter is $\sigma_{LN} = \mathbf{SD}[\ln(N(T))]$.

In Section 5.2.2 we show the closed form solution of the location and scale parameters by solving $\mathbf{E}[\ln(1 + g)]$ and $\mathbf{SD}[\ln(1 + g)]$. Additionally, we test the multiplicative growth model (5.3) and are able to reproduce the standard log-normal distribution with $\mu = 0$ and $\sigma = 1$.

5.2.2 Multiplicative Growth Model with Constant Growth Rate

Closed form solution of the parameters of log-normal distribution We have shown in Section 5.2.1 that the law of proportionate effect is governed by the following rule

$$N(t) = N(t-1) \cdot (1 + g(t)), \tag{5.7}$$

where $\{g(t)\}_{1 \leq t \leq T}$ is a set of mutually independent and identically distributed random variables with mean $\mathbf{E}[g]$ and standard deviation $\mathbf{SD}[g]$, and t is a natural positive number, $t \in \mathbb{N}^+$. Next, we have shown that after T iterations of (5.7), $N(T)$ follows the *log-normal distribution*, whose location and scale parameters are as follows

$$\begin{aligned} \mu_{LN} &= \ln(N(0)) + T \cdot \mathbf{E}[\ln(1+g)], \\ \sigma_{LN} &= \sqrt{T} \cdot \mathbf{SD}[\ln(1+g)]. \end{aligned} \tag{5.8}$$

In this section, we find the closed form solution of μ_{LN} and that of σ_{LN} by finding the expected value and the standard deviation of the function $\ln(1+g)$.

The expected value of a measurable function of some random variable X, $a(x)$, given that X has a probability density function $f(X)$, is given by the the inner product of f and a

$$\mathbf{E}[a(X)] = \int_{-\infty}^{\infty} a(x) f(x) dx. \tag{5.9}$$

In probability theory, it is possible to approximate the moments of a function a of a random variable X using Taylor expansions provided that a is differentiable and that the moments of X are finite. This allows us to not solve the integral in (5.9) and use the following approximations

$$\begin{aligned} \mathbf{E}[a(X)] &\approx a(\mu_X) + \frac{a''(\mu_X)}{2} \sigma_X^2, \\ \mathbf{SD}[a(X)] &\approx a'(\mu_X) \sigma_X, \end{aligned} \tag{5.10}$$

where $\mu_X = \mathbf{E}[X]$ and $\sigma_X = \mathbf{SD}[X]$.

If $X = g$, then $f(X) = f(g)$ is the probability density function of the distribution of the random variable g with the expected value $\mu_X = \mathbf{E}[g]$ and the standard deviation $\sigma_X = \mathbf{SD}[g]$. If $X = g$ and $a(X) = \ln(1+X)$, then $a(X) = a(g) = \ln(1+g)$ and $a(\mu_X) = \ln(1+\mu_X) = \ln(1+\mathbf{E}[g])$. We substitute $X = g$ in (5.10) and keep μ_X and σ_X for readability and get the moments of the distribution of the function of g, $a(g) = \ln(1+g)$:

$$\begin{aligned} \mathbf{E}[\ln(1+g)] &\approx \ln(1+\mu_X) - \frac{\sigma_X^2}{2(1+\mu_X)^2}, \\ \mathbf{SD}[\ln(1+g)] &\approx \frac{\sigma_X}{1+\mu_X}. \end{aligned} \tag{5.11}$$

Finally, we subsitute (5.11) into (5.8) and obtain the closed form solution of the parameters of the log-normal distribution of $N(T)$ generated by the law of proportionate effect with the growth rate g of $\mathbf{E}[g] = \mu_X$ and $\mathbf{SD}[g] = \sigma_X$

$$\begin{aligned} \mu_{LN} &= T \cdot (\ln(1+\mu_X) - \frac{\sigma_X^2}{2(1+\mu_X)^2}) + \ln(N(0)), \\ \sigma_{LN} &= \sqrt{T} \cdot (\frac{\sigma_X}{1+\mu_X}). \end{aligned} \tag{5.12}$$

Closed form solution of the 1st and 2nd moments of growth rate If we are interested to reproduce the log-normal distribution governed by the law of proportionate effect (5.7) and described by location μ_{LN} and scale σ_{LN}, we need to find the analytical solution of the first and second moments of the growth rate g. We derive the form of the mean μ_X and standard deviation σ_X from (5.12) by solving the system of equations

$$\begin{cases} T \cdot \left(\ln(1+\mu_X) - \frac{\sigma_X^2}{2(1+\mu_X)^2}\right) + \ln(N(0)) & = \mu_{LN}, \\ \sqrt{T} \cdot \left(\frac{\sigma_X}{1+\mu_X}\right) & = \sigma_{LN}. \end{cases}$$

Substituting $\frac{\sigma_X}{1+\mu_X} = \frac{\sigma_{LN}}{\sqrt{T}}$ into the first equation, we get

$$\begin{aligned} T \cdot (\ln(1+\mu_X) - \frac{\sigma_{LN}^2}{2T}) + \ln(N(0)) &= \mu_{LN}, \\ \ln(1+\mu_X) &= \frac{\mu_{LN} - \ln(N(0))}{T} + \frac{\sigma_{LN}^2}{2T}, \\ \mu_X &= N(0)^{-\frac{1}{T}} \cdot e^{\frac{2\mu_{LN}+\sigma_{LN}^2}{2T}} - 1. \end{aligned}$$

Let $a = N(0)^{-\frac{1}{T}} \cdot e^{\frac{2\mu_{LN}+\sigma_{LN}^2}{2T}}$, then the first and second moments of the growth rate are

$$\begin{aligned} \mu_X &= a - 1, \\ \sigma_X &= a \cdot \frac{\sigma_{LN}}{\sqrt{T}}. \end{aligned} \tag{5.13}$$

Model of the standard log-normal distribution We test the model (5.7) to reproduce the standard log-normal distribution, *i.e.* the one with location $\mu_{LN} = 0$ and scale $\sigma_{LN} = 1$, $\ln(N \sim \ln\mathcal{N}(0,1)$. To instantiate the growth rate, we use the derived closed form solution of the growth rate moments (5.13), which results in $\mu_X = e^{\frac{1}{2T}} - 1$ and $\sigma_X = \frac{1}{\sqrt{T}} \cdot e^{\frac{1}{2T}}$. Therefore, the final model goes as follows

$$\begin{aligned} &N(t) = N(t-1) \cdot (1+g(t)), \text{ where } \mathbf{E}[g] = e^{\frac{1}{2T}} - 1 \text{ and } \mathbf{SD}[g] = \frac{1}{\sqrt{T}} \cdot e^{\frac{1}{2T}} \\ &N(0) = 1, \text{ the initial value, and } t \in [1, T]. \end{aligned} \tag{5.14}$$

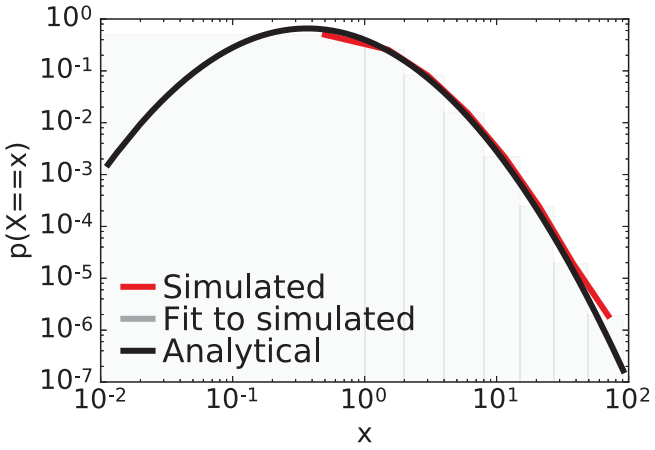


Figure 5.2: Fit of simulated results to the log-normal distribution.

$\ln \mathcal{N}(\mu, \sigma)$	KS (D, p)
$\ln \mathcal{N}(0.002, 1)$	$D = 0.005$, $p = 0.154$

Table 5.2: Parameters of the log-normal fit of the simulated results. Kolmogorov-Smirnov test between the simulated results and the analytical log-normal distribution with $\mu = 0$ and $\sigma = 1$.

We have run the model (5.14) 1000 time steps with 100 observations. Table 5.2 and Figure 5.2 show the fit of the simulated results to the log-normal distribution and the Kolmogorov-Smirnov test to the standard log-normal distribution $\ln \mathcal{N}(0, 1)$. We observe that the model (5.14) recovers the location and scale parameters, $(0, 1)$, and the K-S test confirms that the simulated results follow the log-normal distribution of $(0, 1)$ parameters yielding the distance statistic of ≤ 0.1 and the p-value ≥ 0.1.

5.3 Growth Model of Evaluations on YouTube

In Section 5.2.1 we have presented the multiplicative growth model, based on the law of proportionate effect, that generates the log-normal distribution. In Chapter 4, we have observed that online evaluations follow the log-normal distribution on an aggregated level. Therefore, in this section we will use the introduced dynamics (5.3) to explain the growth of `likes` and that of `dislikes` on `YouTube` platform.

Dataset description Our crawl was able to obtain the time series of `likes` and that of `dislikes` of 545,447 videos. However, the amount of the consecutive time series points is not constant among the videos. Therefore, we have reduced the videos to the subset of those that span the same amount of days. This has yielded 354 videos each with the 100 days points since the upload date. Figure 5.3 shows the total number of evaluations of 354 videos over 100 days. The lines are shaded according to the final number of `likes`

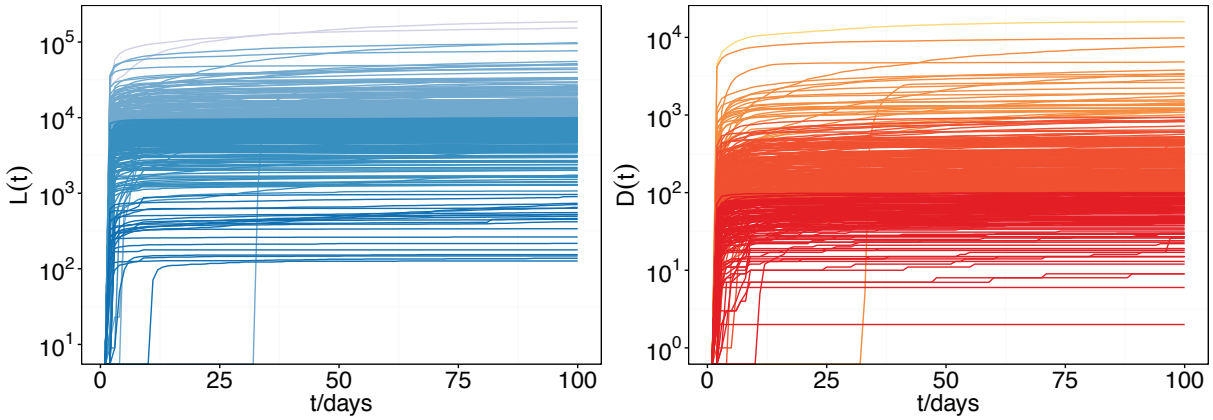


Figure 5.3: The growth of the number of `likes` and `dislikes` for the sample of videos. The lines are shaded according to the total number of `likes` (`dislikes`) that the video received at the end of the collection period. Light to dark colour corresponds to the following categories: $10^6 < N(t)$, $10^5 < N(t) \leq 10^6$, $10^4 < N(t) \leq 10^5$, $10^3 < N(t) \leq 10^4$ and $10^2 < N(t) \leq 1$, where $N(t)$ is the number of `likes` or `dislikes` respectively.

(`dislikes`) at the 100th day. We can easily observe that videos obtain the high number of evaluations very shortly since the upload, and a few days after the upload date the growth of evaluations slows down sharply. This observation is a first evidence that online evaluations do not grow in a constant way.

From the dynamics (5.3), we first derive the formulation of the growth rate. Let $L(t-1)$ be the number of `likes` to a video at time $t-1$, and $D(t-1)$ – that of `dislikes`. Then our assumption of the multiplicative growth implies that the number of `likes` and that of `dislikes` at the next interval of time is determined by:

$$L(t) = L(t-1)(1+g_L(t)), \qquad (5.15) \qquad D(t) = D(t-1)(1+g_D(t)), \qquad (5.16)$$

where $g_L(t)/g_D(t)$ is the (relative) growth rate of the number of `likes` and that of `dislikes` at time t respectively. From (5.15) and (5.16) we obtain the formulation of the (relative) growth rate of `likes` and that of `dislikes` respectively:

$$g_L(t) = \frac{L(t) - L(t-1)}{L(t-1)}, \qquad (5.17) \qquad g_D(t) = \frac{D(t) - D(t-1)}{D(t-1)}. \qquad (5.18)$$

Time T represents the age of the video since its upload, therefore, $L(T)$ and $D(T)$ are the number of `likes` and that of `dislikes` of the video of age T. The growth rate shows how many new `likes` (`dislikes`) the video gets at time t relative to the absolute number of `likes` (`dislikes`) at the previous time step $t-1$.

The law of proportionate effect is based on the assumption that the growth rates are mutually independent and *identically* distributed random variables. We check whether this assumption holds for our data a) by computing the time series of the growth rate of `likes` and that of `dislikes` and b) by computing the distribution of the growth rate at each time step.

5.3.1 Time Series and Statistical Analysis of the Growth Rate

For the time series analysis of the growth rate, we use the mean relative growth rate. In (5.17) and (5.18), the growth rate is calculated for one observation. We define the mean relative growth rate $\langle g_L(t)\rangle$ for N observations as follows:

$$\langle g_L(t)\rangle = \frac{\langle L_i(t)\rangle - \langle L_i(t-1)\rangle}{\langle L_i(t-1)\rangle}, \tag{5.19}$$

where $\langle L_i(t)\rangle$ and $\langle L_i(t-1)\rangle$ is the mean number of `likes` at time t and time $t-1$ respectively, $i \in [1, N]$ and $N = 354$. In the same way, we define the mean relative growth rate of `dislikes` $\langle g_D(t)\rangle$. In Figure 5.4 we show the time series of the mean relative growth rate of `likes` and that of `dislikes`. Our main observation is that the growth rate is not constant at different ages of the video. The growth rate is maximum (close to 1) at the early ages of the video and decays sharply as the video continues to exist on `YouTube`. Namely, 10 days after the upload date, the growth rate reduces to 1% of its initial value. Such an early rapid growth and a following significant decay is a signature of the "collective attention" decay suggested by Wu and Huberman (2007), which occurs for different reasons, for example retention of a video in a local network, the lack of viral spread, or reaching the system size limit (Yasseri *et al.*, 2013).

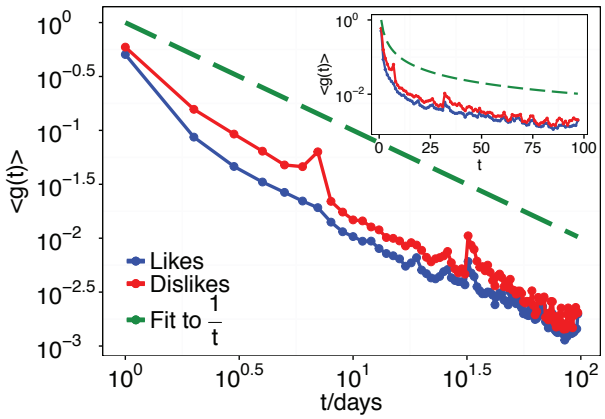


Figure 5.4: The time series of the relative growth rate of `likes` and that of `dislikes`, $\langle g_L(t)\rangle$ and $\langle g_D(t)\rangle$. Base figure is a log-log plot, and inset figure is a semi-log plot, where x-axis is plotted on a linear scale and y-axis is on a logarithmic scale.

In Figure 5.5, we show the fits of the time series to various functions, and parameters and the goodness of the fits are displayed in Table 5.3. A power function $\frac{1}{t^\beta}$ with negative exponent ≈ -1 is seen to be the best fit to the growth rate time series of both `likes` and `dislikes`. This result goes inline with previous study by Asur *et al.* (2011), where powerlaw decay with exponent -1 has been obsered in attention decay to Twitter trends.

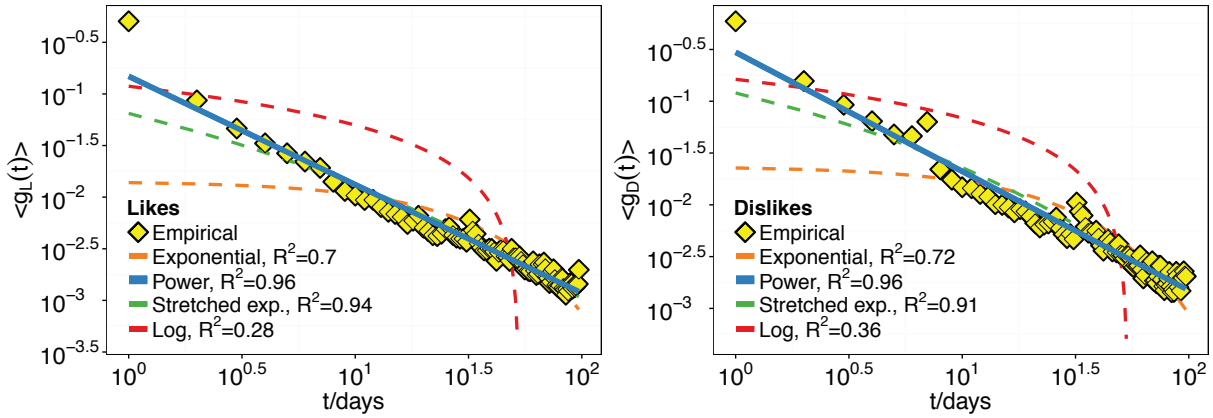


Figure 5.5: The fitting of the relative growth rate of `likes` and that of `dislikes` to various functions. The maximum R^2 next to the fitted function indicates that the best fitting function of both $\langle g_L(t)\rangle$ and $\langle g_D(t)\rangle$ is the power function: $g(t) \sim \frac{1}{t}$.

Function type	Fit of $\langle g_L(t)\rangle$ to function	R^2	Fit of $\langle g_D(t)\rangle$ to function	R^2
Exponential	$e^{-4.25-0.03\cdot t}$	0.7	$e^{-3.75-0.03\cdot t}$	0.72
Power	$\mathbf{e^{-1.91}\cdot t^{-1.05}}$	**0.96**	$\mathbf{e^{-1.21}\cdot t^{-1.15}}$	**0.96**
Stretched exp.	$e^{-e^{1.01}\cdot t^{0.2}}$	0.94	$e^{-e^{0.75}\cdot t^{0.25}}$	0.91
Logarithmic	$0.12-0.03\cdot\ln(t)$	0.28	$0.16-0.04\cdot\ln(t)$	0.36

Table 5.3: The coefficients of the fitted functions of the relative growth rate of `likes` and that of `dislikes`.

Analysis of the growth rate distribution We have shown that on the aggregated level the expected value of the growth rate of `likes` and that of `dislikes` decays over time. We obtain individual differences between the growth rates by analysing the growth rate distribution at every time step $P(g(t))$. Preliminary analysis showed that the growth rates are right-skewed, therefore in Figure 5.6 we plot the density distribution of the logarithm of the $g(t)$, $P(\ln(g(t)))$, of every 10th day. Dashed vertical line marks the distribution mean and the solid line projects the time series of the distribution mean on a 2-dimensional scale. We can easily realize that over time the distribution mean of both growth rates decreases. However, the distribution range of `likes` shrinks faster than that of `dislikes`, which reveals that as video exists on `YouTube` platform, the rate of new `dislikes` is higher than that of new `likes`.

Given that growth rates are both highly skewed, we fit $g_L(t)$ and $g_D(t)$, where $t \in [1, 100]$, to known heavy-tailed and exponential statistical distributions: log-normal, powerlaw, stretched exponential and exponential, and we assess the goodness of the fit via the log-likelihood ratio and its p-value using the `powerlaw` python package (Alstott *et al.*, 2014). Figure 5.7 shows a sample of the probability density plots of $g_L(t)$ and $g_D(t)$ for six selected days. Log-normal distribution has been identified as the best fit for majority of time steps t for both growth rate distributions of `likes` and that of `dislikes`, see Table 5.4.

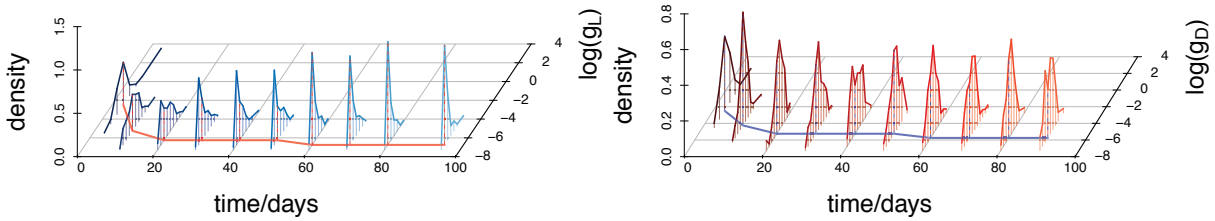


Figure 5.6: Distribution of growth rates at selected days. Dashed vertical line indicates the distribution mean, while the solid line marks the time series of the distribution mean on a 2-dimensional scale.

Day 3	Likes		Dislikes	
	R	p-value	R	p-value
$\ln(\frac{L_{LN}}{L_{PL}})$	68.96	0.00	176.97	0.00
$\ln(\frac{L_{LN}}{L_{SEXP}})$	279.62	0.02	239.03	0.01
$\ln(\frac{L_{LN}}{L_{EXP}})$	455.91	0.02	190.76	0.01
$\ln(\frac{L_{PL}}{L_{SEXP}})$	210.66	0.11	62.06	0.56
$\ln(\frac{L_{PL}}{L_{EXP}})$	386.95	0.07	13.80	0.88
$\ln(\frac{L_{SEXP}}{L_{EXP}})$	176.29	0.00	-48.27	0.00

Day 5	Likes		Dislikes	
	R	p-value	R	p-value
$\ln(\frac{L_{LN}}{L_{PL}})$	223.88	0.00	156.46	0.00
$\ln(\frac{L_{LN}}{L_{SEXP}})$	9666.92	0.30	87.25	0.09
$\ln(\frac{L_{LN}}{L_{EXP}})$	359.09	0.25	35.17	0.06
$\ln(\frac{L_{PL}}{L_{SEXP}})$	9443.04	0.32	-69.20	0.27
$\ln(\frac{L_{PL}}{L_{EXP}})$	135.20	0.70	-121.29	0.00
$\ln(\frac{L_{SEXP}}{L_{EXP}})$	-9307.84	0.00	-52.09	0.00

Day 7	Likes		Dislikes	
	R	p-value	R	p-value
$\ln(\frac{L_{LN}}{L_{PL}})$	246.35	0.00	99.22	0.00
$\ln(\frac{L_{LN}}{L_{SEXP}})$	340.78	0.26	273.88	0.25
$\ln(\frac{L_{LN}}{L_{EXP}})$	66.44	0.19	127.95	0.25
$\ln(\frac{L_{PL}}{L_{SEXP}})$	94.43	0.76	174.67	0.50
$\ln(\frac{L_{PL}}{L_{EXP}})$	-179.90	0.01	28.74	0.82
$\ln(\frac{L_{SEXP}}{L_{EXP}})$	-274.34	0.00	-145.93	0.00

Day 10	Likes		Dislikes	
	R	p-value	R	p-value
$\ln(\frac{L_{LN}}{L_{PL}})$	155.26	0.00	104.82	0.00
$\ln(\frac{L_{LN}}{L_{SEXP}})$	124.28	0.18	49.53	0.17
$\ln(\frac{L_{LN}}{L_{EXP}})$	35.09	0.22	16.57	0.12
$\ln(\frac{L_{PL}}{L_{SEXP}})$	-30.98	0.76	-55.29	0.21
$\ln(\frac{L_{PL}}{L_{EXP}})$	-120.17	0.00	-88.25	0.00
$\ln(\frac{L_{SEXP}}{L_{EXP}})$	-89.19	0.00	-32.96	0.00

Day 20	Likes		Dislikes	
	R	p-value	R	p-value
$\ln(\frac{L_{LN}}{L_{PL}})$	25.52	0.00	68.32	0.00
$\ln(\frac{L_{LN}}{L_{SEXP}})$	-2.29	0.00	6.26	0.16
$\ln(\frac{L_{LN}}{L_{EXP}})$	17.43	0.33	3.30	0.20
$\ln(\frac{L_{PL}}{L_{SEXP}})$	-27.81	0.00	-62.06	0.00
$\ln(\frac{L_{PL}}{L_{EXP}})$	-8.09	0.66	-65.02	0.00
$\ln(\frac{L_{SEXP}}{L_{EXP}})$	19.72	0.00	-2.96	0.01

Day 50	Likes		Dislikes	
	R	p-value	R	p-value
$\ln(\frac{L_{LN}}{L_{PL}})$	1.03	0.85	50.15	0.00
$\ln(\frac{L_{LN}}{L_{SEXP}})$	12.65	0.01	6.78	0.07
$\ln(\frac{L_{LN}}{L_{EXP}})$	49.92	0.00	2.89	0.32
$\ln(\frac{L_{PL}}{L_{SEXP}})$	11.61	0.24	-43.38	0.00
$\ln(\frac{L_{PL}}{L_{EXP}})$	48.89	0.02	-47.26	0.00
$\ln(\frac{L_{SEXP}}{L_{EXP}})$	37.27	0.00	-3.88	0.01

Table 5.4: Statistical analysis of the growth rate. The sign of the log-likelihood ratio (column 1) between the two candidate distribution tells the direction of the winning distribution, while the value of the ratio indicates the strength of the "win". If the ratio is positive, then the distribution in the numerator is the winning one. The significance of the ratio test is assessed via the p-value. In our test, the likelihood of the log-normal distribution is always in the numerator, and mostly all the ratios are positive. This allows us to confirm that the log-normal distribution is the best fitting distribution of the growth rate.

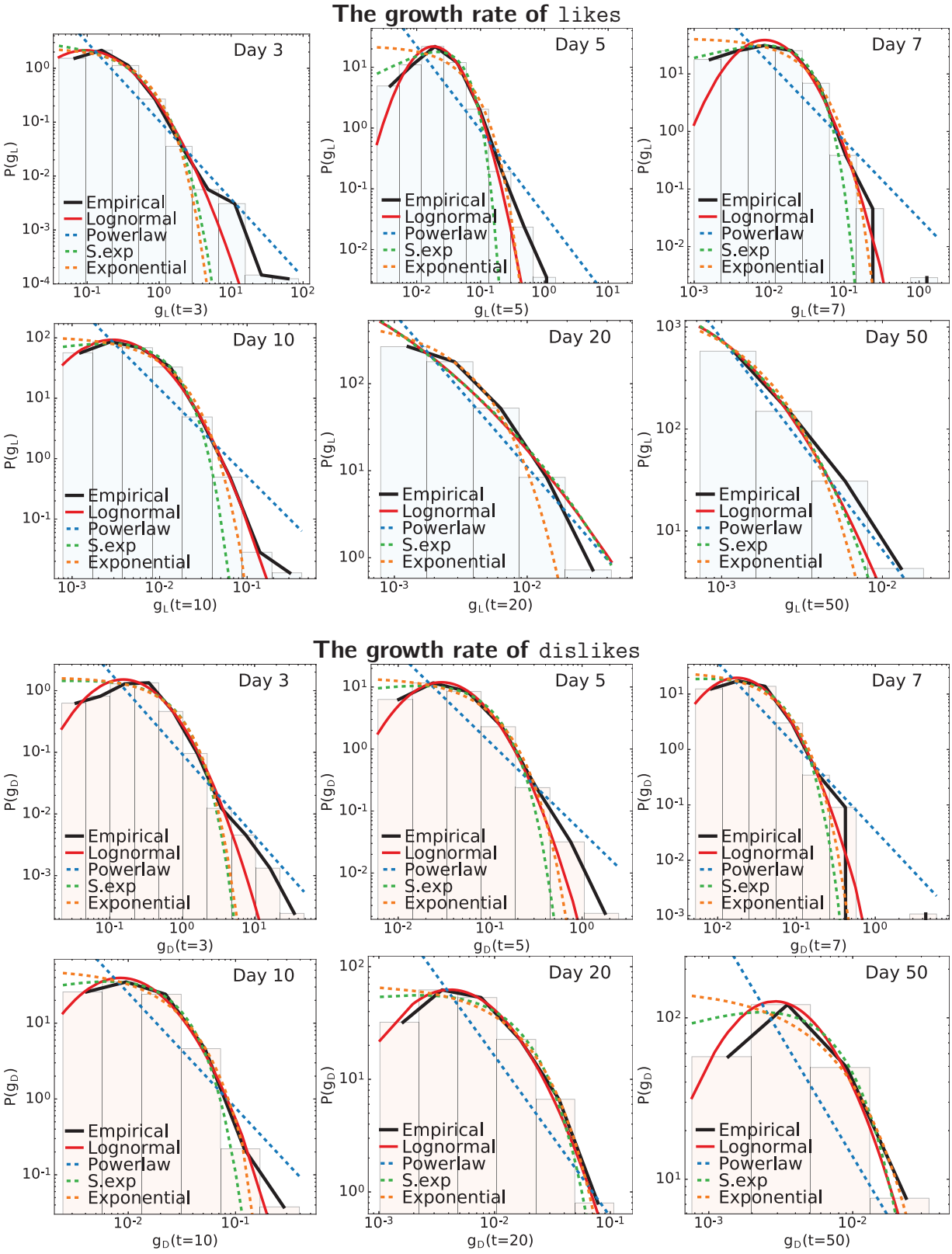


Figure 5.7: Distribution of the relative growth rate of `likes`, $P(g_L(t))$, and that of `dislikes`, $P(g_D(t))$, at a sample of days $t \in [3, 5, 7, 10, 20, 50]$. At each time step t, the growth rate follows the log-normal distribution: $g_L(t) \sim \ln\mathcal{N}(\mu_L(t), \sigma_L(t))$ and $g_D(t) \sim \ln\mathcal{N}(\mu_D(t), \sigma_D(t))$.

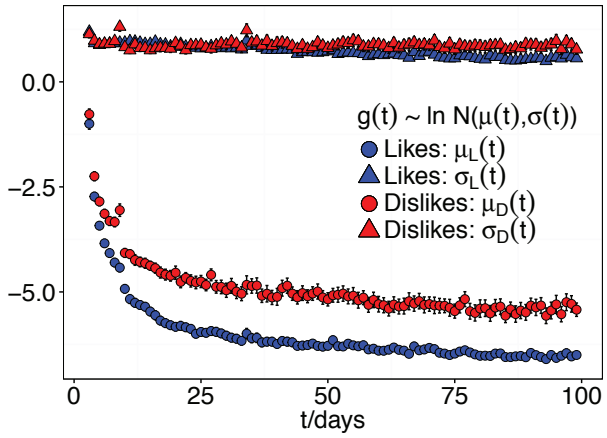


Figure 5.8: The time series of the parameters of the log-normal distribution of the growth rate of `likes` and that of `dislikes`. The parameters μ_L and μ_D decay, while the parameters σ_L and σ_D are constant, $\langle\sigma_L\rangle = 0.72$ and $\langle\sigma_D\rangle = 0.88$.

Function type	Fit of $\mu_L(t)$ to function	R^2	Fit of $\mu_D(t)$ to function	R^2
Exponential	$-e^{1.545+0.004\cdot t}$	0.29	$-e^{1.313+0.005\cdot t}$	0.34
Power	$-e^{0.964}\cdot t^{0.219}$	0.58	$-e^{0.692}\cdot t^{0.238}$	0.63
Stretched exp.	$-e^{e^{0.2}\cdot t^{0.102}}$	0.7	$-e^{e^{-0.024}\cdot t^{0.129}}$	0.75
Logarithmic	$\mathbf{-2.478 - 0.953\cdot \ln(t)}$	**0.8**	$\mathbf{-1.746 - 0.855\cdot \ln(t)}$	**0.84**

Table 5.5: The coefficients of the fitted functions of the μ parameter of log-normal growth rate of `likes` and that of `dislikes`.

Given that the growth rates follow the log-normal distribution at different time steps, we plot in Figure 5.8 the time series of the parameters μ and σ of the log-normal distribution of the growth rate. We observe that parameters $\mu_L(t)$ and $\mu_D(t)$ of both `likes` and `dislikes` are not constant and decay with time, while parameters $\sigma_L(t)$ and $\sigma_D(t)$ are constant and fluctuate around 0.72 and 0.88 for `likes` and `dislikes` respectively. We determine the functional form of parameter μ with time as an independent variable, see Table 5.5 and Figure 5.9. The logarithmic function, $\mu(t) \sim \ln(t)$, is identified as the best fit, yielding the maximum adjusted coefficient determination R^2.

Furthermore, we can confirm analytically that if the growth rate is log-normally distributed at each time step, $g(t) \sim \ln\mathcal{N}(\mu(t), \sigma(t))$, and the expected value of the growth rate follows the power function, $\mathbf{E}[g(t)] \sim \alpha \cdot \frac{1}{t^\beta}$, (Figure 5.5), then only the logarithmic function $\mu(t) \sim \ln(t)$ is the function of the μ parameter of the growth rate:

$$\left.\begin{aligned} &g(t) \sim \ln\mathcal{N}(\mu(t), \sigma(t)),\\ &\mathbf{E}[g(t)] = e^{\mu(t)+\frac{\sigma(t)^2}{2}},\\ &\mathbf{E}[g(t)] = \alpha \cdot \tfrac{1}{t^\beta}, \end{aligned}\right\} \text{, therefore } \mu(t) = (\ln\alpha - \frac{\sigma(t)^2}{2}) - \beta\cdot\ln t$$

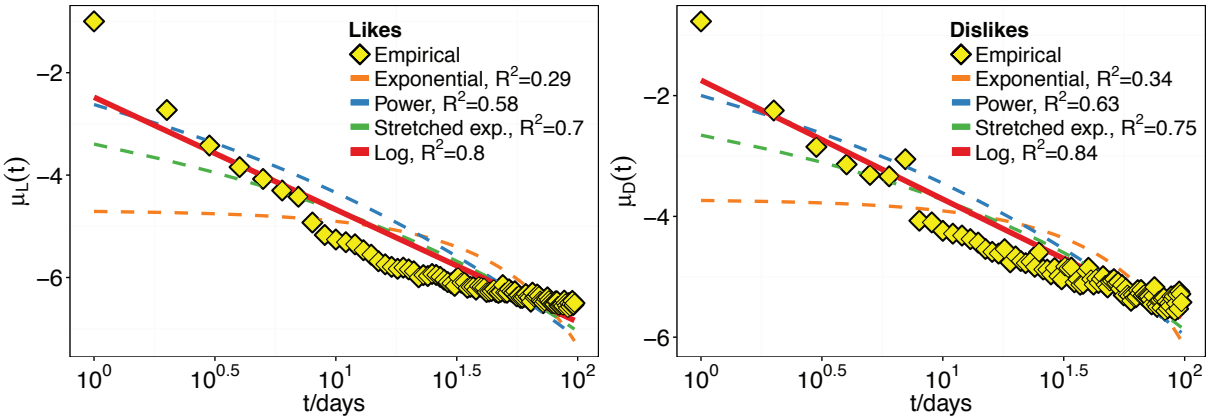


Figure 5.9: The fitting of the parameter $\mu(t)$ of `likes` (left) and that of `dislikes` (right) to various functions. The figures a semi-log plot, where x-axis is plotted on a linear scale and y-axis is on a logarithmic scale; the lin-lin plot is shown in Figure 5.8. The maximum R^2 next to the fitted function indicates that the best fitting function of both $\mu_L(t)$ and $\mu_D(t)$ is the logarithmic function: $\mu(t) \sim \ln(t)$.

To this point, we observe in our dataset that the growth rate of evaluations that leads to the log-normal distribution of evaluations under multiplicative growth model is itself log-normally distributed with decaying expected value. Obtained empirical finding is contrary to previous work where the growth rate has been shown to have constant expected value at every time step. In order to use the multiplicative growth dynamics (5.3), we need to show that the non-constant growth rate of evaluations also leads to the log-normal distribution of evaluations.

5.3.2 Multiplicative Growth Model with Non-constant Growth Rate

We have observed that the growth rate $g(t)$ is not constant at each time step t. Wu and Huberman (2007) captures the decay in attention by introducing the decay factor $r(t)$. Decay is time-dependent and consists of a series of decreasing positive numbers with the property that $r(t) = 1$ and $r(t) \to 0$, when $t \to \infty$. With this additional parameter, the full stochastic dynamics of the video evaluations dynamics is governed by

$$\begin{aligned} N(t) &= N(t-1)(1 + r(t)g(t)), \text{ or} \\ N(t) &= N(t-1)(1 + g^*(t)), \end{aligned} \tag{5.20}$$

where $g^*(t) = r(t)g(t)$.

With this formulation, the growth rate $g^*(t)$ is independent but *not identically distributed* compared to $g(t)$. At each time step, $g^*(t)$ is sampled from the distribution with finite mean and variance, but is not correlated between different time steps.

When $t \to T$ and T is large, we obtain

$$\begin{aligned} N(T) &= N(0)\prod_{t=1}^{T}(1+g^*(t)), \\ \ln(N(T)) &= \ln(N(0)) + \sum_{t=1}^{T}\ln(1+g^*(t)), \text{ or} \\ \ln(N(T)) &= \ln(N(0)) + \sum_{t=1}^{T}X(t), \end{aligned} \tag{5.21}$$

where $X(t) = \ln(1+g^*(t))$.

Lyapunov Central Limit Theorem, see Appendix A.1, states that if a sequence of random variables $\{X_1, X_2, \dots\}$ are independent *but not necessarily identically distributed*, each with finite expected value μ_i and variance σ_i^2, then if $s_n^2 = \sum_{i=1}^{n}\sigma_i^2$, we get that

$$\frac{1}{s_n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{d} \mathcal{N}(0,1). \tag{5.22}$$

From (5.22) it follows that the sum of random variables X_i is normally distributed

$$\sum_{i=1}^{n}X_i \simeq \mathcal{N}(\sum_{i=1}^{n}\mu_i, \sum_{i=1}^{n}\sigma_i^2). \tag{5.23}$$

Substituting $X(t)$ for X_i, we get

$$\sum_{t=1}^{T}X(t) \xrightarrow{d} \mathcal{N}(\sum_{t=1}^{T}\mathbf{E}[X(t)], \sum_{t=1}^{T}\mathbf{VAR}[X(t)]), \tag{5.24}$$

where each of $X(t)$ has expected value $\mathbf{E}[X(t)]$ and variance $\mathbf{VAR}[X(t)]$.

Substituting back $X(t) = \ln(1+g^*(t))$ in (5.24), we obtain

$$\sum_{t=1}^{T}\ln(1+g^*(t)) \xrightarrow{d} \mathcal{N}(\sum_{t=1}^{T}\mathbf{E}[\ln(1+g^*(t))], \sum_{t=1}^{T}\mathbf{VAR}[\ln(1+g^*(t))]). \tag{5.25}$$

If $\mu^* = \sum_{t=1}^{T}\mathbf{E}[\ln(1+g^*(t))]$ and $\sigma^{*2} = \sum_{t=1}^{T}\mathbf{VAR}[\ln(1+g^*(t))]$, then

$$\sum_{t=1}^{T}\ln(1+g^*(t)) \xrightarrow{d} \mathcal{N}(\mu^*, \sigma^{*2}). \tag{5.26}$$

Substituting (5.26) in (5.21), we obtain that the distribution of the log-transformed number of observations at time T is normally distributed

$$\ln(N(T)) \xrightarrow{d} \mathcal{N}(\ln(N(0)) + \mu^*, \sigma^{*2}). \tag{5.27}$$

Therefore, we have shown that the the number of observations $N(T)$ is log-normally distributed and the location and scale parameters are as follows

$$\begin{aligned} \text{location } \mu_{LN} \quad & \ln(N(0)) + \sum_{t=1}^{T} \mathbf{E}[\ln(1 + g^*(t))], \\ \text{scale } \sigma_{LN} \quad & \sqrt{\sum_{t=1}^{T} \mathbf{VAR}[\ln(1 + g^*(t))]}. \end{aligned} \tag{5.28}$$

We have analytically confirmed that log-normally distributed growth rate with different finite expected mean and standard deviation is able to give rise to the log-normal distribution under the law of proportionate effect. Our next step is to simulate the multiplicative growth model with the decay in the collective attention to the video, where the decay is empirically obtained from the decay in the growth rate of evaluations.

5.3.3 Simulation

Simulation model We have observed that the decay in the log-normally distributed growth rate is driven by the decreasing μ parameter. Therefore, we use the values of the fitted function of μ and keep σ constant when sampling the growth rate in our modified data-driven multiplicative growth model (5.3). The update rule for the number of evaluations goes as follows:

The new number of `likes` at the time step t, ΔL_t, is a fraction $g_L(t)$ of the number of `likes` at time step $t-1$, such that the fraction is determined by the log-normal distribution with the following parameters:

$$\begin{aligned} &L(t) = L(t-1)(1 + g_L(t)), \text{ or} \\ &L(t) = L(t-1) + g_L(t) \cdot L(t-1) = L(t-1) + \Delta L_t, \text{ where} \\ &g_L(t) \sim \ln\mathcal{N}(\mu_L(t), \langle\sigma_L\rangle), \text{ such that} \quad \mu_L(t) = -2.48 - 0.95 \cdot \ln t, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \langle\sigma_L\rangle = 0.72, \\ &L(0) \sim P(L(t=1)), \text{ sampling the initial value from their respective distribution,} \end{aligned}$$

and similarly, the number of `dislikes` at time t is determined by the following model:

$D(t) = D(t-1)(1 + g_D(t))$, or
$D(t) = D(t-1) + g_D(t) \cdot D(t-1) = D(t-1) + \Delta D_t$, where
$g_D(t) \sim \ln\mathcal{N}(\mu_D(t), \langle\sigma_D\rangle)$, such that $\mu_D(t) = -1.75 - 0.86 \cdot \ln t$,
$\langle\sigma_D\rangle = 0.88$,
$D(0) \sim P(D(t=1))$, sampling the initial value from their respective distribution,

where location parameters $\mu_L(t)$ and $\mu_D(t)$ are time-dependent and follow the fitted logarithmic function $\mu(t) = -\alpha - \beta \cdot \ln(t)$, and scale parameters are constant and have the following values: $\langle\sigma_L\rangle = 0.72$ and $\langle\sigma_D\rangle = 0.88$. And, the initial values are sampled from their respective distribution. Figure 5.10 shows the histogram of the frequency distribution of initial values of `likes` and that of `dislikes`. Both distributions are right-skewed. The minimum initial value of `dislikes` is 1 and that of `likes` is 9.

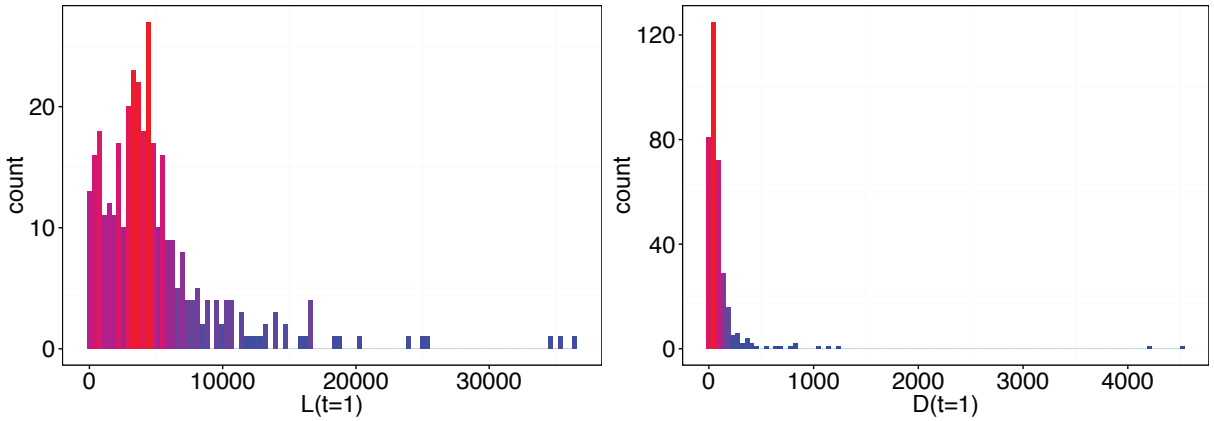


Figure 5.10: The histogram of the initial values of `likes` and that of `dislikes`. Zeros are removed. Range of `likes` is $[9, 36444]$ and that of `dislikes` is $[1, 4524]$.

Simulation results We simulate the above models 100 times corresponding to 100 day length of the empirical time series and with 354 items coresponding to 354 crawled videos. Figure 5.11 depicts the distribution of the number of evaluations obtained via the model simulations in red colour and that of the empirical values in blue. Dashed lines denote the fit to the log-normal distribution.

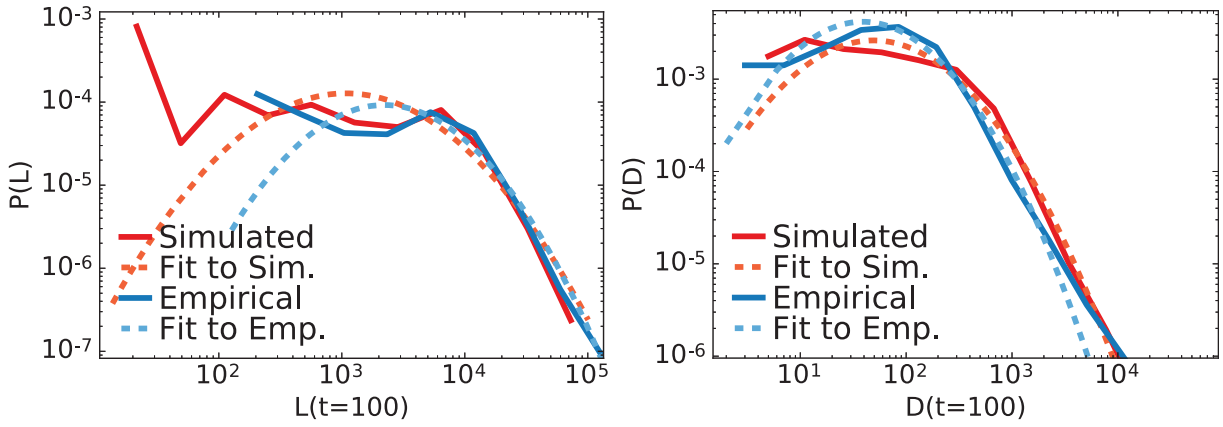


Figure 5.11: Distribution of the number of `likes` and that of `dislikes`. Empirical observations are shown in blue, the simulated values are plotted in red, and the fits to the log-normal distribution are displayed with a dashed line.

We observe that simulated results reproduce the range of empirical values and the log-normal shape. However, the simulated number of `likes` recovers the empirical `likes` well in the tail of the distribution, *i.e.* for large values, but differs in the head. The distribution of the number of `dislikes`, however, is not recovered.

First, we statistically assess whether the model produces the log-normal distribution of the number of `likes` and that of `dislikes`. We perform the Kolmogorov-Smirnov test of the empirical and simulated values to theoretical log-normal distribution, see Table 5.6. We observe that for both types of evaluations the D-statistic is approximately ≤ 0.1. K-S results allow us to confirm that the simulated distributions are log-normal. Indeed, the proposed multiplicative growth model returns the log-normally distributed values. However, the μ parameter of the empirical distributions is not recovered for the number of `dislikes` (5.611 vs. 5.095) and slightly differs for the number of `likes` (8.621 vs 8.879).

Type of observation		$\ln\mathcal{N}(\mu,\sigma)$	KS $(D, p\text{-value})$	
`Likes`	Simulated	$\ln\mathcal{N}(8.621, 1.279)$	$D = 0.161$	$p = 0$
	Empirical	$\ln\mathcal{N}(8.879, 1.085)$	$D = 0.125$	$p = 2.6 \cdot 10^{-5}$
`Dislikes`	Simulated	$\ln\mathcal{N}(5.611, 1.310)$	$D = 0.056$	$p = 1.361 \cdot 10^{-97}$
	Empirical	$\ln\mathcal{N}(5.095, 1.199)$	$D = 0.075$	$p = 0.035$

Table 5.6: Fit of simulated and empirical observations to the log-normal distribution.

Next, we assess the statistical similarity between the simulated and the empirical values by performing the Kolmogorov-Smirnov test, see Table 5.7. Between the simulated and empirical number of `likes`, the D-statistic is < 0.1, which indicates that we cannot reject the hypothesis that the simulated number of `likes` come from the same distribution as the empirical number of `likes`. Contrary to positive evaluations, the D-statistic of `dislikes` is $\gg 0.1$, therefore we can reject the hypothesis that the simulated and empirical number of `dislikes` come from the same distributions.

	KS $(D, p\text{-value})$	
`Likes`	$D = 0.099$	$p = 0.002$
`Dislikes`	$D = 0.289$	$p = 2.908 \cdot 10^{-26}$

Table 5.7: K-S test between the simulated results and the empirical values of `likes` and that of `dislikes`.

Finally, we evaluate the relationship between the simulated `likes` and `dislikes`. In Figure 5.12 we observe that the scatterplot of evaluations forms a cloud, and does not show any correlation. In fact, the coefficient of determination R^2 of a linear model, $\ln(\texttt{dislikes}) \sim f(\ln(\texttt{likes}))$, is almost 0, which confirms no relation between the produced `likes` and `dislikes`.

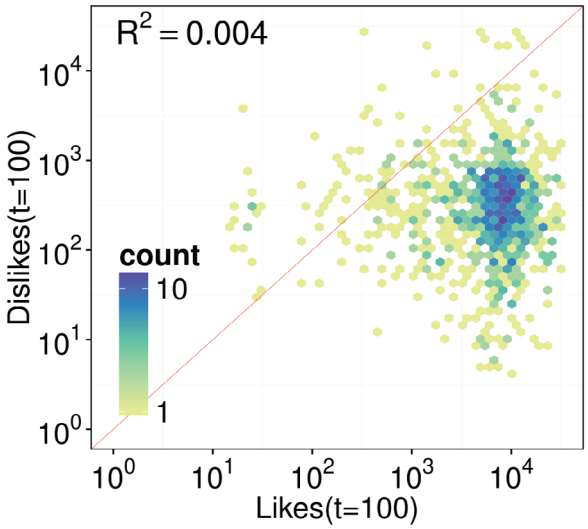


Figure 5.12: Relationship between the number of `likes` and that of `dislikes`. No dual regime in the relationship between the simulated evaluations can be observed.

5.3.4 Model Limitations

We have first introduced the multiplicative growth model based on the law of proportionate effect (5.3) that governs the emergence of the log-normal distribution. Our empirical results, however, revealed deviations from the assumptions of the original model, namely the growth rate of the `YouTube` evaluations is log-normally distributed with the decaying expected value. The early exponential growth in evaluations, followed by the significant slow down at later days after the upload date of the video, has been explained by the decay in the collective attention. To account for the empirical observations, first, we have analytically shown that not identically distributed growth rate also leads to the log-normal distribution. Second, we have identified the functional form of the parameters of the growth rate. And finally, we have devised and run simulations of the modified multiplicative growth model that considers for the decay in the collective attention.

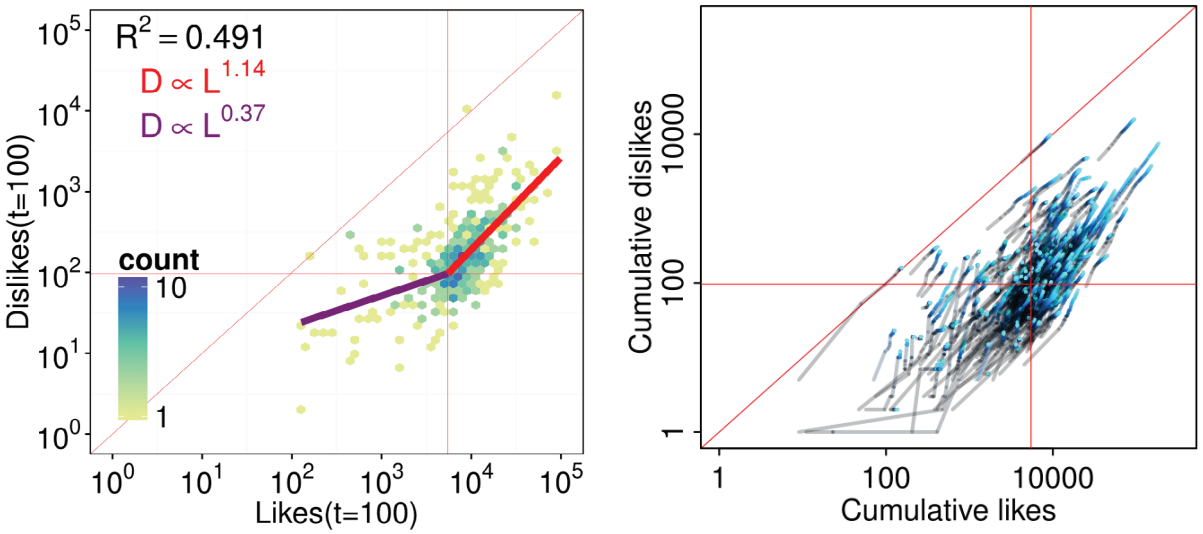
Statistically, we have confirmed that the proposed two-parametric growth model is able to explain the dynamics of positive evaluations, but not that of negative evaluations. The mismatch between the empiricial and simulated number of `dislikes` suggests that there is an additional rule that determines their growth, for instance a coupled growth with `likes`. We have statistically confirmed that the proposed multiplicative growth models do not reproduce the correlated relationship between the evaluations. The idea of a coupled growth of `likes` and `dislikes` stems from the empirical observations in Chapter 4, where we have observed that the relationship between positive and negative evaluations is non-linear. After a video reaches a threshold of the number of `likes`, the number of `dislikes` grows faster than before the threshold, indicating the burst of the positivity bubble, or the bubble of local popularity. In Section 5.4, we propose the model that accounts for the interaction between the positive and negative evaluations and results in the non-linearity between the evaluations.

5.4 Coupled Growth Model of Evaluations on YouTube

5.4.1 Analysis of the Relationship between `Likes` and `Dislikes`

We first explore the relationship between the empirical number of `likes` and that of `dislikes` in our dataset of 354 videos each with the 100-day long time series, see Figure 5.13a. Our first observation is that the number of `dislikes` are always lower in absolute values than the number of `likes` – all points are below the diagonal line. Secondly, the evaluations are strongly correlated with the Pearson correlation coefficient $\rho = 0.66$. Following the method in Chapter 4, we apply the `MARS` model (Friedman, 1991, 1993) and obtain the coefficient of determination $R^2 = 0.49$, which is higher than that of a linear model $R^2 = 0.43$. The `MARS` model returns the cutoff value in the number of `likes` at around 5000, which separates one linear piece of the relationship from another. The lower quadrant is characterized by a slower growth of `dislikes` compared to `likes`, which tells that `YouTube` users at this stage mostly like the video. The upper quadrant, however, shows that the relationship between the evaluations evens off, so that half of `YouTube` users give a `like` and another half clicks a `dislike`. Skeweness towards positivity fades.

We confirm that the non-linearity between `likes` and `dislikes` is also present in our subset of timestamped dataset of 354 videos. This empirical finding motivates us to zoom in the evaluation dynamics of each individual video. In Figure 5.13b, we reveal



(a) `MARS` model applied to the number of `likes` versus `dislikes`. Vertical and horizontal red lines are situated at the cutoff in the number of `likes` and that of `dislikes`. Colour shade of the points shows the joint frequency of evaluations.

(b) Evaluations dynamics of each video. Each segment represents the dynamics of 1 of the 354 videos. The colour shade of a segment marks the number of evaluations at a specific time: light grey denotes day 1 since upload, light blue – day 100 since upload.

Figure 5.13: Relationship between the empirical collective evaluations of a time-stamped data.

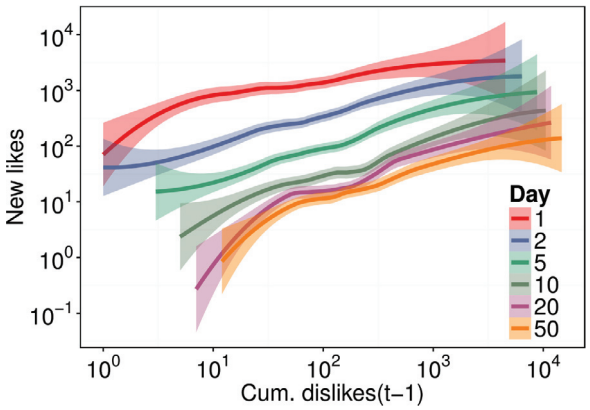
the change of the cumulative number of `likes` together with `dislikes` in time. Each segment represents the dynamics of evaluations of each of the 354 videos. Additionally, each segment is gradually shaded from light grey to blue, such that light grey denotes day 1 since the video upload and blue marks day 100 since upload. Red vertical and horizontal lines indicate the cutoff in the number of `likes` and that of `dislikes` as found by the `MARS` model. Our main observation is that dynamics of evaluations of most of the videos follow a linear relationship, namely the segments are parallel to the diagonal line. However, in the lower quadrant, one can occasionally observe vertical segments, such that the number of `likes` is constant or increases slowly and the number of `dislikes` of that video soars. We hypothesize that such dynamics is a characteristic of the filter bubble phenomenon.

Insights of Figure 5.13b motivate us to look at how the relationship between the growth of opposite signals changes over time. We have observed that there are some videos where a small increase in the number of `likes` produces a strong growth in the number of `dislikes`, featuring a video bursting out of the local bubble. In Figure 5.14a, we plot the LOESS curves of the relationship between the growth in the number of `likes` and the cumulative number of the opposite signal at different days of the video lifetime. Similarly, in Figure 5.14b we depict the growth in the number of `dislikes` and the respective cumulative number of the opposite signal at selected days since video upload. LOESS regression (Cleveland and Devlin, 1988) stands for the locally weighted scatterplot smoothing, where at each subset of the input data a local weighted polynomial regression is fit. As it is a non-parametric regression, we use LOESS to visualise the trends in the relationship between the dependent and predictor variables. Main difference between Figures 5.14a and 5.14b is that at every time step the growth in positive signal (`new likes`) is linear to negative signal – straight lines, while the growth in negative signal (`new dislikes`) is non-linear to `likes` – parabolic lines. In Figure 5.14a, we observe a soft positive trend, where the amount of `new likes` slightly increases together with the increasing amount of `dislikes`. In Figure 5.14b, however, we see that the amount of `new dislikes` is uncorrelated for smaller number of `likes` (see flat lines), but becomes highly correlated after a threshold in the number of `likes` has been reached. The threshold found by `MARS` of 5000 `likes` matches the position of the bend of the lines in Figure 5.14b.

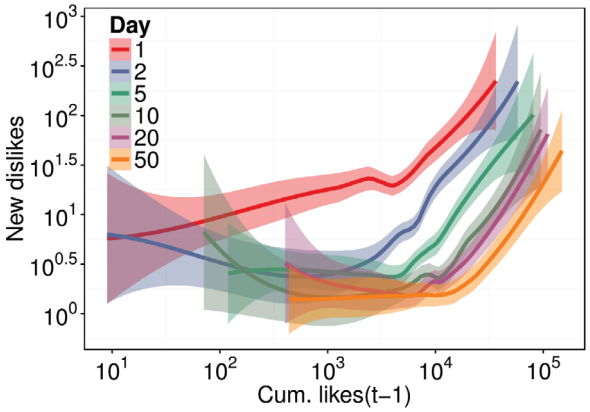
Empirical observation up to this point can serve as building blocks for the data-driven model of `new likes` and that of `new dislikes`. Namely, the growth of `likes` depend on the previous amount of `likes` *and* the previous amount of `dislikes`. Similarly, the growth of `dislikes` is dependent on the previous amounts of *both* signals. Figures 5.14a and 5.14b show plots on a log-log scale, therefore in our model we take this into account as follows:

$$\begin{aligned} \ln(\Delta L_t) &= I^L + \alpha^L \ln(L_{t-1}) + \beta^L \ln(D_{t-1}) + \varepsilon^L, \\ \ln(\Delta D_t) &= I^D + \alpha^D \ln(L_{t-1}) + \beta^D \ln(D_{t-1}) + \varepsilon^D, \end{aligned} \tag{5.29}$$

where ε are 0-mean normally-distributed error terms.



(a) LOESS curves between the new number of `likes` on day t, $\Delta L(t)$, and the cumulative number of `dislikes` on the previous day, $D(t-1)$. Each curve shows the relationship: $\Delta L(t) \sim D(t-1)$, where t $\in$ [1,2,5,10,20,50].



(b) LOESS curves between the new number of `dislikes` on day t, $\Delta D(t)$, and the cumulative number of `likes` on the previous day, $L(t-1)$. Each curve shows the relationship: $\Delta D(t) \sim L(t-1)$, where t $\in$ [1,2,5,10,20,50].

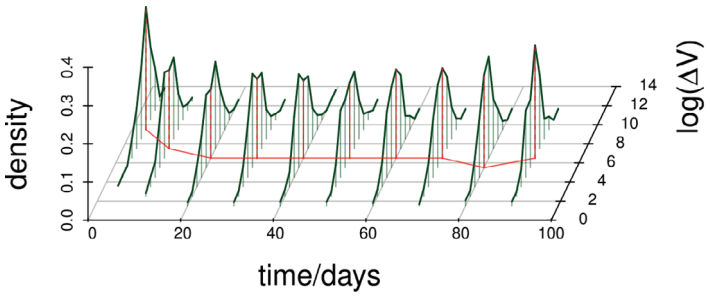
Figure 5.14: Relationship between `new likes` and previous `dislikes`, and `new dislikes` and previous `likes`.

The main drawback of this initial model (5.29), however, is that the decay in attention is not captured. Increasing `likes` and increasing `dislikes` lead to permanent growth in `likes` or `dislikes`, which is not what happens in empirical data. As we have seen in the beginning of the chapter, in Figure 5.3, both evaluations saturate as the video stays longer online. To account for this, we will introduce an additional term in our model that serves as a stopping criterion for the ever increasing number of `likes` and `dislikes`.

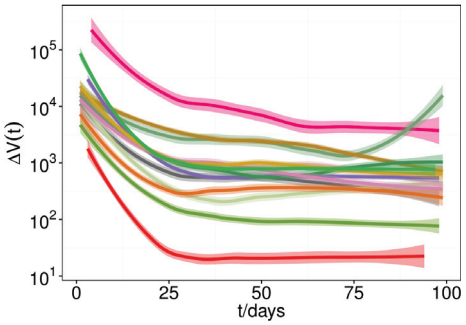
5.4.2 Growth Rate of Views

Decay in attention is easily explained with decaying popularity of a video. Up to this point, we have been overlooking this important metric in video statistics. Previous studies by Szabo and Huberman (2010) have captured the decay in video popularity by analyzing the `views` count over time, instead of the number of `likes` or `dislikes`. Our timestamped dataset of 354 videos also contains the time series of `views` of each video, which allows us to check whether the amount of `views` saturate over time. We do this by examining the distribution of `new views` – the difference between the cumulative number of `views` at two consecutive days $\Delta V_t = V_t - V_{t-1}$. In our preliminary analysis, we have observed that ΔV_t is highly skewed, therefore, in Figure 5.15a we plot the time series of the distribution of the *logarithm transformed* value, $P(\ln(\Delta V_t))$. Dashed vertical line marks the distribution mean and the solid line projects the time series of the distribution mean on a 2-dimensional scale. We can easily realize that over time the expected value of the distribution of `new views` decreases, which confirms the existence of the decay in attention to a video as the video stays longer on the platform.

We next look at the decay of `new views` over time of each video. In Figure 5.15b, we plot the LOESS curves of the time series of ΔV for a sample of 5% of the dataset. We



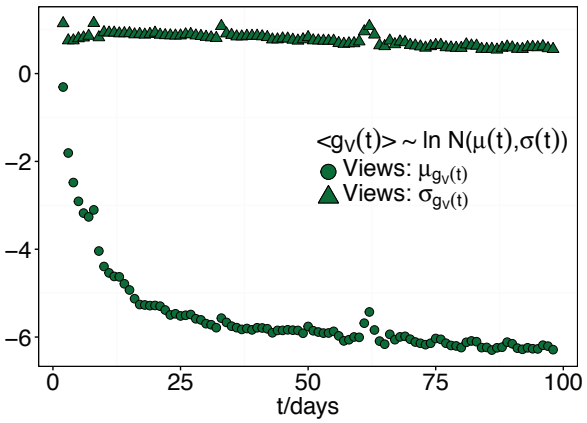
(a) Distribution of log-transformed `new views` at selected days. Dashed vertical line indicates the distribution mean, while the solid line projects the time series of the distribution mean on a 2-dimensional scale.



(b) Decay of the amount of `new views` for a sample of 5% of the videos of 354 videos dataset.

Figure 5.15: Time series of the distribution of the growth rate of `views` on `YouTube`.

observe that the amount of `new views` is negatively correlated between the consecutive timestamps, and the convergence of `new views` at later time steps is some fraction of `new views` at early days: $\Delta V(t = 100) = \alpha \cdot \Delta V(t = 2)$, where $\alpha \in [0, 1)$. We capture the negative time correlation between the `new views` by analysing the distribution of the `new views` normalized by the previous cumulative amount of `views`, namely: $g_V(t) = \frac{\Delta V}{V_{t-1}}$. The distribution of $g_V(t)$ is highly skewed at each time step (not shown here), therefore we have fitted $g_V(t)$ for each $t \in [1, 100]$ to known heavy-tailed and exponential statistical distributions: log-normal, powerlaw, stretched exponential and exponential, using the `powerlaw` python package (Alstott *et al.*, 2014). Log-normal distribution has been identified as the best fit for majority of time steps t. In Figure 5.16, we plot the time series of the parameters μ and σ of the log-normal distribution of $g_V(t)$. We observe that μ decays with time, while parameter σ is constant and fluctuates around 0.76.



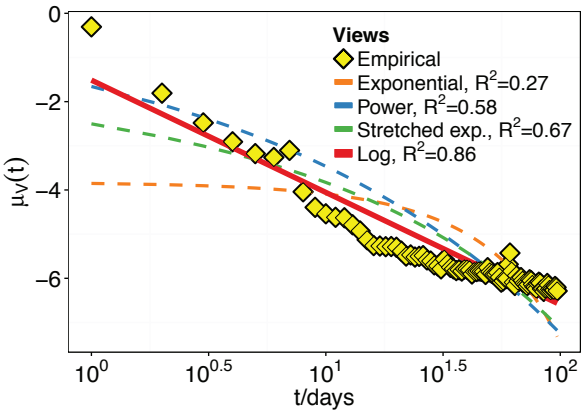


Figure 5.16: Left: The time series of the parameters of the log-normal distribution of the growth rate of `views`. The parameter μ_V decays, while the parameters σ_V is constant, $\sigma_V = 0.76$. Right: The fitting of the parameter μ_V to various functions. The figures a semi-log plot, where x-axis is plotted on a linear scale and y-axis is on a logarithmic scale; the lin-lin plot is shown in (a). The maximum R^2 next to the fitted function indicates that the best fitting function is the logarithmic function.

Function type		Fit of $\mu_V(t)$ to function	R^2
Exponential	$\mu(t) \sim e^{-\alpha \cdot t}$	$e^{1.340.01 \cdot t}$	0.27
Power	$\mu(t) \sim \beta \cdot t^{-\alpha}$	$e^{0.5} \cdot t^{0.32}$	0.58
Stretched exp.	$\mu(t) \sim e^{-\alpha \cdot t^{\beta}}$	$e^{-e^{-0.09} \cdot t^{0.17}}$	0.67
Logarithmic	$\mu(t) \sim \alpha - \beta \cdot \ln(t)$	$\mathbf{-1.51 - 1.1 \cdot \ln(t)}$	**0.86**

Table 5.8: The coefficients of the fitted functions of the μ parameter of log-normal growth rate of `views`.

We determine the functional form of μ with time as an independent variable, see Table 5.8 and Figure 5.16, and identify that the logarithmic function, $\mu(t) \sim \ln(t)$, is the best fit, yielding the maximum adjusted coefficient determination R^2. Finally, by knowing the distribution of $g_V(t)$ we are able to find the number of `new views` at time step t correlated with the number of `views` at the previous time step $t-1$, namely: $\Delta V_t = g_V(t) \cdot V_{t-1}$.

5.4.3 Statistical Model of Coupled Growth

Having introduced the ΔV_t term which ensures that `likes` and `dislikes` do not grow infinitely, we update the initial model (5.29). The new model takes into account the interaction between the opposite signals and adds the decay in attention to a video, which is captured by the decay in the amount of `new views`. Additionally, we introduce a dummy variable for each video that measures the *individuality* or an *intrinsic quality* of a video. The updated model of `new likes` and that of `new dislikes` goes as follows:

$$\begin{aligned}\ln(\Delta L_t) =& I_1^L + I_2^L + \cdots + I_{N-1}^L + I_N^L + \alpha^L \ln(L_{t-1}) + \beta^L \ln(D_{t-1}) + \theta^L \ln(\Delta V_t) + \varepsilon^L,\\ \ln(\Delta D_t) =& I_1^D + I_2^D + \cdots + I_{N-1}^D + I_N^D + \alpha^D \ln(L_{t-1}) + \beta^D \ln(D_{t-1}) + \theta^D \ln(\Delta V_t) + \varepsilon^D,\end{aligned} \tag{5.30}$$

where α is a weight of a positive signal, β is a weight of a negative signal, θ is a coefficient of the attention decay, ε is a zero mean normally distributed error term and $I_{i\in[1,N]}$ is a dummy variable that can be either 0 or 1 and encompasses an intrinsic video quality of each of the $N = 354$ videos.

We apply the proposed statistical model on the timestamped dataset of 354 videos. The coefficient of determination of the `new likes` model is 0.986 and that of the `new dislikes` is 0.882. Additionally, in Table 5.9 we compare the proposed model and its variations by excluding one of the predictor terms. The proposed model still yields the highest R^2 and the lowest BIC – the goodness of fit metric which penalizes for the number of parameters in the model.

`New likes` model	R^2	BIC
Complete (5.30)	0.986	27063
Exclude Intercepts ($-I$)	0.943	54969
Exclude Views ($-\Delta V_t$)	0.952	55037

`New dislikes` model	R^2	BIC
Complete (5.30)	0.882	21895
Exclude Intercepts ($-I$)	0.755	26939
Exclude Views ($-\Delta V_t$)	0.776	29186

Table 5.9: Comparison of the goodness of the fit of the proposed model and models excluding one of the predictor terms.

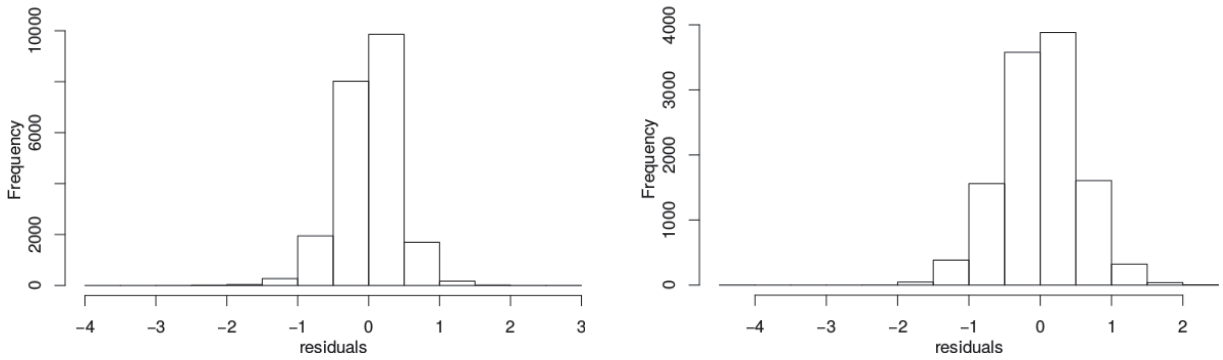


Figure 5.17: Distribution of the residuals of `new likes` and `new dislikes` models.

Finally, we validate the model by examining the diagnostic plot of the model residuals. In Figure 5.17, we confirm that the distribution of the residuals of both models is normal and is centered around zero.

5.4.4 Simulation

Simulation model We turn the proposed statistical model (5.30) into simulation model. First, we obtain the update rule for the growth of `likes` and that of `dislikes` by exponentiating Equation 5.30:

$$\begin{aligned} \Delta L_t &= e^{\alpha^L} L_{t-1} \cdot e^{\beta^L} D_{t-1} \cdot e^{\theta^L} \Delta V_t, \\ \Delta D_t &= e^{\alpha^D} L_{t-1} \cdot e^{\beta^D} D_{t-1} \cdot e^{\theta^D} \Delta V_t. \end{aligned} \tag{5.31}$$

And, therefore, the number of `likes` and that of `dislikes` at time step t is:

$$\begin{aligned} L_t &= L_{t-1} + \Delta L_t, \\ D_t &= D_{t-1} + \Delta D_t. \end{aligned} \tag{5.32}$$

We sample the initial values of `likes`, `dislikes` and `views`, and a corresponding video quality through their joint distribution: $P(L_{t=1}, D_{t=1}, V_{t=1}, I)$. Finally, the update rule for the amount of `new views` has been obtained empirically in Section 5.4.2 and goes as follows:

$$\begin{aligned} &\Delta V_t = g_V(t) \cdot V_{t-1}, \text{ where} \\ &g_V(t) \sim \ln \mathcal{N}(\mu_V(t), \langle \sigma_V \rangle), \text{ such that} \quad \mu_V(t) = -1.51 - 1.1 \cdot \ln t, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \langle \sigma_V \rangle = 0.76. \end{aligned}$$

Simulation results We simulate the above models 100 times corresponding to 100 day length of the empirical time series and with 354 items coresponding to 354 crawled videos.

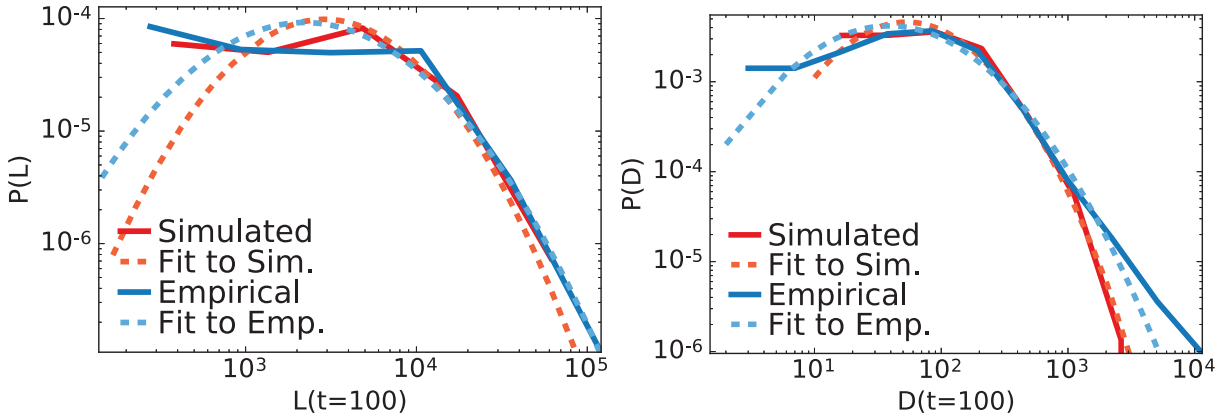


Figure 5.18: Distribution of the number of `likes` and that of `dislikes`. Empirical observations are shown in blue, the simulated values are plotted in red, and the fits to the log-normal distribution are displayed with a dashed line.

Type of observation		$\ln\mathcal{N}(\mu,\sigma)$	KS $(D, p\text{-value})$	
`Likes`	Simulated	$\ln\mathcal{N}(8.814, 0.918)$	$D = 0.121$	$p = 5.4 \cdot 10^{-5}$
	Empirical	$\ln\mathcal{N}(8.879, 1.085)$	$D = 0.125$	$p = 2.6 \cdot 10^{-5}$
`Dislikes`	Simulated	$\ln\mathcal{N}(4.965, 0.994)$	$D = 0.057$	$p = 0.189$
	Empirical	$\ln\mathcal{N}(5.095, 1.199)$	$D = 0.075$	$p = 0.035$

Table 5.10: Fit of simulated and empirical observations to the log-normal distribution.

Figure 5.18 depicts the distribution of the number of evaluations obtained via the model simulations in red colour and that of the empirical values in blue. Dashed lines denote the fit to the log-normal distribution. We observe that simulated results reproduce the range of empirical values and the log-normal shape.

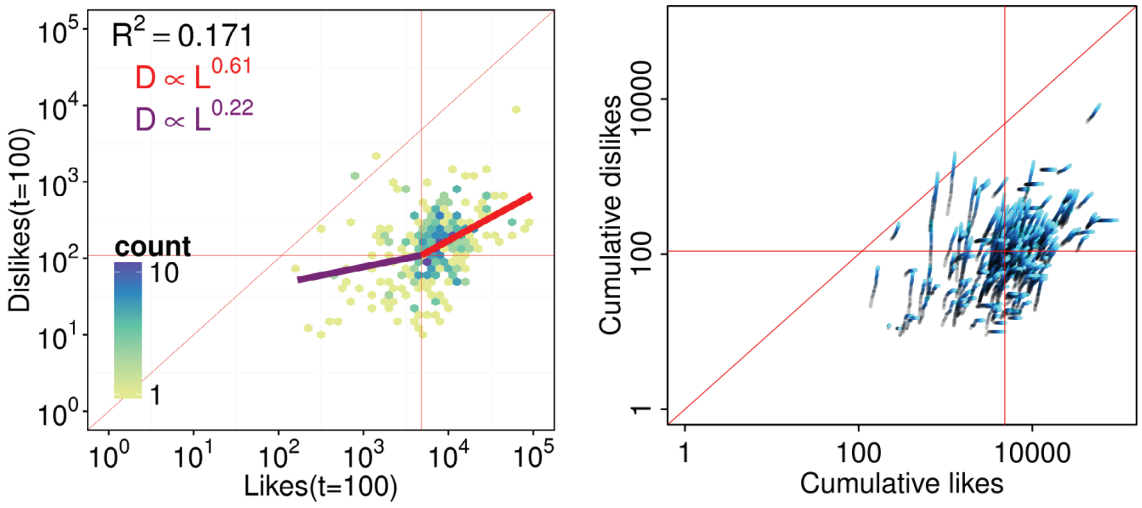
First, we statistically assess whether the model produces the log-normal distribution of the number of `likes` and that of `dislikes`. We perform the Kolmogorov-Smirnov test of the empirical and simulated values to theoretical log-normal distribution, see Table 5.10. We observe that for `likes` the D-statistic is approximately 0.1, while for `dislikes` it is even lower – 0.05. These robust K-S test results allow us to confirm that the simulated distributions are log-normal. Indeed, the proposed coupled growth model returns the log-normally distributed values and importantly recovers the parameters of the empirical distributions up to first digit after the decimal point: 8.814 vs. 8.879 for the number of `likes` and 4.965 vs. 5.095 for the number of `dislikes`.

Next, we assess the statistical similarity between the simulated and the empirical values by performing the Kolmogorov-Smirnov test, see Table 5.11. The D-statistic of the test of `likes` and that of `dislikes` is strongly < 0.1 yielding also high p-value> 0.05. Such sturdy K-S test results indicate that we cannot reject the hypothesis that the simulated number of `likes` come from the same distribution as the empirical number of `likes`. Same statement applies to the simulated and empirical number of `dislikes`.

	KS $(D, p\text{-value})$	
`Likes`	$D = 0.093$	$p = 0.087$
`Dislikes`	$D = 0.073$	$p = 0.285$

Table 5.11: K-S test between the simulated results and the empirical values of `likes` and that of `dislikes`.

Finally, we evaluate the relationship between the simulated `likes` and `dislikes`. In Figure 5.19a we show the scatterplot of evaluations and piecewise functions obtained through `MARS` model. We recover the cutoff in the number of `likes` in the simulated results, which matches the cutoff obtained in the empirical values. The coefficient of the relationship between `likes` and `dislikes` in the lower quadrant is smaller than in the upper quadrant $0.22 < 0.61$. Therefore, the obtained cutoff distinguishes between the two regimes – one where `dislikes` grow slower and another where `dislikes` catch up with `likes`. Our simulated results, however, do not fully recover two empirical observations: a) the coefficient in the upper quadrant 0.61 vs 1.14 in the empirical values, and b) the stronger correlation between evaluations 0.171 vs 0.491 in the empiricial values. Figure 5.19b depicts the dynamics of evaluations of each of the simulated video. It is clear that close to the cutoff border we are able to reproduce the videos that simulate the burst out of the positivity bubble, see vertical segments. However, we cannot fully recover the trajectory of the videos in the upper quadrant. Figure 5.19b shows evidently the limitations of the proposed model, namely the interaction after the bubble burst between the evaluations is not fully captured, which also explains why we obtain lower coefficient in the upper quadrant.



(a) `MARS` model applied to the number of `likes` versus `dislikes`. Vertical and horizontal red lines are situated at the cutoff in the number of `likes` and that of `dislikes`.

(b) Evaluations dynamics of each of the 354 simulated video. The colour shade of a segment marks the number of evaluations at a specific time.

Figure 5.19: Dual regime in the relationship between the simulated number of `likes` and that of `dislikes`.

5.5 Discussion

We have build a statistical multiplicative growth model of coupled evaluations and are able to reproduce statistical parameters of empirical log-normal distribution and certain coefficients of the relationship between the two evaluations. In the proposed model, the amount of positive/negative evaluations at each time step depends on the amount of positive signal in the previous time step, the amount of negative signal in the previous time step and the aging factor of the video captured by the decaying number of new views over time. Coupling in this model means that evaluations are dependent on both `likes` and `dislikes`, and multiplicative growth ensures that evaluations grow according to the Law of Proportionate Effect, (5.15) and (5.16).

However, clearly from the simulated results we observe that the resulting relationship does not exactly reproduce the trajectories of the dynamics of empirical videos and the coefficient of the relationship after the burst of the filter bubble. This mismatch between the empirical and simulated values poses as a limitation of the proposed statistical model, namely sole statistical model is not able capture the microdynamics of each video. Our proposed attempt in this direction is to deduce the so-called video quality or novelty of each video incorporated in the dummy variables of each video in the regression model.

Instead of having a video-centric approach, another solution is to incorporate the emotional expression of video titles and comments as a trigger for positive and negative evaluations. We have presented empirical results in Chapter 5, where arousal and negative emotions expressed in `YouTube` video titles and in other platforms have shown to be a significant predictor for an online item to reach polarized or global regime. Along this line, we can further implement a user-centric approach, where a user's emotional state expressed through valence and arousal modifies his opinion, *e.g.* `like` or `dislike`, on the video. For instance, a user that experiences a high arousal, such as $E_a > 4$ from Chapter 3, might respond to a video by expressing his opinion. This agent-based model will combine emotion and opinion dynamics, defining the relationship between opinions and emotions at the individual level, and will remove the limitations currently present in the statistical model.

Chapter 6

Conclusions

This thesis attempted to study in details different instances of opinion polarization: from political to social network domain, from elite to mass polarization. We closed the gap in the political science research by developing a social interactions based approach of studying polarization instead of polarization solely based on ideological stances. By leveraging on the high precision data from Swiss political social network platform, we learn that the measures of polarization applied in the classic two-party system and resulting conclusions cannot be adopted and reproduced in multi-party system. Through our statistical data analysis we revealed the universal properties of collective opinion polarization in online communities such as the log-normal distribution of collective evaluations.

Apart from the contribution to computational social science community, we made an important step towards testing one of the fundamental theories in psychology of emotions. Thanks to the state-of-the-art sentiment analysis techniques, we validated the Cacioppo's theory of asymmetric emotional responses on massive online datasets. This finding revealed how users react to emotional online content, namely they are likely to respond to posts that elicit mixed or negative emotions or emotions of high arousal. Furthermore, the study on the burst of the filter bubble presented an empirical evidence that negativity and arousal expressed in texts are factors to negative backlash of online community which eventually leads to extreme polarized responses. Finally, we proposed a statistical model of coupled evaluations to reproduce the digital traces of the filter bubble burst.

6.1 Contributions in a Nutshell

In the next sections, we list in brief details the outcomes of each of the open research questions that were outlined in the Introduction. The data used in this thesis might be available upon request.

6.1.1 Political Polarization in Online Communities

We explored the behavioral aspects of political polarization by measuring network polarization of three different types of online social interactions by politicians on `politnetz.ch`. We first constructed a multiplex network with politicians as nodes and three layers of directed links: `supports`, `comments` and `likes` (Figure 2.2). We measured network polarization in each layer through the party based Q-modularity score, which captures the level of intra-party cohesion with respect to inter-party cohesion. For each layer, we obtained the following results of the network polarization and its evolution over time:

- The layers of `supports` and `likes` showed high modularity, while `comments` layer resulted in low modularity score (Table 2.3). This suggests that the layers of `supports` and `likes` exhibit significant patterns of polarization with respect to party alignment compared to the `comments` layer.
- At the `supports` layer, the partition of politicians into parties is very similar to the maximal partition found by community detection algorithms. This suggests that support link for politicians is an act of partisanship.
- At the `likes` layer, the value of modularity is high but yet it is lower than that of `supports`. This suggests that liking a post is also a signal of an adherence to a party, however cross-party `like` links are more frequent in comparison to cross-party `support` links.
- At the `comments` layer, the community detection algorithms reveal that a partitioning of politicians conveys higher modularity than party alignment, suggesting that groups in `comments` are not party-driven.
- Demodularity and topical analysis revealed that `comments` layer is organized around topics and cross party online debates especially between the opposing parties are very frequent (Table 2.5).
- Time series of network polarization of each layer reveals that polarization in `likes` based on party alignment increases significantly around the federal elections of 2011, compared to moments without electoral campaigns (Figure 2.4). This suggests an existing relation between online activity and political events, in which social interaction becomes more influenced by party membership when elections are close.

We analyzed the social networks of politicians in `politnetz.ch` (Figure 2.5) to explore the relation between ideology and social structures in online interaction and obtained the following patterns:

- Green parties (GLP and Grüne) have a higher in-degree centralization than the rest, indicating that their internal structure is more unequal with respect to popularity.
- Left-aligned parties have lower average path length and higher maximum k-core numbers than right-aligned ones, showing that left parties create social networks

with higher information spreading capabilities than right parties. This result is in contrast with previous findings by Conover *et al.* (2011a) which provide the opposite pattern for the networks of US `Twitter` users.

6.1.2 Emotional Reactions in Online Communities

We test the predictions of the theory of asymmetry of emotional reactions in a real-life social setting, by analyzing 65 millions pairs of messages and their replies from three popular online communities. The following is the complete picture of the emotional response patterns in online communities:

- Kolmogorov-Smirnov test of the distribution of waiting time, the time between receiving and replying to a message, conditional to expressed valence in the stimuli or reply message, outputs the results based on which we cannot reject the hypothesis that the distributions of the waiting time of different emotions are identical (Figure 3.2 and Tables 3.2– 3.7).
- Contrary to our expectation, the temporal aspect of the negativity bias theory in online settings is not validated, namely that on the Internet users do not assign higher priority to the response to negative stimuli, when the priority to reply is measured as the time to reply to a message. A possible explanation is that the Internet medium changes the time scale of emotional interaction such that rapid emotional responses are not observable.
- When the priority to reply is measured through the likelihood of a message to receiving a reply, then the negativity bias theory is confirmed. On `4chan` and `Reddit` an original message expressing extreme negative emotions is 7.3 times more likely to receive a reply (Figure 3.6). However, an expression of positive emotions either does not have an effect on receiving a reply or even reduces by halve the chance of a message to be responded to.
- A threshold of emotional arousal have been found. Messages that were replied at least once show higher likelihood of expressing high arousal values than the messages that were ignored. This result sheds light on the role of emotional arousal in emotional reactions, *i.e.* arousal triggers actions. The detected threshold of arousal is in line with previous findings on `Twitter` retweets by Pfitzner *et al.* (2012).
- The findings suggest that what makes us react online is more salient than when we react, which highlights the difference between emotional interaction in offline face-to-face interaction and in online computer mediated communication.
- Emotions of replies depend on the emotions in the stimuli messages (Figure 3.7). Users not only react to emotional messages but they react differently. On `YouTube`, negatitivity provokes more negativity.

- Our findings revealed a slight positivity offset only among `YouTube` postings (Figure 3.5A), but not on `4chan` or `Reddit`, which potentially can be explained by the difference in the functional usage of these online platforms.

6.1.3 Emotional Polarization in Online Communities

We explore the digital traces of the burst of the "filter bubble", which in our research is instantiated in the dual regime of collective evaluations. Namely, an uploaded online item initially obtains mainly positive evaluations since it circulates within the local network of early adopters, *i.e.* people who discovered the item through recommender system or some social filtering. However, once the item goes beyond the local attention and reaches the "global" audience or different users on the Internet, the collective negativity towards the item backlashes. Our findings on the evidence of the burst of the filter bubble show that collective evaluations across various online media have statistical regularities in the distributions of evaluations and their relationships. We briefly describe three main observations:

- The distributions of the amounts of likes and dislikes per item are well fitted by log-normal distributions, a result that gives insights into the properties of the process that creates evaluations (Figure 4.2).
- By applying multivariate regression splines model, we find robust evidence of the existence of a local and a global regime that is consistent with our hypotheses about the burst of filter bubbles (Figure 4.3).
- Results of linear and logistic regression models demonstrated that arousal, an emotional dimension of activation, is a factor in creation of polarization and in the access to the global regime (Table 4.5), lending support for psychology theories about the role of affect in the polarization of opinions.

6.1.4 Multiplicative Growth Model of Collective Evaluations

We propose the coupled model of human appraisal. Motivated by the results on the universalities in distribution properties of collective evaluations, we started by applying the classic multiplicative growth model with identically distributed growth rate. By following the data-driven approach, we quickly realized that the assumptions necessary for the existing model do not hold against the empirical observations on a time-stamped data of evaluations on `YouTube`, and therefore the model would not reproduce the parameters of the empirical distributions of `likes` and that of `dislikes` of videos. We proposed several modifications and provided an analytical solution for the model validity, which resulted in the following outcomes:

Growth model without interaction of evaluations on `YouTube`

- Relative growth rate of evaluations shows a rapid decay (Figure 5.4) which is a signature of the collective attention decay, *i.e.* as the video stays longer on `YouTube` the amount of new evaluations becomes smaller, video is not novel anymore. Power function is the best fitting function to decaying growth rate (Table 5.3) which is in line with previous research on `Twitter` (Asur *et al.*, 2011).
- At each time step relative growth rate of evaluations follows the log-normal distribution, contrary to previous works by Adamic (2001) and Wilkinson and Huberman (2007), where growth rate was normally distributed. Furthermore, the μ parameter of the distribution is decaying and σ is constant over time (Figure 5.5).
- Based on Lyapunov Central Limit Theorem, we provided an analytical solution that the growth model based on the law of proportionate effect and log-normally distributed growth rate produces values that are log-normally distributed (Section 5.3.2).
- The simulation model based on the growth model with log-normally distributed growth rate with parameters from empirical data resulted in the log-normally distributed `likes` and `dislikes`. However, the dual regimes in the relationship between the evaluations cannot be reproduced (Figure 5.12), and empirical parameters of `dislikes` distribution are also not recovered.

Motivated by the results on the traces of the burst of the filter bubble, namely the non-linearity in the relationship between `likes` and `dislikes`, we proposed to include a coupling of both signals in the existing model to account for the interaction between evaluations. This development resulted in the following:

Coupled growth model of evaluations on `YouTube`

- LOESS regression between the new `likes` and previous `dislikes` showed linear relationship (Figure 5.14a), while LOESS between the new `dislikes` and previous `likes` reveals non-linear, parabolic trend (Figure 5.14b).
- To account for the decay in attention factor, we explored the saturating number of `views` of a video over time. Decay in attention is easily explained with decaying popularity of a video which was in previous research was measured through the views count.
- Empirical analysis revealed that growth rate of `views` is indeed decaying over time (Figure 5.15a) and also follows the log-normal distribution. Given that `views` appear before `likes` or `dislikes`, this result suggests that mechanisms behind the growth of `views` are driving mechanisms of evaluations.

- We propose a statistical model of new `likes` and that of new `dislikes` as a function of positive and negative signals, and that of video popularity.
- The model recovers well the parameters of the empirical distributions of `likes` and that of `dislikes` on `YouTube`. Importantly, the dual regime in collective evaluations together with the empirical threshold of collective attention is reproduced (Figure 5.19a).
- The proposed coupled model has several limitations, *e.g.* the dynamics of each video is not fully reproduced (Figure 5.19b), and the relationship between evaluations in the polarized regime is not as strong as in the empirical data.

6.2 Future Work

In our thesis we have presented aspects in formation of collective opinion polarization, like emotions or social interactions, that we consider novel in the study of opinion diversity on the Internet. However, there are aspects of polarization that have already been considered and yet remain active research areas. In the next sections, we provide results of our studies in this direction. Furthermore, we supplement the results shown in previous chapters, *e.g.* in `politnetz.ch`, with results on other datasets, like `Twitter` and `Facebook`.

6.2.1 Demographical Characteristics of Polarization

Our main focus in polarization study was social interactions aspect of polarization. In one of our preliminary studies, we have also explored the area of ideological polarization. The question of whether users that sort themselves into similar political communities, also exhibit similar demographical traits is a widely asked question in political science, given the rising amount of evidence that voting behaviour is influenced by such factors, like race, age, sex, geographical location and educational level (Cornell University, 2012). Following the same reasoning, we asked a more general question of whether polarized content online, *e.g.* a video with high amount of likes and dislikes, is shared and popularized by users of certain demographical traits.

To answer this question, we combined user-centric data from `Twitter` with video-centric data from YouTube to build a rich picture of who watches and shares what on `YouTube`. We studied 87,000 `Twitter` users, 5.6 millions `YouTube` videos and 15 millions video sharing events from user-, video- and sharing-event-centric perspectives. From `Twitter` data we obtained demographics characteristics of users that share videos, and from `YouTube` data we extracted metrics of video popularity, video categories and topics.

	male	female	urban	rural	student	mother	father	US
`views`	0	$+^{**}$	$-^{*}$	0	0	$-^{*}$	$-^{*}$	$-^{*}$
polarization	0	$-^{*}$	0	$-^{*}$	$-^{*}$	$-^{*}$	$-^{**}$	0
lag, Δt	$-^{**}$	$+^{*}$	$-^{*}$	$-^{*}$	0	$+^{**}$	$-^{*}$	$-^{*}$

Table 6.1: Demographics. A + indicates a positive deviation from the general population, - negative and 0 not statistically significant. ** indicates that the significance was based on δ being in the bottom/top 0.5%, * for the bottom/top 2.5%.

To understand the significance of the influence of variables such as gender or occupation on (i) the number of views, (ii) the polarization or controversiality,[1] and (iii) the lag we applied a so-called "permutation test" (Good, 2005), which unlike other tests does not make assumptions on the distribution type of the observed variables. To test, say, the impact of stating "student" in the Twitter bio on the number of views we first computed the average view count for all views by the "student" group and compared this with the average for the complement "non-student" group. Omitting the details, permutation test then allows to mark the significane of δ, which is the observed empirical difference between the groups. In Table 6.1, a ** indicates a high statitical significance, * – lesser extent of significance and 0 no statistical significance. Table 6.1 shows correlations with respect to the per-user median (i) number of views of shared videos, (ii) polarization/controversiality of shared videos and (iii) of inter-event times or how long it takes for a `Twitter` user to share a `YouTube` video. One of the demographic differences that can be spotted is that men compared to women share less popular (fewer views) videos earlier (smaller lag). Interesting gender and geographical differences can be seen in sharing polarized videos – women and users from rural locations tend to not share controversial videos. It is also interesting how sharing of controversial video becomes strongly negatively correlated when family status of men changes (father group). All the correlation analysis findings however can be left for further interpetation.

Additionally, we analyze the relationship between the same `YouTube` video features and categories of videos (Music, Sports and Politics) and three dimensions of online user behaviour, *e.g.* social, influential or sharing. As a simple analysis tool we computed Spearman's rank correlation coefficient for each pair of features, and assess its statistical significance using the method described earlier. To simplify the presentation, we group the Twitter features into four classes. First, to see how "social" a user is we look at (i) the number of friends, and (ii) the number of distinct users mentioned. Second, to see how common "sharing" is for a user we included the fraction of tweets that (i) are retweets, (ii) contain a hashtag, (iii) contain a YouTube URL, and (iv) contain a non-YouTube URL.

[1] We calculate the polarization that a YouTube video creates on its viewers through its amounts of likes L_v, dislikes D_v, and total views V_v, through the equation $Pol_v = \frac{L_v}{V_v^{0.849}} \cdot \frac{D_v}{V_v^{0.884}}$. The rationale behind this calculation is the rescaling of the likes and dislikes ratio based on the fact that they do not grow linearly with each other. The exponents correspond to the base rates of the logarithmically transformed amounts of views, likes and dislikes. This way we standardize the ratio over their nonlinear relation.

	views	polarization	lag, Δt	Music	Sports	News&Politics
Social	- -	- -	- -	+ 0	0 +	+ +
Sharing	0 - **- -**	0 **-** 0 **-**	**- - - -**	0 - 0 **-**	0 + + 0	+ + + +
Influence	**- -** + 0	**- -** + 0	0 - - +	+ 0 + 0	0 0 + 0	+ + - +

Table 6.2: Each cell in the table links a Twitter user feature group (row) with a particular YouTube video feature (columns). The three symbols in the cell indicate "+" = significant (at 1% using a permutation test as previously described) and positive, "-" = significant and negative, and "0" = not significant or below 0.05. Bold symbols indicate an absolute value of Spearman's Rank correlation coefficient > 0.1 The symbols are in the order of the features listed above in the text.

Finally, we look at notions of "influence" that includes (i) the number of Twitter followers, (ii) the fraction of a user's tweets that are retweeted, (iii) the average retweet count of tweets that obtained at least one retweet, and (iv) the average number of followers of a user's followers.

Table 6.2 links a Twitter user feature group (row) with a particular YouTube video feature (columns). Certain general observations can be made. For example, all of our notions of "social" correlate with a drop in lag time, and out of the topics considered, News & Politics is the one that is most consistently linked with users who actively share. But other observations are more complex and, for example, only some but not other notions of influence correlate positively with a large number of views. This descriptive results can potentially be used for formulating further research questions.

6.2.2 Mass Polarization on Twitter and YouTube

In Chapter 2, we have thoroughly investigated the concept of *elite polarization* on the example of the multi-party network of Swiss politicians. One of our outcomes is the differences in the topological structure of left-, and right- aligned parties. In line with this research, one of our preliminary studies on *mass polarization* in the US has also demonstrated the differences in the sharing activity and the contents of what is shared and viewed between differently aligned users on `Twitter` and `YouTube`.

Our motivation of studying preferences of politically aligned users and differences in their online behaviour was generated by the study by Conover *et al.* (2011a). To see the differences between the sets of politically aligned users on both Twitter and YouTube, we had the following questions in mind: a) which political user groups share more politically charged content, b) what is the most frequent content of each political user group.

To separate users into political groups we followed a US bipartite system with audience divided into left (L) and right (R) users (Conover *et al.*, 2011a). Twitter users that followed more of the 13 left seed users were marked as left-leaning, users that followed more of the 19 right seed users were marked as right-leaning and users with a split preference or not

following any seed user were marked as apolitical. Our approach resulted in three disjoint sets of left users U_L ($|U_L| = 11,217$), right U_R ($|U_R| = 1,046$) and apolitical users U_A ($|U_A| = 57,672$).

We addressed question a) by looking at how much L, R, A users share YouTube videos in the category *News & Politics*. If left-leaning user u_L shared set of videos V_L with a subset of videos in the category *News & Politics*, $V_L^{News\&Politics} \in V_L$; then we looked at the distribution of ratio of number of political video shares to total amount of shares per each u_L, u_R and u_A: $r_{\{u_L,u_R,u_A\}} = \frac{|V_{\{L,R,A\}}^{News\&Politics}|}{|V_{\{L,R,A\}}|}$. On average mean ratio of videos with political content for each user population is: $\mu_L = 0.06$, $\mu_R = 0.29$, $\mu_A = 0.05$, which confirms right users share more news and politics related videos compared to left users and apolitical users. Figure 6.1 shows density plot of $r_{\{u_L,u_R,u_A\}}$ on lin-lin scale with linear binning. Right users share more news and politics related videos with distribution of ratio of politics video staying uniform over various values r_{u_R}, on average 30 out 100 shared videos are in politics category. Left users show similar behaviour as apolitical users with distribution of ratio of politics video decreasing when $r_{\{u_L,u_A\}}$ increases; on average 5-6 out of 100 shared videos are in politics category.

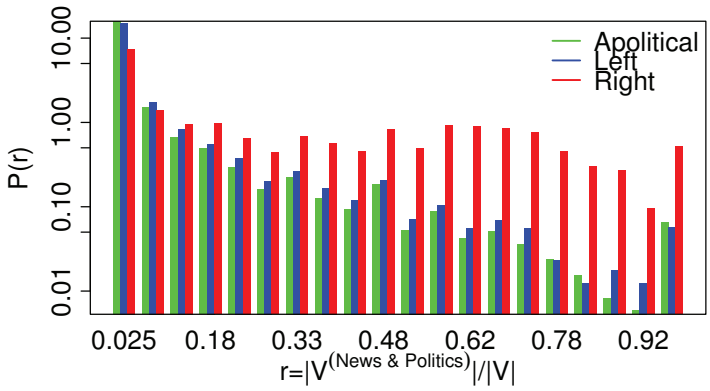


Figure 6.1: Density of ratio of video shares amount in category *News & Politics* to total number of video shares for each politically leaned user: left u_L (blue), right u_R (red) and apolitical u_A (green). Average for each user population: $\mu_L = 0.06$, $\mu_R = 0.29$, $\mu_A = 0.05$.

To answer question b) we calculated topic distributions of videos per each political user category and rank topics in each user group according to their frequency. In order to statistically compare the ranking of topics across groups, we applied the distance between ranks of topics method by Havlin (1995). If $R_1(\lambda)$ is the rank of topic λ in user group 1 and $R_2(\lambda)$ is the rank of the same topic λ in user group 2, distance $r_{12}(\lambda)$ *between the ranks of topic* λ *in two user groups* is $r_{12}(\lambda) = |R_1(\lambda) - R_2(\lambda)|$. Thus, the distance *between two user groups* is defined as the mean square root distance between the ranks of all common topics: $r_{12} = (\frac{1}{N}\sum_\lambda r_{12}^2(\lambda))^{\frac{1}{2}}$, where N is the number of common topics across user groups. We summarize the distance metric across four user groups: Left, Right, Apolitical and all population (Left, Right and Apolitical) with $N = 23,844$ and $R_{max} = 281,265$ in Table 6.3.

	Left	Right	Apolitical	All
Left	-	35733.87	33807.2	25722.16
Right	-	-	49314.69	37913.44
Apolitical	-	-	-	23879.92

Table 6.3: Topical distance across political, apolitical and all `Twitter` user groups.

We find that the distances from right users is maximum to left, apolitical and all, and left and apolitical are close to each other in terms of distance. This suggests that right users have their own hierarchy of topics distinguished from left and apolitical users, while latter groups have more similar topics. To support our findings in distance between topic ranks, we look at the most 20 frequent Freebase topics for each user group, see Table 6.4. Right users share more politically charged content including politicians (Barack Obama, Alex Jones, Ron Paul), news channels (Russia Today, The Young Turks), military-related keywords (Gun, Police) and concepts (USA). Conversely, left-leaning users have similar interests as apolitical, giving priority to entertainment videos. For example, "Barack Obama" topic (Freebase ID */m/02mjmr*) is placed 30th popular among left users and 1st among right population.

Rank	Left	Right	Apolitical
1	Minecraft	Barack Obama	Minecraft
2	Call of Duty II	Alex Jones	Call of Duty II
...	...	...	...
6	Film	Ron Paul	Hip hop music
...	...	...	...
13	Album	Police	Call of Duty
14	Call	Mitt Romney	Video blog
15	Song	RT	Episode
...	...	...	...
25	Heavy metal	Boston	NBA
...	...	...	...
27	Episode	US NSA	Super Junior
...	...	...	...
29	Justin Bieber	Bomb	Pokemon
30	Barack Obama	Train	Music

Table 6.4: Topics of political user groups on `YouTube` and `Twitter`. Right-aligned users share more politically-charged content with mean ratio of videos with political content: $\mu_{\text{LEFT}} = 0.06$, $\mu_{\text{RIGHT}} = 0.29$, $\mu_{\text{APOL}} = 0.05$. Ranking of `Freebase` topics across user groups is stat.compared via distance between topics ranks. We find distance from right users is maximum to left and apolitical.

Results of a) and b) support each other and give the following picture on political engagement of L/R/A user groups. For left users, a) says they act as apolitical users and on average do not share much political videos, with b) confirming that among top 20 video topics of left users none relate to politics. And for right users, a) states that they share more political content which is supported by b) where 9 out of top 20 topics have government, news, politics related concepts.

6.2.3 Elite Polarization on Twitter and Facebook

In Chapter 2, we focused on elite polarization in Swiss polticians social network site – `politnetz.ch`. In this brief descriptive study, we looked at the online activity of Swiss politicians on three worldwide known online platforms, namely `Twitter`, `Facebook` and `YouTube`. We have obtained data on politicians activity on `Twitter`, such as their tweets, followers, retweets, favorites and replies, of a total of 689 politicians. Similarly, we collected the `Facebook` activity of 827 politicians, including their public profiles, posts, comments and likes. On `YouTube`, we focused on per party aggregation, namely collecting video statistics and comments of party channels.

On `YouTube`, we have observed that one of the parties (BDP) with the smallest representation on online media (by the amount of politicians, 6%) has the largest traffic of posted online content on `YouTube` (40%) among all other parties. This suggests that parties with few members might employ actively this media channel compared to other parties, as a way to be more visible online and as a way to engage with their supporters and spread their political stances.

On `Twitter`, we have observed a highly modular network structure of party followers. Namely, Swiss politicians mostly follow politicians from their own party. Furthermore, the most "unsocial" party by looking at the share of the number of followers across different parties is the far-right and also the largest party of Switzerland, namely SVP (grey colour). SVP follows other parties the least, see that the area of grey triangle is the smallest across different pie charts.

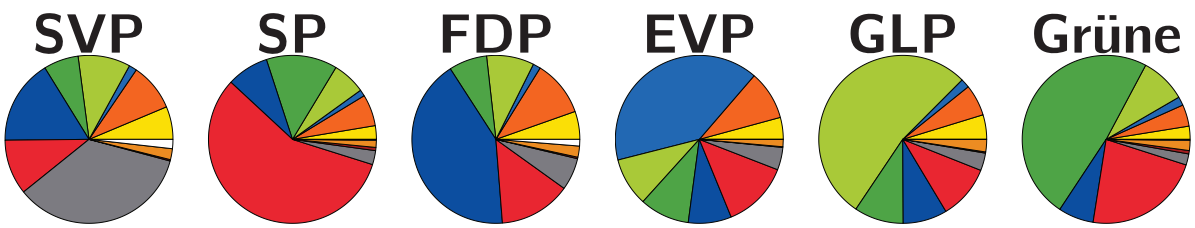


Figure 6.2: Pie charts of the followers of the party on `Twitter`. Color scheme of parties is shown in Table 2.1, except that Christian parties are split into CVP (light blue) and EVP (orange).

On Facebook we construct an aggegated network of politicians' likes to posts, see Figure 6.3. We clearly observe a highly modular structure of Facebook likes to posts, which confirms our findings on politnetz.ch data. Additionally, the position and relation of parties to each other is reinforced by Fruchterman-Reingold graph visualisation layout (Fruchterman and Reingold, 1991), which places further apart nodes that have less number of interactions or links between them. This way we observe a greater social interaction distance between the green parties block (green colour) and christian parties (orange colour). Similarly, we see that both largest and ideologically opposite parties like SVP (grey) and SP (red) are placed diametrally opposite to each other. The ideological differences between these parties are further reinforced through the social interactions between them. Ideological differences might not be the only reasons why parties are split far apart. BDP (yellow) and SVP (grey) are both right parties but are located in opposition to each other, so that no cross-likes to posts appear between them. One possible explanation is the scandal of expulsion of some members of SVP party that eventually formed BDP. These descriptive findings further confirm the presence of not only ideological but also activity based polarization among politicians in other online platforms.

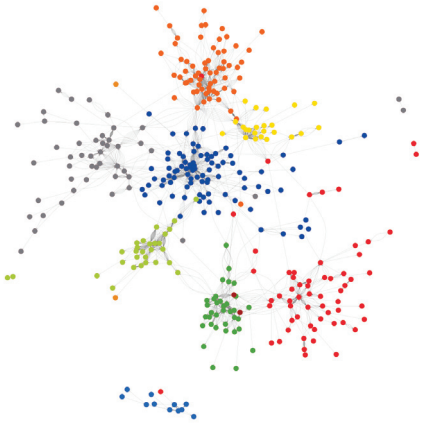


Figure 6.3: The network of likes among politicians on Facebook.

6.2.4 Success and Failure of Politicians

Finally, we are planning to explore factors that determine success and failure of individual politicians and not the parties in elections. Previous elections in Switzerland have unexpectedly resulted in non-election of favorable candidates, and similarly unexpected vote on outsiders. As predictor factors we consider three broad categories: political stress, online behaviour and activity of politicians, and mass media coverage and popular reactions.

A proxy for the mass media coverage can be Twitter accounts of journalists, newspaper and TV. Hashtags and mentions of users, as well as the emotions expressed in online comments from the local most read newspaper 20min.ch can serve as proxies of popular reactions. Stress level of politician encompasses the frequency of scandals politician has been involved and his winning expectancy. One of the hypothesis is that politicians that have 50% chance to win election will have higher stress level than those that have no chance or very high chance to be elected. And therefore, politicians with high stress level will engage in more emotional online activity by creating posts of high emotional arousal or negative valence. Cacioppo's theory predicts that emotional posts will trigger

emotional polarized response from public. Through this study we are able to realize one of our additional goals to test Cacioppo's theory in political domain.

6.3 The Role of Collective Evaluations

In Section 6.2, we have mostly outlined the direction of future work in the domain of political polarization. In the two following sections, we open the discussion on the role of collective evaluations in online platforms. As we have seen in Chapter 4, large amount of both `likes` and `dislikes` manifests polarized opinion towards online item. In Section 6.3.1, we present how a lightweight activity such as liking has become an important social interaction tool among individuals. We call the first section "Generation ``Like''" (Lee, 2015) to reflect the naming pattern of demographic cohorts, like Generation X, Generation Y or Generation Z. In Section 6.3.2, we look at the opposite of `like` – a `dislike` button – and present ideas on how disliking on social platforms can influence user participation.

6.3.1 Generation ``Like''

The Internet offers plethora of various types of information and an abundance of ways for people to interact socially and professionally without big costs and efforts. Diverse opportunities offered by the Internet and the rise and ease of gadgets motivate individuals to stay connected.

Below is an excerpt from the report of the Pew Research Center's Internet & American Life Project dated back to 2001 (Lenhart *et al.*, 2001). The report centered around the life and experiences of teenagers online at the beginning of the new millenium:

> *The Internet is the telephone, television, game console, and radio wrapped up in one for most teenagers and that means it has become a major "player" in many American families. Teens go online to chat with their friends, kill boredom, see the wider world, and follow the latest trends. Many enjoy doing all those things at the same time during their online sessions. Multitasking is their way of life. And the emotional hallmark of that life is enthusiasm for the new ways the Internet lets them connect with friends, expand their social networks, explore their identities, and learn new things.*

12 years forward, the Pew Center research by Madden *et al.* (2013) has provided an updated statistics on the usage of online media and the Internet and found that eight in ten online young adults and teens now use social media sites, and nine in ten young people are online.

Online communication and exposure to diverse and new information on the Internet amends individuals' behaviour and characteristics. Recent report by Microsoft Canada

(Consumer Insights) (2015) have found that the overload of information on the Internet and the usage of mobile devices lead to shortening of the human attention span. How many times has the reader been distracted by mobile phone, online news or a social network message while reading this thesis? Human attention span is defined as the amount of concentrated time one can spend on a task without becoming distracted. In 2000 it was measured to be 12 seconds, while in 2015 it has reduced to 8 seconds. Digital lifestyle made us being less focused than a goldfish (attention span of 9 seconds).

> *What information consumes is the attention of its recipients. Hence a wealth of information creates a poverty of attention.*
>
> Herbert Simon, Nobel winner in Economics (1978)

With the current development of shortened focus and rapid filtering of online content, the question of how it is possible for users to hijack attention of others becomes very relevant. Users find different ways to "diversify" their social media portfolio (Madden *et al.*, 2013) in order to attract the Internet audience to their online content. In chase of increasing the web clicks on their content, they encourage other individuals to like and comment on their posts, to rate their photos and to engage in various other social interactions.

"Like" is already officially embedded in various social media under various instances, *e.g.* `Facebook` likes, `YouTube` likes, `Google+` "+1"s, `Twitter` favorites, `Flickr` favorites, `Reddit` upvotes, Pinterest "Re-pin"s, `Amazon` star ratings. Apart from different names, it has different representations: from binary $\{0, 1\}$, ternary $\{-1, 0, +1\}$ to N-ary, *e.g.* ratings [1..10]. Liking online content might carry slightly different meanings, such as "I saw it or I was here", preference, endorsement, agreement, subscription, self-expression, or reciprocal relationship (Lee, 2015), nevertheless the common characteristic of all different connotations of "like" is the expression of a positive association with the online content (Kosinski *et al.*, 2013). Liking is a lightweight activity, it is less costly than a comment, it connects people that share the same interests, and it carries positive connotation, therefore in the battle for online attention it becomes an important metric for the popularity of online content.

"Like" is a also relatively basic digital record of human behaviour (Kosinski *et al.*, 2013). It has been shown that patterns of likes in social media are highly correlated with users' personal traits (Kosinski *et al.*, 2013; Youyou *et al.*, 2015). Based on the dataset of `Facebook` likes, Kosinski *et al.* (2013) have successfully predicted personal and highly sensitive traits of individuals, such as sexual orientation, ethnicity, religious and political views, use of addictive substances, and even parental separation.

"Like" also serves as a manifestation of opinion. Facebook, Inc. (2012), for instance, defines liking activity as an act of expression of an idea and a form of speech. Political campaigns demonstrate the extent to which online media are quintessential in communicating and exchanging opinions and mobilization of users. In our study, we have also shown the

strength of cross party debates, how politicians mobilize and stop liking posts of opponents several months before the elections or how politicians on Facebook only like Facebook pages of in-party candidates. "Where once a neighbor would show allegiance to a political candidate by staking a sign on the front lawn, a user now clicks Like on a candidate's Facebook Page instead" (Robbins, 2013). However, the meaning of "Like" as an online expression of opinion creates gaps in the legal system. In 2009, a deputy sheriff was fired for liking the Facebook page of his boss's political rival. Furthermore, a U.S. district court decreed that the termination did not violate the deputy sheriff's First Amendment rights to free speech (Robbins, 2013), arguing that free-speech protections cannot be applied when someone doesn't actually say something. This and many more cases concerning the expression of opinion on the Internet and potential private or professional consequences should be carefully reconsidered in the legal system.

6.3.2 Influence of `Dislike` Button on Users Participation

As we have seen in Chapter 4, online popularity might bring its adversary side. Platforms, that offer a functionality to dislike content, often have popular items that attain not only great amount of upvotes but simultaneously great amount of downvotes. For instance, viral videos, such as "Baby" by Justin Bieber or "Gagnam style" by PSY, top not only the list of the most viral and popular videos on YouTube, but also the list of the most hated videos on the Internet.

The question of the influence of the dislike button on user participation, which is currently excluded from the design of some of the leading social networking sites, *e.g.* `Facebook` (Cashmore, 2010), still remains. The effect of the social rating system has been well demonstrated in the episode "Nosedive" of the "Black Mirror" TV series (2016). Main character lives in a world enhanced by technological gadgets where anyone is able to rate any person on a positive or a negative scale. While initially possessing a high approval rate, the main character rolls down in the social rating system due to several unfortunate encounters with other strangers. This triggers a snowball effect, with unknown people not willing to help her based on her low rating and downvoting her even further. In 2015, a mobile application for rating people based on their personal, professional and dating relationships has been proposed. Described as "Yelp for People", the developers received a viral negative backlash and a wide criticism over concerns of cyberbullying and online harassment. So far, any attempts to introduce negative evaluation directed at user have been suppressed.

The rationale behind this decision is often attributed to the straightforward assumption that, when users receive explicit dislikes by other users, their participation decreases. Under different assumptions, the presence of the dislike button might be critically necessary for other social media. An example is the "Digg collapse" (Walker and Ante, 2012), in

which a massive amount of users stopped using Digg to start using Reddit, following a platform redesign that disabled the option to dislike content (Tassi, 2012).

An alternative view on the arguments mentioned above is that the possibility to dislike content does not decrease user participation. The main difference between the examples of Facebook and Digg is the nature of the content posted in the online medium. While the content shared in Facebook is very close to the identity of the user that posts it (*e.g.* profile pictures), the content shared on Digg is usually composed of web links that are not authored by the posting user. This divides the role of likes and dislikes as *subject evaluation* in the former case, versus *object evaluation*. As these dislikes are completely anonymous, the social pressure they exert might not be enough to hinder the intrinsic motivation of a user. This still remains an open research question but can be tested with the online data.

6.4 Relevance in Other Fields

Apart from the relevance of this study within research communities, we expect that the scientific output of this project is of use in a broader context.

Our approach of study interaction-based polarization in the Swiss website politnetz.ch has triggered further scientific collaborations with local political scientists who research on direct democracy, e-voting and Voting Advice Applications. Our approach can be later applied in other international case studies with multi-party political configuration. Future work on the success and failure of politicians can shed light on the effectiveness of social media in the mobilization of users during political campaigns. Additionally, by performing the sentiment analysis of the comments section in digitally available newspapers, we can pulse the public opinion on every politician and measure the role of popular reactions on the success or failure of politicians during election campaigns.

Incorporating tools for sensing the emotionality and polarization level of discussions on the Internet gives rise to the question of the design of online communication platforms. One aim is to optimize user participation and engagement, and some social network providers have made certain steps to allow users to express the fine-grained level of the emotions they feel when reading online posts. For instance, in 2016 Facebook changed its "like" button to six different emotional reactions, including the two negative emotions of different arousal, *i.e.* angry and sad. Second aim is to monitor the level of negativity and polarization in the discussion, for instance early detection of cyberbullying, rising social conflict or aggression, online harassement or misogyny might mitigate potentially dangerous consequences for a personal well-being of a participant of online discussion. We expect that research on the mechanisms of the online platform design to either increase involvement and engagement of users or to detect early irrevertable conflicts will be continued.

Finally, management sciences can improve work group cohesion by understanding the role of emotional interaction in decision making. Economics and finance can improve models of systemic risk by introducing the effect of emotional influence in agent decisions. And our work can be also used to provide understanding of viral marketing and how emotions shape opinions towards products and brands.

Appendices

Appendix A

Mathematical Definitions

A.1 Statistical Metrics

Correlation Estimation

The Pearson correlation coefficient is a measure of the association between two variables X and Y, and is defined as the covariance of the two variables divided by the product of their standard deviation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X}) \cdot (y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{Y})^2}},$$

where $\overline{X}$ and $\overline{Y}$ are the mean of each of the random variables, and x_i, y_i are the i-th outcome of the random variables X and Y respectively.

The Jaccard Similarity

The coefficient is used to compare similarity and diversity between two sets. Its value for sets X and Y is defined as a division of the intersection of two sets over the union of these sets. For computing one-sided overlaps, the *Partial Jaccard coefficient* normalizes over one set only. For instance, the J_{XY} normalized over Y gives the fraction of the set Y which is attributed to the set X.

$$J_{XY} = \frac{|X \cap Y|}{|X \cup Y|}, \qquad \text{Partial } J_{XY} = \frac{|X \cap Y|}{|Y|}.$$

Shannon's Entropy

Entropy stands for the measure of uncertainty of the information content, or how much of the uncertainty will be reduced when information is received. If the result of a random variable is difficult to predict, then the entropy, unpredictability or uncertainty, of the outcome is high; conversely, if the result is predictable, it's entropy is low. Entropy is at its highest value, when the outcomes of a random variable are equally probable. Shannon (1948) proved that only a logarithmic function satisfies the properties necessary to measure the entropy, and proposed the following formula:

$$H(A) = -\sum_{i=1}^{m} p(a_i) \cdot \log_b p(a_i),$$

where A is a discrete random variable taking m possible values $(a_1, a_2, \ldots, a_m)$, $p(a_i)$ is the probability of A taking value a_i, and b is the base of the logarithm used which determines the information units of the entropy. For $b = 2$, the entropy is measured in *the number of bits per random variable outcome*. Therefore, *information* can be expressed in terms of the entropy as:

$$I = k \cdot H(A),$$

where k is the number of events. Therefore, with the logarithm of base 2, information is measured in *bits*. Thus, a biased coin that always falls heads, *i.e.* $p(X_{\text{biased}} = \text{Heads}) = 1$ and $p(X_{\text{biased}} = \text{Tails}) = 0$, has the entropy of $H(X_{\text{biased}}) = -(0 \cdot \log_2 0 + 1 \cdot \log_2 1) = 0$ bits per outcome which gives us no information after each throw, *i.e.* $I = 1 \cdot 0 = 0$ bits. After toss of a biased coin, we learned no new information as outcome is known *a priori*, thus no uncertainty is reduced. On the other hand, the entropy of an unbiased coin, *i.e.* $p(X_{\text{unbiased}} = \text{Heads}) = \frac{1}{2}$ and $p(X_{\text{unbiased}} = \text{Tails}) = \frac{1}{2}$, is $H(X_{\text{unbiased}}) = -(\frac{1}{2} \cdot \log_2 \frac{1}{2} + \frac{1}{2} \cdot \log_2 \frac{1}{2}) = 1$ bit per outcome, and the information received is $I = 1 \cdot 1 = 1$ bit.

Miller-Madow Entropy Estimator

Calculation of the entropy from the data samples requires the knowledge on the probabilities of the outcome of the random variable as seen from the Shannon's formula. In practice, $p(a_i)$ is unknown and must be *estimated* from the observed counts of the i-th outcome in the sequence of trials. The simplest and widely used estimator of the entropy is the *Maximum Likelihood estimator (ML)* (Hausser and Strimmer, 2009):

$$\hat{H}^{ML}(A) = -\sum_{i=1}^{m} \hat{p}^{ML}(a_i) \cdot \log_2 \hat{p}^{ML}(a_i),$$

where $\hat{p}^{ML}(a_i)$ is an estimate of the probability of the i-th possible outcome a_i of the random variable, often calculated as $\hat{p}^{ML}(a_i) = \frac{y_i}{n}$, where $y_i \geq 0$ is the number of the observed counts of the i-th outcome out of the n number of observations. When the number of the possible outcomes of the random variable is much lower than the number of observations, $m \ll n$, the ML method gives the optimal estimation of the entropy. On the other hand, the finite sample sizes can leave some outcomes unobserved, and the ML method can underestimate the true entropy.

To overcome the sample size bias, corrections and other estimators of the entropy have been developed (Hausser and Strimmer, 2009). For our empirical estimation of the information content, we employ the *Miller-Madow entropy estimator* (Miller, 1955). In essence, it is the ML entropy with the bias correction:

$$\hat{H}^{MM}(A) = \hat{H}^{ML}(A) + \frac{p_{>0} - 1}{2 \cdot n},$$

where $p_{>0}$ is the number of outcomes with $y_i > 0, \forall i \in [1, m]$, *i.e.* how many of the m possible outcomes were observed in n trials.

Mutual Information

If X and Y are two random variables, then the mutual information between them is expressed as:

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y),$$

where H(X), H(Y) are the marginal entropies, or simply the entropies of X and Y; $H(X,Y)$ is the joint entropy of X and Y, and $H(X|Y)$ is the conditional entropy of X given Y, which is the amount of the uncertainty that remains in X when the value of Y is known, the relationship between the metrics is shown in Figure A.1. The mutual information measures the average reduction of the uncertainty in X that results from learning Y; and vice versa (MacKay, 2002). The measure is symmetric, $I(X;Y) = I(Y;X)$.

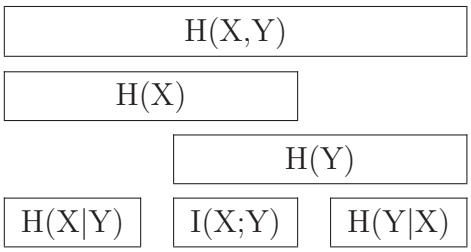


Figure A.1: The relationship between the joint entropy, the marginal entropy, the conditional entropy and the mutual information (MacKay, 2002).

The mutual information can be measured in terms of the probabilities of the outcomes of random variables:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \cdot \log_2 \left(\frac{p(x,y)}{p(x) \cdot p(y)} \right),$$

where $p(x,y)$ is the joint probability distribution of X and Y, and $p(x)$ and $p(y)$ are the marginal probabilities of the outcomes x and y. We compute the entropy and the mutual information using the `Entropy` library of the **R** programming language developed by Hausser and Strimmer (2009).

Normalized Mutual Information

This metric is also known as the *uncertainty coefficient* (Press *et al.*, 2007) of X with respect to Y, and is given by:

$$\mathrm{NMI}(X|Y) = \frac{I(X;Y)}{H(X)} = \frac{H(X) - H(X|Y)}{H(X)}.$$

In essence, the uncertainty coefficient of the dependent variable X with respect to the independent variable Y is the mutual information of both variables $I(X;Y)$ normalized over the entropy of the dependent variable, $H(X)$ or $H(Y)$. The measure lies between 0 and 1, and is interpreted as follows: if the $\mathrm{NMI}(X|Y) = 0$, then there is no relation between X and Y; if the $\mathrm{NMI}(X|Y) = 1$, then the knowledge of Y fully predicts or determines X, or 100% of the information content of X is captured by Y. A value in-between 0 and 1 gives the fraction of the information gain of X when Y is known. Interchanging X and Y in the uncertainty coefficient, $\mathrm{NMI}(Y|X)$, will define the dependence of Y with respect to X, and the normalization is performed over the entropy of Y.

Pointwise Mutual Information

It is the mutual information for *the pairs of outcomes* of two random variables rather than all possible values:

$$\mathrm{PMI}(X = x; Y = y) = \log_2 \frac{p(x,y)}{p(x) \cdot p(y)} = \log_2 \frac{p(x|y)}{p(x)} = \log_2 \frac{p(y|x)}{p(y)},$$

where $p(x,y)$ is the joint probability distribution of the outcomes x and y of the random variables X and Y respectively, and $p(x)$ and $p(y)$ are the marginal probability distributions of the outcomes x and y of the random variables X and Y respectively.

The Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test (K-S test) is a non-parametric test that is used to assess the statistical similarity of two samples, or of a one sample to a reference distribution.

In a two-sample test, the null hypothesis H_0 for the test is that two samples are generated from the identical distribution. The result of the test outputs the observed test statistic D_{obs}, which measures the maximum distance between the two empirical distribution functions, and the p-value, which gives the probability of observing test statistic D equal to or more extreme than what was actually observed given the null hypothesis is correct, $P(D \geq D_{obs}|H_0)$.

The null hypothesis is rejected, if this probability is smaller than or equal to a pre-defined significance level α, which is usually set to 0.05. The significance level α is the probability of falsely rejecting the null hypothesis given that it is correct, known as the type I error. For example, with the p-value $= 0.02$, there is a 2% chance to make an error by rejecting the null hypothesis given that it is correct.

Lyapunov Central Limit Theorem

If $X_1, \ldots, X_n$ are independent but not identically distributed random variables with finite mean and variance $\mathbf{E}[X_i] = \mu_i$ and $\mathbf{VAR}[X_i] = \sigma_i^2$, then

$$
\begin{aligned}
& s_n^2 = \sum_{i=1}^{n} \sigma_i^2, \\
& \frac{1}{s_n} \sum_{i=1}^{n} (X_i - \mu_i) \xrightarrow{d} \mathcal{N}(0,1), \\
& \sum_{i=1}^{n} (X_i - \mu_i) \simeq \mathcal{N}(0, s_n^2), \text{ since } \mathbf{VAR}[aX] = a^2 \mathbf{VAR}[X], \\
& \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i \simeq \mathcal{N}(0, s_n^2), \\
& \sum_{i=1}^{n} X_i \simeq \mathcal{N}(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2), \\
& \sum_{i=1}^{n} X_i \simeq \mathcal{N}(\sum_{i=1}^{n} \mathbf{E}[X_i], \sum_{i=1}^{n} \mathbf{VAR}[X_i]).
\end{aligned}
$$

A.2 Network Theory Definitions

Q-modularity

In the network theory, a network shows a modular structure if it can be partitioned into the groups that are densely interconnected inside and loosely connected to the other groups. The *Q-modularity* metric quantifies the quality of a partition in terms of a modular structure (Newman, 2006), comparing the fraction of links of nodes within the groups to the expected value if links were distributed purely at random, but preserving the nodes' degrees. The formula for a *directed* network is given by:

$$Q = \frac{1}{m} \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} \left[A_{ij} - \frac{k_i^{\text{out}} \cdot k_j^{\text{in}}}{m} \right] \delta(i,j),$$

where N and m are the amounts of nodes and links in the network, k_i^{out} and k_i^{in} are the out-degree and the in-degree of node i, A is the adjacency matrix of the network, and $\delta(i,j)$ is a function that takes the value 1 if nodes i and j are in the same group, and 0 otherwise. For the case of a weighted network, the amount of links is replaced with the sum of weights of all links, and the adjacency matrix has entries corresponding to the weight of each link.

In-degree Centralization

Freeman (1978) introduced the concept of degree centralization, where the average difference between the node degree centralities is normalized over the value of a star network.

$$C_{in} = \frac{\sum_{i=1}^{n} [k_*^{in} - k_i^{in}]}{\max \sum_{i=1}^{n} [k_*^{in} - k_i^{in}]},$$

where k_i^{in} is the in-degree of node i, k_*^{in} is the largest in-degree of the network and $\max \sum_{i=1}^{n} [k_*^{in} - k_i^{in}]$ is the maximum possible sum of differences in the degree centrality, which corresponds to the value of a star network. The numerator represents the sum of differences between the highest degree in the graph given by node $v*$ and the degrees of the other nodes, measuring the extent to which the most central node $v*$ exceeds the in-degree of the other nodes. The denominator stands for the maximum possible value of such difference in the network with the same number of nodes. Normalized over the denominator, the degree centrality is a value in the interval $[0, 1]$ and represents the average deviation of nodes in the network from the most central node.

Average Path Length

This social network metric measures the efficiency of the information transportation in a network:

$$\ell = \frac{1}{n \cdot (n-1)} \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} d(v_i, v_j), \forall i \neq j,$$

where $d(v_i, v_j)$ is the length of the shortest path between nodes v_i and v_j, and n is the number of nodes in the network. In essence, the average path length is the sum of the lengths of the shortest paths between all pairs of nodes in the network divided over the maximum number of all possible pairs in the network where a path can exist, therefore the normaliztion factor is $\frac{1}{n \cdot (n-1)}$. If the network consists of the disconnected components, the normalization factor becomes the number of existing paths in the network:

$$\ell = \frac{1}{\sum_{i=1}^{n} \sum_{j=1}^{n} \delta(i,j)} \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} d(v_i, v_j), \forall i \neq j,$$

where $\delta(i,j) = 1$ if there exists a path between the nodes i and j, 0 – otherwise.

k-core Decomposition

A k-core of a network is a sub-network in which all nodes have a degree $\geq k$. The k-core decomposition is a procedure of finding all k-cores, $\forall k > 0$, by repeatedly pruning nodes with degrees k. Therefore, it captures not only the direct, but also the indirect impact of users leaving the network. As an illustration consider Figure A.2, which shows the process of finding the k-core decomposition:

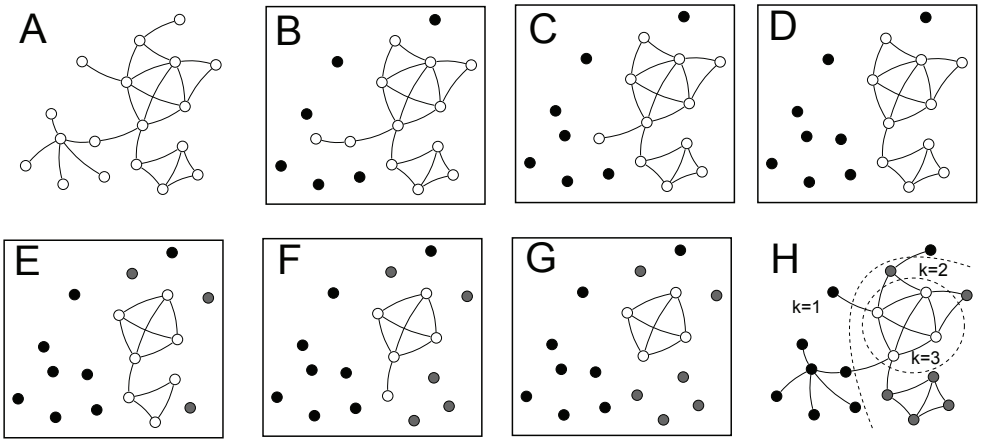


Figure A.2: Effects of node removals on network connectivity as captured by degree only (A → B) and k-core decomposition (A → C → D → E)

Starting again from A, and applying the k-core procedure, will repeatedly remove nodes of degree less than 2, until only those with degree of at least 2 remain in panel D. The

removed nodes up to this point are in the 1-core of the network but not in the 2-core. In a second iteration, all nodes of degree less than 3 are removed, colored dark gray in panel G. These form part of the 2-core of the network, but not of the 3-core, which is composed of the last four nodes. Hence, supposing that users leave a community when they are left with less than 3 friends, the k-core decomposition captures the full cascading effect that departing users have on the network as a whole. More details on the empirical calculation of the k-core decomposition can be found elsewhere (García *et al.*, 2013a).

Demodularity

We introduce an inter-group measure called *demodularity*, which quantifies the relationship across different groups rather than within the group, and can be defined as a property of a network where nodes of one group preferentially attach to the nodes of the other group. Negative scores of the demodularity from group f to group t indicate that nodes of f strongly avoid interactions with nodes of t, contrary to positive scores that show cross-community interactions. The demodularity, $\overline{Q}_{ft}$, from community f to community t in a *directed* network is defined as:

$$\overline{Q}_{ft} = \frac{1}{m_f} \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} \left[A_{ij} - \frac{k_i^{\text{out}} \cdot k_j^{\text{in}}}{m} \right] \delta(C(i), f) \cdot \delta(C(j), t),$$

where $C(i)$ is the group to which node i belongs, $\delta(C(i), f)$ is a function such that $\delta(C(i), f) = 1$ if the group of node i is f, and 0 otherwise, and the rest of the notation is consistent with the earlier definition of Q-modularity.

Political Distance

The Euclidean distance between two points $\mathbf{X}$ and $\mathbf{Y}$ is the length of the line segment connecting them: $\overline{\mathbf{XY}}$. In Cartesian coordinates, if $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$ are two points in the Euclidean n-space, then the distance between $\mathbf{X}$ and $\mathbf{Y}$ is defined as:

$$\mathbf{d}(\mathbf{X}, \mathbf{Y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

Each of the parties $\mathbf{X}$ and $\mathbf{Y}$ has two political position coordinates: *Left-Right (lr)* and *Liberal-Conservative (lc)*. If $\mathbf{X} = (x_{lr}, x_{lc})$ and $\mathbf{Y} = (y_{lr}, y_{lc})$, then the Euclidean party distance is given by:

$$\mathbf{d}(\mathbf{X}, \mathbf{Y}) = \sqrt{(x_{lr} - y_{lr})^2 + (x_{lc} - y_{lc})^2}.$$

Appendix B

Supplementary Material to Chapter 3

B.1 Data Examples

Authors would like to warn the readers of the *explicit language content* in the following data samples.

Mixed emotions I The following messages are the replied messages that were categorized as those expressing both positive and negative feeling to an object or an event, or an ambivalent feeling of "being torn". The `SentiStrength` library outputs to these texts highly positive and highly negative emotion scores of $(+5, -5)$:

> *I fucking love dogs, but I've been bitten by one pretty hard. She broke skin and I have a neat little scar. I still love dogs, but no matter the size, I still get extremely afraid if a dog starts barking (you know that meaner kind of bark?) and tugging at the leash...even if it's not coming at me.*
>
> a comment on `4chan`

> *Actually yes. The great gay riot of 77. An organized group of gay and gay supporters got together to kidnap the Governor of Arizona's son, holding him hostage for a whole sunday morning. It devastating and extremely fabulous.*
>
> a comment in politics thread on `Reddit`

> *Fucking amazing and fucking terrifying.*
>
> a comment on a picture on `Reddit`

Mixed emotions II The following messages are the replied messages that were categorized as those expressing an in-group support (positive emotions) and an out-group hate (negative emotions). The `SentiStrength` library outputs to these texts highly positive and highly negative emotion scores of $(+5, -5)$:

I fucking love ugly people so much! They're so talented! I'm ugly myself, so I can say this! Epic video! I love uglies! Good looking people are fucking morons!

a comment on a video song on YouTube

HATERS VS LOVERS!!! WHO WIN? Bieber have 100.000.000.000 haters (i'm one of them) and 1.000.000 lovers!!! So thumbs up, if you are hater!!!

a comment on a video song in YouTube

Mixed emotions III The following messages are the replied messages that were categorized as those expressing irony. The SentiStrength library outputs to these texts highly positive and highly negative emotion scores of $(+5, -5)$, however words expressing highly positive emotions carry negative connotation, or words of highly negative emotions carry positive meaning:

holy fuck, i lost fucking terrible to this one man. fucking hilarious, you're a genius. stay the fuck in here please.

a comment on 4chan

Excellent! I really hate that movie now. (And I haven't even seen it)

a comment on a trailer of a movie "Star Trek Into Darkness" in YouTube

What a fucking moron you are. You are so retarded it is amazing, wow!

a comment on a comedy video in YouTube

I gave in and now it's on repeat! I fucking love this album! Damn I got the Christmas Spirit like a motherfucker!!!

a comment on a comedy thread in Reddit

Mixed emotions IV The following messages are the replied messages that were categorized as those expressing positive emotions to feeling negative emotion to an object or an event (*e.g.* love to hate), or vice versa – expressing negative feeling to feeling positive emotion (*e.g.* hate to love):

I fucking hate fundamentalists. I hate them with a fucking passion because they're like a fucking plague.

a comment on a video of the channel "The Young Turks" in YouTube

... Who cares about the deaths. Great characters and it's fucking sad, But that is why i fucking love this manga. the sadness and epicness ! Well Played Kishimoto RESPECT !

a comment on a video in YouTube

Your tears taste so sweet But in all seriousness, I fucking love our rivalry. You guys played great today. The best I've seen the Wolverines play all year.

a comment on a football fans discussion thread in Reddit

B.2 Likelihood of Emotions across Online Categories

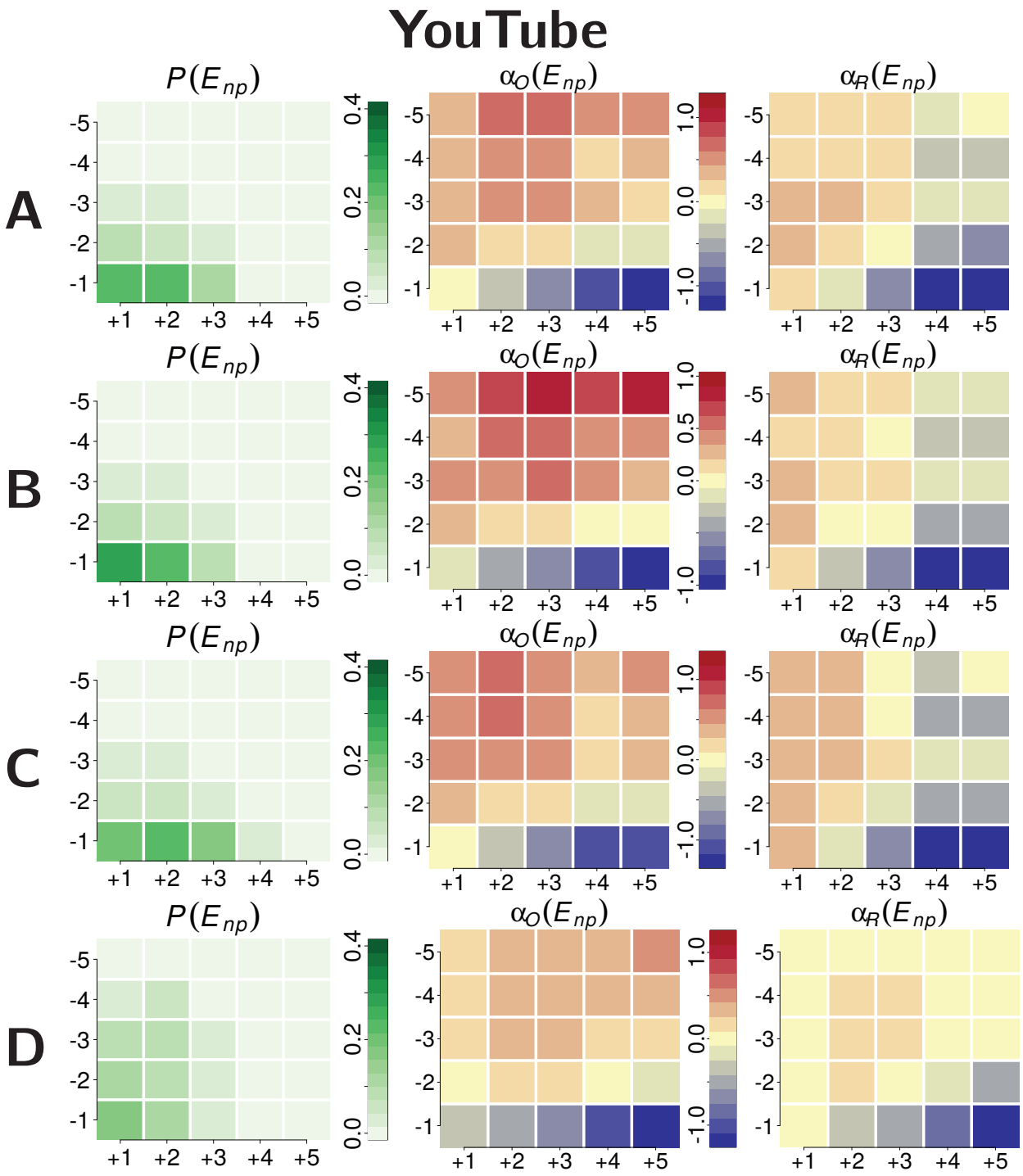


Figure B.1: Normalized log-likelihood of co-occurrence of both positive and negative emotions E_{np} of original message $\alpha_O(E_{np})$ and reply message $\alpha_R(E_{np})$ across responses in YouTube categories: A) Entertainment ($3 \cdot 10^6/20 \cdot 10^6$), B) Gaming ($4 \cdot 10^6/23 \cdot 10^6$), C) Music ($2 \cdot 10^6/13 \cdot 10^6$), D) News & Politics ($4 \cdot 10^6/12 \cdot 10^6$).

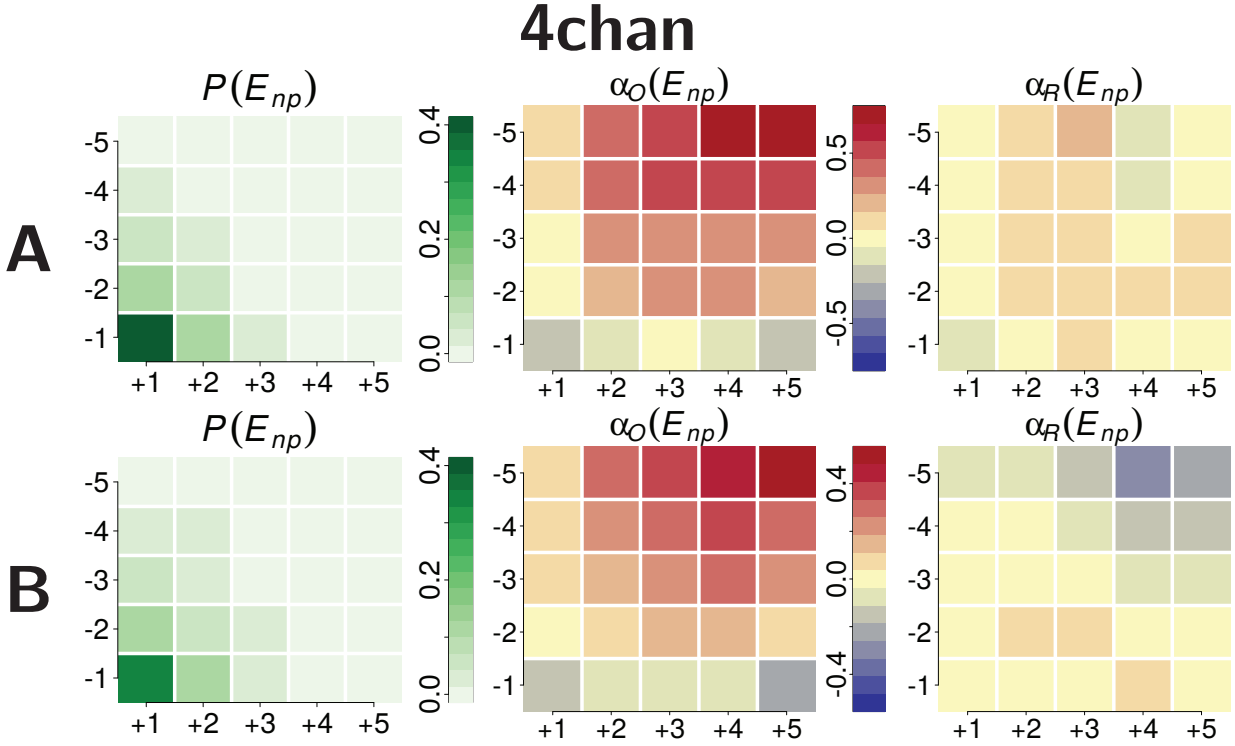


Figure B.2: Normalized log-likelihood of co-occurrence of both positive and negative emotions E_{np} of original message $\alpha_O(E_{np})$ and reply message $\alpha_R(E_{np})$ across responses in `4chan` categories: /b/ $(8.5 \cdot 10^6/25 \cdot 10^6)$, non-/b/ $(7.8 \cdot 10^6/18 \cdot 10^6)$.

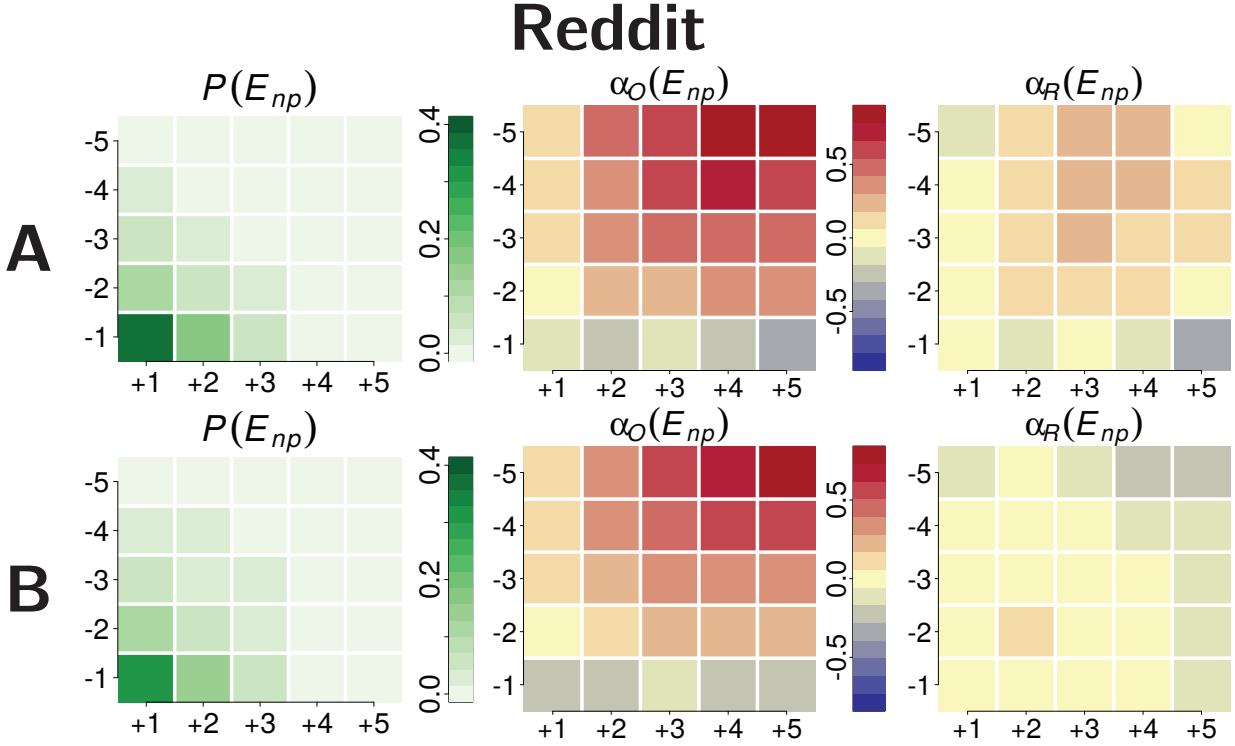


Figure B.3: Normalized log-likelihood of co-occurrence of both positive and negative emotions E_{np} of original message $\alpha_O(E_{np})$ and reply message $\alpha_R(E_{np})$ across responses in `Reddit` categories: A) Funny $(3 \cdot 10^6/6.9 \cdot 10^6)$, B) Rest $(21 \cdot 10^6/44 \cdot 10^6)$.

List of Figures

List of Tables

Bibliography

Abbasi, A.; Hassan, A.; Dhar, M. (2014). Benchmarking Twitter Sentiment Analysis Tools. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Abelson, R. P. (1964). Mathematical models of the distribution of attitudes under controversy. *Contributions to mathematical psychology* **14**, 1–160.

Abisheva, A.; Garimella, V. R. K.; García, D.; Weber, I. (2014). Who watches (and shares) what on youtube? and when?: using twitter to understand youtube viewership. In: *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, pp. 593–602.

Adamic, L. A. (2001). *Network dynamics: The world wide web*. Ph.D. thesis, Citeseer.

Adamic, L. A.; Glance, N. (2005). The political blogosphere and the 2004 US election: divided they blog. In: *Proceedings of the 3rd international workshop on Link discovery*. ACM, pp. 36–43.

Alstott, J.; Bullmore, E.; Plenz, D. (2014). powerlaw: a Python package for analysis of heavy-tailed distributions. *PloS one* **9(1)**, e85777.

Amendola, L.; Marra, V.; Quartin, M. (2015). The evolving perception of controversial movies. *Palgrave Communications* **1**.

Aragón, P.; Kappler, K. E.; Kaltenbrunner, A.; Laniado, D.; Volkovich, Y. (2013). Communication dynamics in twitter during political campaigns: The case of the 2011 Spanish national election. *Policy & Internet* **5(2)**, 183–206.

Asur, S.; Huberman, B. A.; Szabò, G.; Wang, C. (2011). Trends in Social Media: Persistence and Decay. In: *ICWSM*. The AAAI Press.

Axelrod, R. (1997a). Advancing the art of simulation in the social sciences. In: *Simulating social phenomena*, Springer. pp. 21–40.

Axelrod, R. (1997b). The dissemination of culture a model with local convergence and global polarization. *Journal of conflict resolution* **41(2)**, 203–226.

Baldassarri, D.; Bearman, P. (2007). Dynamics of political polarization. *American sociological review* **72(5)**, 784–811.

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature* **435(7039)**, 207–211.

Berrios, R.; Totterdell, P.; Kellett, S. (2015). Eliciting mixed emotions: a meta-analysis comparing models, types, and measures. *Frontiers in psychology* **6**.

Blau, I.; Caspi, A. (2010). Studying invisibly: Media naturalness and learning. In: *Evolutionary Psychology and Information Systems Research*, Springer. pp. 193–216.

Blau, P. M. (1977). *Inequality and heterogeneity: A primitive theory of social structure*, vol. 7. Free Press New York.

Blumer, H. (1951). Collective behavior. *New outline of the principles of sociology* , 166–222.

Bollen, J.; Mao, H.; Pepe, A. (2011). Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In: *International AAAI Conference on Weblogs and Social Media*.

Bradley, M. M.; Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings. Tech. rep.*, Technical report C-1, the center for research in psychophysiology, University of Florida.

Burke, M.; Marlow, C.; Lento, T. (2010). Social network activity and social well-being. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 1909–1912.

Cacioppo, J. T.; Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin* **115**, 401–423.

Cacioppo, J. T.; Gardner, W. L. (1999). EMOTION. *Annual Review of Psychology* **50(1)**, 191–214.

Cacioppo, J. T.; Gardner, W. L.; Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: the case of attitudes and evaluative space. *Personality and Social Psychology Review* **1(1)**, 3–25.

Cardillo, A.; Gómez-Gardenes, J.; Zanin, M.; Romance, M.; Papo, D.; del Pozo, F.; Boccaletti, S. (2012). Emergence of network features from multiplexity. *arXiv preprint arXiv:1212.2153* .

Cashmore, P. (2010). Should Facebook add a dislike button? In: *CNN articles*.

Centola, D.; Gonzalez-Avella, J. C.; Eguiluz, V. M.; San Miguel, M. (2007). Homophily, cultural drift, and the co-evolution of cultural groups. *Journal of Conflict Resolution* **51(6)**, 905–929.

Chmiel, A.; Sienkiewicz, J.; Thelwall, M.; Paltoglou, G.; Buckley, K.; Kappas, A.; Hołyst, J. A. (2011a). Collective emotions online and their influence on community life. *CoRR* **abs/1107.2647**.

Chmiel, A.; Sobkowicz, P.; Sienkiewicz, J.; Paltoglou, G.; Buckley, K.; Thelwall, M.; Hołyst, J. A. (2011b). Negative emotions boost user activity at BBC forum. *Physica A: statistical mechanics and its applications* **390(16)**, 2936–2944.

Christophe, V.; Rimé, B. (1997). Exposure to the social sharing of emotion: Emotional impact, listener responses and secondary social sharing. *European Journal of Social Psychology* **27(1)**, 37–54.

Cleveland, W. S.; Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83(403)**, 596–610.

Conover, M.; Ratkiewicz, J.; Francisco, M. R.; Gonçalves, B.; Menczer, F.; Flammini, A. (2011a). Political Polarization on Twitter. *ICWSM* **133**, 89–96.

Conover, M. D.; Gonçalves, B.; Flammini, A.; Menczer, F. (2012). Partisan asymmetries in online political activity. *EPJ Data Science* **1(1)**, 1.

Conover, M. D.; Gonçalves, B.; Ratkiewicz, J.; Flammini, A.; Menczer, F. (2011b). Predicting the political alignment of twitter users. In: *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, pp. 192–199.

Conway, B. (2015). Why do we care about the colour of the dress? `https://www.theguardian.com/commentisfree/2015/feb/27/colour-dress-optical-illusion-social-media`. Accessed: 2016-11-08.

Cornell University (2012). How groups voted. `http://ropercenter.cornell.edu/polls/us-elections/how-groups-voted/`. Accessed: 2016-11-08.

Crane, R.; Schweitzer, F.; Sornette, D. (2010). Power law signature of media exposure in human response waiting time distributions. *Physical Review E* **81(5)**, 056101.

Crane, R.; Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences* **105(41)**, 15649–15653.

Crow, E. L.; Shimizu, K. (1988). *Lognormal distributions: Theory and applications*, vol. 88. M. Dekker New York.

Cuddeback, G.; Wilson, E.; Orme, J. G.; Combs-Orme, T. (2004). Detecting and statistically correcting sample selection bias. *Journal of Social Service Research* **30(3)**, 19–33.

De Choudhury, M.; Counts, S.; Gamon, M. (2012). Not All Moods are Created Equal! Exploring Human Emotional States in Social Media. In: *ICWSM*.

De Choudhury, M.; De, S. (2014). Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. In: *ICWSM*.

De Choudhury, M.; Gamon, M.; Counts, S.; Horvitz, E. (2013). Predicting Depression via Social Media. In: *ICWSM*. p. 2.

Deffuant, G.; Neau, D.; Amblard, F.; Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems* **3(01n04)**, 87–98.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association* **69(345)**, 118–121.

Derks, D.; Fischer, A. H.; Bos, A. E. (2008). The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior* **24(3)**, 766–785.

DiMaggio, P.; Evans, J.; Bryson, B. (1996). Have American's Social Attitudes Become More Polarized? *American Journal of Sociology* **102(3)**, 690–755.

Dodds, P. S.; Danforth, C. M. (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies* **11(4)**, 441–456.

Duch, J.; Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Phys Rev E* **72(027104)**.

Easley, D.; Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

Elliot, A. J. (2015). Color and psychological functioning: a review of theoretical and empirical work. *Frontiers in psychology* **6**, 368.

Esteban, J.; Schneider, G. (2008). Polarization and conflict: Theoretical and empirical issues. *Journal of Peace Research* **45(2)**, 131–141.

Facebook, Inc. (2012). Brief of Facebook, Inc. as Amicus Curiae in Support of Plaintiff-Appellant Daniel Ray Carter, Jr. and in Support of Vacatur at 1, Bland v. Roberts, 857 F. Supp. 2d 599 (E.D. Va. 2012) (No. 12-16771). `http://www.aclu.org/files/assets/bland_v._roberts_appeal_-__facebook_amicus_brief.pdf`. Accessed: 2016-11-08.

Feldman, L. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology* **69**, 153–166.

Festinger, L. (1957). Cognitive dissonance theory. *1989) Primary Prevention of HIV/AIDS: Psychological Approaches. Newbury Park, California, Sage Publications* .

Finucane, M. L.; Alhakami, A.; Slovic, P.; Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making* **13(1)**, 1–17.

Fiorina, M. P.; Abrams, S. J. (2008). Political polarization in the American public. *Annu. Rev. Polit. Sci.* **11**, 563–588.

Flache, A.; Macy, M. W. (2011). Small worlds and cultural polarization. *The Journal of Mathematical Sociology* **35(1-3)**, 146–176.

Flache, A.; Mäs, M. (2008). How to get the timing right? A computational model of how demographic faultlines undermine team performance and how the right timing of contacts can solve the problem. *Comput Math Organ Theory* **14**, 23–51.

Fontaine, J. R.; Scherer, K. R.; Roesch, E. B.; Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science* **18(12)**, 1050–1057.

Ford, D. (2015). What color is this dress? `http://edition.cnn.com/2015/02/26/us/blue-black-white-gold-dress/`. Accessed: 2016-11-08.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks* , 215.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics* , 1–67.

Friedman, J. H. (1993). *Fast MARS. Tech. rep.*, Dept. of Statistics, Stanford University.

Fruchterman, T. M. J.; Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience* **21(11)**, 1129–1164.

Garas, A.; García, D.; Skowron, M.; Schweitzer, F. (2012). Emotional persistence in online chatting communities. *Scientific Reports* , 402–436.

García, D.; Garas, A.; Schweitzer, F. (2011). Positive words carry less information than negative words. *CoRR* **abs/1110.4123**.

García, D.; Garas, A.; Schweitzer, F. (2012a). Positive words carry less information than negative words. *EPJ Data Science* **1(1)**, 3.

García, D.; Mavrodiev, P.; Schweitzer, F. (2013a). Social resilience in online communities: The autopsy of friendster. In: *Proceedings of the first ACM conference on Online social networks*. ACM, pp. 39–50.

García, D.; Mendez, F.; Serdült, U.; Schweitzer, F. (2012b). Political polarization and popularity in online participatory media: an integrated approach. In: *Proceedings of the first edition workshop on Politics, elections and data*. ACM, pp. 3–10.

García, D.; Tanase, D. (2013). Measuring cultural dynamics through the eurovision song contest. *Advances in Complex Systems* **16(08)**, 1350037.

García, D.; Zanetti, M. S.; Schweitzer, F. (2013b). The Role of Emotions in Contributors Activity: A Case Study on the GENTOO Community. *CoRR* **abs/1306.3612**.

Gayo Avello, D.; Metaxas, P. T.; Mustafaraj, E. (2011). Limits of electoral predictions using twitter. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence.

Gentzkow, M.; Shapiro, J. M. (2010). *Ideological segregation online and offline. Tech. rep.*, National Bureau of Economic Research.

Germann, M.; Mendez, F.; Wheatley, J.; Serdült, U. (2014). Spatial maps in voting advice applications: The case for dynamic scale validation. *Acta Politica* .

Gibrat, R. (1930). -. *Bulletin de la Statistique Générale de la France* **19(469)**.

Gibrat, R. (1931). Les Inégalités économiques. Paris, France.

Gil, S.; Le Bigot, L. (2015). Colour and emotion: children also associate red with negative valence. *Developmental science* .

Gilbert, E.; Karahalios, K. (2010). Widespread Worry and the Stock Market. In: *ICWSM*. pp. 59–65.

Golder, S. A.; Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* **333(6051)**, 1878–1881.

Gómez, S. (2011). Radatools software, Communities detection in complex networks and other tools. `http://deim.urv.cat/~sgomez/radatools.php`. [Online; accessed 27-October-2013].

Goncalves, B.; Perra, N.; Vespignani, A. (2011). Validation of Dunbar's number in Twitter conversations. *arXiv preprint arXiv:1105.5170* .

Gonzalez-Bailon, S.; Banchs, R. E.; Kaltenbrunner, A. (2010). Emotional Reactions and the Pulse of Public Opinion: Measuring the Impact of Political Events on the Sentiment of Online Discussions. *CoRR* **abs/1009.4019**.

Good, P. I. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer, 3 edn.

Gorn, G.; Pham, M. T.; Sin, L. Y. (2001). When arousal influences ad evaluation and valence does not (and vice versa). *Journal of consumer Psychology* **11(1)**, 43–55.

Groeber, P.; Schweitzer, F.; Press, K. (2009). How groups can foster consensus: The case of local cultures. *Journal of Aritifical Societies and Social Simulation* **12(2)**, 1–22.

Gruzd, A.; Roy, J. (2014). Investigating political polarization on twitter: A canadian perspective. *Policy & Internet* **6(1)**, 28–45.

Guerra, P. H. C.; Meira Jr, W.; Cardie, C.; Kleinberg, R. (2013). A Measure of Polarization on Social Media Networks Based on Community Boundaries. In: *ICWSM*.

Hausser, J.; Strimmer, K. (2009). Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. *J. Mach. Learn. Res.* **10**, 1469–1484.

Havlin, S. (1995). The distance between Zipf plots. *Physica A Statistical and Theoretical Physics* **216(1-2)**, 148–150.

Heath, C.; Bell, C.; Sternberg, E. (2001). Emotional selection in memes: the case of urban legends. *Journal of personality and social psychology* **81(6)**, 1028.

Hefte, R. (2015). How can we benefit from consensus decision-making? `http://www.extension.umn.edu/community/civic-engagement/tip-sheets/consensus-decision-making/`. Accessed: 2016-11-08.

Hegselmann, R.; Krause, U.; *et al.* (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation* **5(3)**.

Herman, E. S.; Chomsky, N. (2010). *Manufacturing consent: The political economy of the mass media*. Random House.

Hermann, M.; Städler, I. (2014). Wie sich die SVP aus dem Bürgerblock verabschiedet hat. WWW page.

Hetherington, M. J. (2009). Review article: Putting polarization in perspective. *British Journal of Political Science* **39(02)**, 413–448.

Hoang, T.-A.; Cohen, W. W.; Lim, E.-P.; Pierce, D.; Redlawsk, D. P. (2013). Politics, sharing and emotion in microblogs. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, pp. 282–289.

Huberman, B. A.; Adamic, L. A. (1999). Internet: growth dynamics of the world-wide web. *Nature* **401(6749)**, 131–131.

Iosub, D.; Laniado, D.; Castillo, C.; Morell, M. F.; Kaltenbrunner, A. (2014). Emotions under discussion: Gender, status and communication in online collaboration. *PloS one* **9(8)**, e104880.

Isen, A. M.; Shalker, T. E.; Clark, M.; Karp, L. (1978). Affect, accessibility of material in memory, and behavior: A cognitive loop? *Journal of personality and social psychology* **36(1)**, 1.

Jiang, Z.-Q.; Xie, W.-J.; Li, M.-X.; Podobnik, B.; Zhou, W.-X.; Stanley, H. E. (2013). Calling patterns in human communication dynamics. *Proceedings of the National Academy of Sciences* **110(5)**, 1600–1605.

Jo, H.-H.; Karsai, M.; Kertész, J.; Kaski, K. (2012). Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics* **14(1)**, 013055.

Kappas, A. (2013). Social regulation of emotion: messy layers. *Frontiers in psychology* **4**.

Kappas, A.; Küster, D.; Theunis, M.; Tsankova, E. (2010). CyberEmotions: Subjective and physiological responses to reading online discussion forums. Poster presented at the 50th Annual Meeting of the Society for Psychophysiological Research, Portland, Oregon.

Kappas, A.; Tsankova, E.; Theunis, M.; Küster, D. (2011). Cyberemotions: Subjective and physiological responses elicited by contributing to online discussion forums. Poster presented at the 51st Annual Meeting of the Society for Psychophysiological Research, Boston, Massachusetts.

Kapteyn, J. (1903). Skew Frequency curves in biology and statistics. Astronomical Laboratory, Noordhoff, Groningen.

Khatib, L.; Dutton, W. H.; Thelwall, M. (2012). Public Diplomacy 2.0: An Exploratory Case Study of the US Digital Outreach Team. *The Middle East Journal* .

Kivran-Swaine, F.; Brody, S.; Diakopoulos, N.; Naaman, M. (2012). Of Joy and Gender: Emotional Expression in Online Social Networks. In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion*. CSCW '12, pp. 139–142.

Konnikova, M. (2013). The psychology of online comments. *The New Yorker* .

Kosinski, M.; Stillwell, D.; Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* **110(15)**, 5802–5805.

Kucuktunc, O.; Cambazoglu, B. B.; Weber, I.; Ferhatosmanoglu, H. (2012). A large-scale sentiment analysis for Yahoo! answers. In: *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, pp. 633–642.

Kuhbandner, C.; Pekrun, R. (2013). Joint effects of emotion and color on memory. *Emotion* **13(3)**, 375.

Kühne, R. (2012). Media-induced affects and opinion formation: How related and unrelated affects influence political opinions. *Living Reviews in Democracy* **3**.

Kühne, R.; *et al.* (2012). Political news, emotions, and opinion formation: Toward a model of emotional framing effects. In: *Annual Conference of the International Communication Association (ICA), Phoenix, AZ*.

Kuppens, P.; Oravecz, Z.; Tuerlinckx, F. (2010). Feelings change: accounting for individual differences in the temporal dynamics of affect. *Journal of personality and social psychology* **99(6)**, 1042.

Kushin, M.; Kitchener, K. (2009). Getting political on social network sites: Exploring online political discourse on Facebook. *First Monday* **14(11)**.

Küster, D.; Tsankova, E.; Theunis, M.; Kappas, A. (2011). Measuring cyberemotions: How do bodily responses relate to the digital world? Poster presented at the 7th Conference of the Media Psychology Division of the Deutsche Gesellschaft für Psychologie, Bremen, Germany.

Lang, K.; Lang, G. E. (1962). Collective Dynamics: Process and Form. In: A. Rose (ed.), *Human behavior and social processes: an interactionist approach*, International library of sociology and social reconstruction, Houghton Mifflin. pp. 340–360.

Larsen, P. (2011). O.C.'s Rebecca Black talks about 'Friday'. `http://www.ocregister.com/articles/-292662--.html`. Accessed: 2016-02-02.

Lee, D. (2015). LIKE and Recommendation in Social Media. `http://pike.psu.edu/publications/tutorial-www15.pdf`. Accessed: 2016-11-08.

Lenhart, A.; Simon, M.; Graziano, M. (2001). *The Internet and Education: Findings of the Pew Internet & American Life Project. Tech. rep.*, Pew Research Center.

Leskovec, J.; Adamic, L. A.; Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)* **1(1)**, 5.

Leskovec, J.; Backstrom, L.; Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 497–506.

Levendusky, M. S. (2013). Partisan Polarization in the U.S. Electorate. *Oxford Bibliographies Online* .

Lietz, H.; Wagner, C.; Bleier, A.; Strohmaier, M. (2014). When politicians talk: Assessing online conversational practices of political parties on twitter. In: *International AAAI Conference on Weblogs and Social Media (ICWSM2014), Ann Arbor, MI, USA*.

Lin, M.; Lucas, H. C.; Shmueli, G. (2013). Too Big to Fail: Large Samples and the p-Value Problem. *Information Systems Research* .

Lin, Y.-R. (2014). Assessing Sentiment Segregation in Urban Communities. In: *Proceedings of the 2014 International Conference on Social Computing*. pp. 9:1–9:8.

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

Lorenz, J. (2009). Universality in movie rating distributions. *The European Physical Journal B* **71(2)**, 251–258.

MacKay, D. J. C. (2002). *Information Theory, Inference & Learning Algorithms*. New York, NY, USA: Cambridge University Press.

Macy, M. W.; Kitts, J. A.; Flache, A.; Benard, S. (2003). Polarization in dynamic networks: A Hopfield model of emergent structure. *Dynamic social network modeling and analysis* , 162–173.

Madden, M.; Lenhart, A.; Duggan, M.; Cortesi, S.; Gasser, U. (2013). *Teens and technology 2013*. Pew Internet & American Life Project Washington, DC.

Malmgren, R. D.; Stouffer, D. B.; Motter, A. E.; Amaral, L. A. (2008). A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences* **105(47)**, 18153–18158.

Mano, H. (1999). The influence of pre-existing negative affect on store purchase intentions. *Journal of Retailing* **75(2)**, 149–172.

Mäs, M. (2012). Understanding and Solving Societal Problems with Modeling and Simulation. `https://www.ethz.ch/content/dam/ethz/special-interest/gess/computational-social-science-dam/documents/education/Fall2012/crowds/Lecture11_social_influence.pdf`. Accessed: 2016-11-08.

Mäs, M.; Flache, A. (2013). Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. *PloS one* **8(11)**, e74516.

Mäs, M.; Flache, A.; Takács, K.; Jehn, K. A. (2013). In the short term we divide, in the long term we unite: Demographic crisscrossing and the effects of faultlines on subgroup polarization. *Organization science* **24(3)**, 716–736.

McCarty, N.; Poole, K. T.; Rosenthal, H. (2006). *Polarized America: The dance of ideology and unequal riches*, vol. 5. mit Press.

McCoy, T. (2015). The inside story of the 'white dress, blue dress' drama that divided a planet. `https://www.washingtonpost.com/news/morning-mix/wp/2015/02/27/the-inside-story-of-the-white-dress-blue-dress-drama-that-divided-a-nation/`. Accessed: 2016-11-08.

McPhee, W. N. (1963). *Formal theories of mass behavior*. London: The Free Press of Glencoe, Collier-Macmillian, 244 pp.

Menichetti, G.; Remondini, D.; Panzarasa, P.; Mondragón, R. J.; Bianconi, G. (2013). Weighted Multiplex Networks. *CoRR* **abs/1312.6720**.

Microsoft Canada (Consumer Insights) (2015). Attention Spans. https://advertising.microsoft.com/en/WWDocs/User/display/cl/researchreport/31966 /en/microsoft-attention-spans-research-report.pdf. Accessed: 2016-11-08.

Miller, G. A. (1955). Note on the bias of information estimates. *Information Theory in Psychology: Problems and Methods* , 95–100.

Miller, M.; Sathi, C.; Wiesenthal, D.; Leskovec, J.; Potts, C. (2011). Sentiment Flow Through Hyperlink Networks. In: *ICWSM*.

Miller, N. E. (1961). Some recent studies on conflict behaviour and drugs. *American Psychologist* **16(1)**, 12–24.

Mitrović, M.; Paltoglou, G.; Tadić, B. (2010). Networks and emotion-driven user communities at popular blogs. *The European Physical Journal B* **77(4)**, 597–609.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet mathematics* **1(2)**, 226–251.

Moscovici, S.; Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of personality and social psychology* **12(2)**, 125.

Myers, D. G.; Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin* **83(4)**, 602.

Nakatani, S. (2010–current). Language Detection Library for Java. `https://github.com/shuyo/language-detection/blob/wiki/ProjectHome.md`.

Newman, M. (2003). Fast algorithm for detecting community structure in networks. *Physical Review E* **69**.

Newman, M. E. (2006). Modularity and community structure in networks. *Proc Natl Acad Sci U S A* **103(23)**, 8577–8582.

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of communication* **24(2)**, 43–51.

Norman, G. J.; Norris, C. J.; Gollan, J.; Ito, T. A.; Hawkley, L. C.; Larsen, J. T.; Cacioppo, J. T.; Berntson, G. G. (2011). Current Emotion Research in Psychophysiology: The Neurobiology of Evaluative Bivalence. *Emotion review* **3(3)**, 349–359.

Oliveira, J. G.; Barabási, A.-L. (2005). Human dynamics: The correspondence patterns of Darwin and Einstein. *arXiv preprint physics/0511006* .

Olson, R. (2015). A data-driven guide to creating successful reddit posts, redux. Accessed: 2016-11-08.

Onnela, J.-P.; Reed-Tsochas, F. (2010). Spontaneous emergence of social influence in online systems. *Proceedings of the National Academy of Sciences* **107(43)**, 18375–18380.

Osgood, C. E. (1969). On the whys and wherefores of E.P. and A. *Journal of Personality and Social Psychology* **12(3)**, 194–199.

Pang, B.; Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval* **2(1-2)**, 1–135.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Paulhus, D. L.; Lim, D. T. (1994). Arousal and evaluative extremity in social judgments: A dynamic complexity model. *European Journal of Social Psychology* **24(1)**, 89–99.

Pennebaker, J. W.; Francis, M. E.; Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* **71**, 2001.

Pfitzner, R.; Garas, A.; Schweitzer, F. (2012). Emotional Divergence Influences Information Spreading in Twitter. In: J. G. Breslin; N. B. Ellison; J. G. Shanahan; Z. Tufekci (eds.), *ICWSM*. The AAAI Press.

Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. New York, NY, USA: Cambridge University Press, 3 edn.

Quercia, D.; Capra, L.; Crowcroft, J. (2012). The Social World of Twitter: Topics, Geography, and Emotions. *ICWSM* **12**, 298–305.

Quercia, D.; O'Hare, N. K.; Cramer, H. (2014a). Aesthetic capital: what makes london look beautiful, quiet, and happy? In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, pp. 945–955.

Quercia, D.; Schifanella, R.; Aiello, L. M. (2014b). The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In: *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, pp. 116–125.

Rafaeli, S.; Sudweeks, F. (1997). Networked Interactivity. *Journal of Computer-Mediated Communication* **2(4)**, 0–0.

Reisenzein, R. (1983). The Schachter theory of emotion: two decades later. *Psychological bulletin* **94(2)**, 239.

Rimé, B. (2009). Emotion Elicits the Social Sharing of Emotion: Theory and Empirical Review. *Emotion Review* **1(1)**, 60–85.

Robbins, I. P. (2013). What Is the Meaning of'Like'?: The First Amendment Implications of Social-Media Expression. *The First Amendment Implications of Social-Media Expression (June 18, 2013)* .

Russell, J. A. (1979). Affective Space is Bipolar. *Journal of Personality and Social Psychology* .

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology* **39**, 1161–1178.

Russell, J. A.; Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology* **76(5)**, 805–819.

Russell, J. A.; Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of research in Personality* **11(3)**, 273–294.

Salganik, M. J.; Dodds, P. S.; Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *science* **311(5762)**, 854–856.

Saunders, K. L.; Abramowitz, A. I. (2004). Ideological realignment and active partisans in the American electorate. *American Politics Research* **32(3)**, 285–309.

Schulz, A.; Roessler, P. (2012). The spiral of silence and the internet: Selection of online content and the perception of the public opinion climate in computer-mediated communication environments. *International Journal of Public Opinion Research* **24(3)**, 346–367.

Schweitzer, F.; Behera, L. (2009). Nonlinear voter models: the transition from invasion to coexistence. *The European Physical Journal B* **67(3)**, 301–318.

Schweitzer, F.; García, D. (2010). An agent-based model of collective emotions in online communities. *The European Physical Journal B* **77(4)**, 533–545.

Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks* **5(3)**, 269 – 287.

Serdült, U. (2014). Referendums in Switzerland. In: *Referendums around the World*, Springer. pp. 65–121.

Shafir, E.; Simonson, I.; Tversky, A. (1993). Reason-based choice. *Cognition* **49(1-2)**, 11–36.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, 379–423, 623–656.

Siegel, J.; Dubrovsky, V.; Kiesler, S.; McGuire, T. W. (1986). Group processes in computer-mediated communication. *Organizational behavior and human decision processes* **37(2)**, 157–187.

Simmel, G. (1906). The Sociology of Secrecy and of Secret Societies. *American Journal of Sociology* **11**, 441–498.

Sobkowicz, P.; Sobkowicz, A. (2010). Dynamics of hate based Internet user networks. *The European Physical Journal B* **73(4)**, 633–643.

Stone, P. J.; Dunphry, D. C.; Smith, M. S.; Ogilvie, D. M. (1966). The General Inquirer: A Computer Approach to Content Analysis.

Sunstein, C. R. (2002). The law of group polarization. *Journal of political philosophy* **10(2)**, 175–195.

Sunstein, C. R. (2003). *Why societies need dissent*. Harvard University Press.

Szabo, G.; Huberman, B. A. (2010). Predicting the popularity of online content. *Communications of the ACM* **53(8)**, 80–88.

Szell, M.; Lambiotte, R.; Thurner, S. (2010). Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences* .

Szell, M.; Thurner, S. (2010). Measuring social dynamics in a massive multiplayer online game. *Social networks* **32(4)**, 313–329.

Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics* **37(2)**, 267–307.

Tan, C.; Lee, L.; Tang, J.; Jiang, L.; Zhou, M.; Li, P. (2011). User-level Sentiment Analysis Incorporating Social Networks. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '11, pp. 1397–1405.

Tassi, P. (2012). Facebook Didn't Kill Digg, Reddit Did. In: *Forbes*.

Taylor, J.; Macdonald, J. (2002). The effects of asynchronous computer-mediated group interaction on group processes. *Social Science Computer Review* **20(3)**, 260–274.

Thelwall, M.; Buckley, K.; Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology* **62(2)**, 406–418.

Thelwall, M.; Buckley, K.; Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology* **63(1)**, 163–173.

Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; Kappas, A. (2010a). Sentiment in Short Strength Detection Informal Text. *J. Am. Soc. Inf. Sci. Technol.* **61(12)**, 2544–2558.

Thelwall, M.; Wilkinson, D.; Uppal, S. (2010b). Data mining emotion in social network communication: Gender differences in MySpace. *Journal of the American Society for Information Science and Technology* **61(1)**, 190–199.

Theunis, M.; Küster, D.; Tsankova, E.; Kappas, A. (2010). CyberEmotions: Online discussion forums elicit subjective emotional response. Poster accepted for presentation at the 3rd European Conference on Emotion, organized by the Consortium of European Research on Emotion, Villeneuve d'Ascq, France. Not presented due to air traffic issues (volcanic ash cloud).

Van Mieghem, P. (2011). Human Psychology of Common Appraisal: The Reddit Score. *IEEE Transactions on Multimedia* **13(6)**, 1404–1406.

Van Mieghem, P.; Blenn, N.; Doerr, C. (2011). Lognormal distribution in the digg online social network. *The European Physical Journal B* **83(2)**, 251–261.

Van Swol, L. M. (2009). Extreme members and group polarization. *Social Influence* **4(3)**, 185–199.

Varol, O.; Ferrara, E.; Ogan, C. L.; Menczer, F.; Flammini, A. (2014). Evolution of online user behavior during a social upheaval. In: *Proceedings of the 2014 ACM conference on Web science*. ACM, pp. 81–90.

Walker, J.; Ante, S. E. (2012). Once a Social Media Star, Digg Sells for $500,000. In: *The Wall Street Journal*.

Wang, C.; Huberman, B. A. (2012). Long trend dynamics in social media. *EPJ Data Science* **1(1)**, 1.

Wang, N.; Kosinski, M.; Stillwell, D.; Rust, J. (2014). Can well-being be measured using Facebook status updates? Validation of Facebook's Gross National Happiness Index. *Social Indicators Research* **115(1)**, 483–491.

Warriner, A.; Kuperman, V.; Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* **45(4)**, 1191–1207.

Waugh, A. S.; Pei, L.; Fowler, J. H.; Mucha, P. J.; Porter, M. A. (2009). Party polarization in congress: A network science approach. *arXiv preprint arXiv:0907.3509* .

Webb, E. J.; Campbell, D. T.; Schwartz, R. D.; Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*, vol. 111. Rand McNally Chicago.

Weninger, T.; Johnston, T. J.; Glenski, M. (2015). Random Voting Effects in Social-Digital Spaces: A Case Study of Reddit Post Submissions. In: *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, pp. 293–297.

West, R.; Paskov, H.; Leskovec, J.; Potts, C. (2014). Exploiting Social Network Structure for Person-to-Person Sentiment Analysis. *Transactions of the Association for Computational Linguistics* **2**, 297–310.

Wikipedia (2015). Introduction to Sociology/Collective Behavior — Wikipedia, The Free Encyclopedia. `https://en.wikibooks.org/wiki/Introduction_to_Sociology/Collective_Behavior`. Accessed: 2016-11-08.

Wikipedia (2016). Twitter Revolution. `https://en.wikipedia.org/wiki/Twitter_Revolution`. Accessed: 2016-11-08.

Wilkinson, D. M.; Huberman, B. A. (2007). Assessing the Value of Coooperation in Wikipedia. *CoRR* **abs/cs/0702140**.

Wilson, G. (2015). What Colors Are This Dress? Kim Kardashian, Miley Cyrus, Justin Bieber And A Bajillion Other Celebs Weigh In. `http://www.mtv.com/news/2091624/dress-blue-black-white-gold-kim-kardashian-miley-cyrus-justin-bieber/`. Accessed: 2016-11-08.

Wu, F.; Huberman, B. A. (2007). Novelty and collective attention. *Proceedings of the National Academy of Sciences* **104(45)**, 17599–17601.

Wu, F.; Huberman, B. A. (2008). Public discourse in the web does not exhibit group polarization. *Available at SSRN 1052321* .

Wu, Y.; Zhou, C.; Xiao, J.; Kurths, J.; Schellnhuber, H. J. (2010). Evidence for a bimodal distribution in human communication. *Proceedings of the national academy of sciences* **107(44)**, 18803–18808.

Yasseri, T.; Hale, S. A.; Margetts, H. (2013). Modeling the rise in internet-based petitions. *arXiv preprint arXiv:1308.0239* .

Youyou, W.; Kosinski, M.; Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* **112(4)**, 1036–1040.

Zajonc, R. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist* **35(2)**, 151–175.

Zaller, J. (1992). *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

Zech, E.; Rimé, B.; Nils, F. (2004). Social sharing of emotion, emotional recovery, and interpersonal aspects. In: *The Regulation of Emotion*, Taylor & Francis. pp. 159–188.

Zillmann, D. (1971). Excitation transfer in communication-mediated aggressive behavior. *Journal of experimental social psychology* **7(4)**, 419–434.