The Cost of Sharing Information in a Social World

Arthi Ramachandran

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

The Cost of Sharing Information in a Social World

Arthi Ramachandran

With the increasing prevalence of large scale online social networks, the field has evolved from studying small scale networks and interactions to massive ones that encompass huge fractions of the world's population. While many methods focus on techniques at scale applied to a single domain, methods that apply techniques across multiple domains are becoming increasingly important. These methods rely on understanding the complex relationships in the data. In the context of social networks, the big data available allows us to better model and analyze the flow of information within the network.

The first part of this thesis discusses methods to more effectively learn and predict in a social network by leveraging information across multiple domains and types of data. We document a method to identify users from their access to content in a network and their click behavior. Even on a macro level, click behavior is often hard to obtain. We describe a technique to predict click behavior using other public information about the social network.

Communication within a network inevitably has some bias that can be attributed to individual preferences and quality as well as the underlying structure of the network. The second part of the thesis characterizes the structural bias in a network by modeling the underlying information flow as a commodity of trade.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am greatly indebted to my advisor Augustin Chaintreau, for his constant support and guidance throughout my PhD. He has been an amazing researcher to learn from – his enthusiasm, his optimism, his ability to look at the big picture, and way of capturing the energy and interest of the audience have been truly inspiring. He has been unwavering in his encouragement and support, especially during the lean patches. I could not have hoped for anyone better as my advisor. He has been my biggest cheerleader and as a result, given me confidence in my own abilities as a researcher. Thanks for taking a chance on me.

I would like to thank the members on my PhD committee: Augustin Chaintreau, Vishal Misra, Krishna Gummadi, Kathleen McKeown and Chris Wiggins for their time, and their useful comments and suggestions. I also thank the fabulous staff at Columbia, especially Jessica Rosa, Cindy Walters and Elias Tesfaye for all their administrative help. I gratefully acknowledge my funding through the Computer Science Chair's Distinguished Fellowship, Microsoft's Graduate Women's Research Program, the National Science Foundation's Graduate Research Fellowship and Augustin Chaintreau's NSF grant CNS-1254035.

I have been very fortunate in having some terrific researchers as my collaborators including Arnaud Legout, Maksym Gabielkov, Patrick Loiseau, Michela Chessa, Gilad Lotan, and Lucy Wang. I am very grateful for all the fun and learning I have had during these interactions. I'd like to thank Rocco Servedio for his valuable advice during some of my more challenging times. Thanks also to Julia Hirschberg for her advice and support, through WiCS as well as in navigating the PhD. I'd also like to thank Adam Cannon for taking me on to run the ESP program.

My years at Columbia have been an exciting journey, and for this I have my friends to thank. Many thanks to my friends and colleagues in the social networks group including Christopher Riederer, Mathias Lecuyer, Max Tucker and Ana-Andreea Stoica. My WiCS cohort, Rebecca, Melanie, Erica, Sara, Anna, and many others, have been a source of friendship and support throughout the degree. My thanks to Daniel who put up with me as a roommate for many years and was always willing to listen to my complaints and give advice. I owe Josh many thanks and chocolates for always giving me a place to crash during the years of remote-working and commuting. I'm grateful to my undergrad friends, Amy, Dan, Irene, and Megan among others, who've been with me on the journey from undergrad days to joining the working world and then going right back to school. Thanks also to my family away from home, Aditya, Aishwarya, Bharat, Divya, Harini, Kalpana, PJ, Ravi, and Shankar, who were always up for a movie, frisbee and, of course, lots of cooking and good food.

I am very thankful to my family, my brother, brother-in-law ,and sister-in-law for their love and encouragement. And of course, to my pattu nephew for bringing me so much joy.

I am deeply indebted to my parents, without whose encouragement, love and hard work, this thesis would not have been possible.

Last but not least, thank you to my husband for his unwavering love and support throughout this long roller-coaster journey of graduate school. Aravind, I never would have made it without you.

Dedicated to my family

Introduction

Social networks research studies the interrelationships between different agents in a system and understand how these interactions can lead to complex phenomena. Such networks include friendship networks (*e.g.,* Facebook), communication networks (*e.g.,* arXiv collaboration networks), communication networks (*e.g.,* Twitter), or even disease transmission networks. Over the last couple of decades, the field has evolved from studying small scale networks and interactions to massive ones that encompass huge fractions of the world's population. Much of this expansion in scale has been made possible by the increased prevalence of online social networks such as Facebook and Twitter among others. The resultant information exchanges and networks have transformed the research questions from primarily sociological in nature to predominantly computational.

With all the different types of networks in use, there are many forms of communication that arise. For example, a collaboration network might be sparser with less frequent communications than an email network. Or the importance of a link in a disease transmission network comes into effect with the first contact while for an online social network, continued contact is more crucial. In spite of these differences between networks, the questions researchers ask in each of these domains are fundamentally very similar – How do information move across the network? What are the drivers as well as consequences of various network structures? What is the impact of behavioral changes?

0.1 Directions in Social Networks

We now give an overview of some of the major directions of research in Social Networks and describe some of the techniques used to address these questions.

How is information shared in social networks? A core function of many networks is to convey information in some fashion. Research questions in this area involve modeling the flow of information through a network. Early research in online social networks focused on measuring and understanding the resultant information diffusion structures (called 'cascades'). The vast majority of such cascades tend to be shallow and narrow *i.e.,* they reach only a few individuals with a minimal number of hops from the starting point (Figure 0.2). This holds true across several contexts including news or other media content on platforms such as Twitter [Goel, Watts, and Goldstein 2012; Goel et al. 2016] and recommendation networks in viral marketing [Leskovec, Singh, and Kleinberg 2006; Leskovec, Adamic, and Huberman 2007]. On the other hand, a small fraction of the cascades become quite large (in either breadth or depth) [Dow, Adamic, and Friggeri 2013; Adamic, Lento, and Fiore 2012]. Such cascades can can gain size through very different means (for instance, relying on a single major source vs. smaller organic diffusion). In terms of shape of cascades, while there are many different structures that emerge, cascades largely fall into a few broad shape classes [Leskovec et al. 2007] (Figure 0.1).

Can we predict influence? This understanding of diffusion structures leads us to another major question studied in information flow - can we *a priori* predict the influence or the spread of a piece of media or information? One key observation in the literature is the importance of the use of time-based features (*e.g.,* the initial cascade details or the change in features over several units of time) in the prediction of cascade size. This time component can be differently accounted for by including a Bayesian approach [Zaman, Fox, and Bradlow 2014], reformulating the problem into several stages [Cheng et al. 2014], and

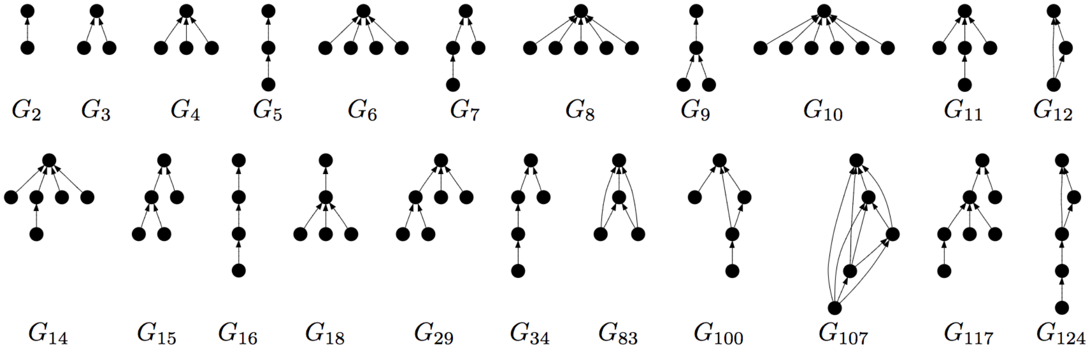

Figure 0.1: Graphical representation of common cascade shapes (order by frequency) [Leskovec et al. 2007]

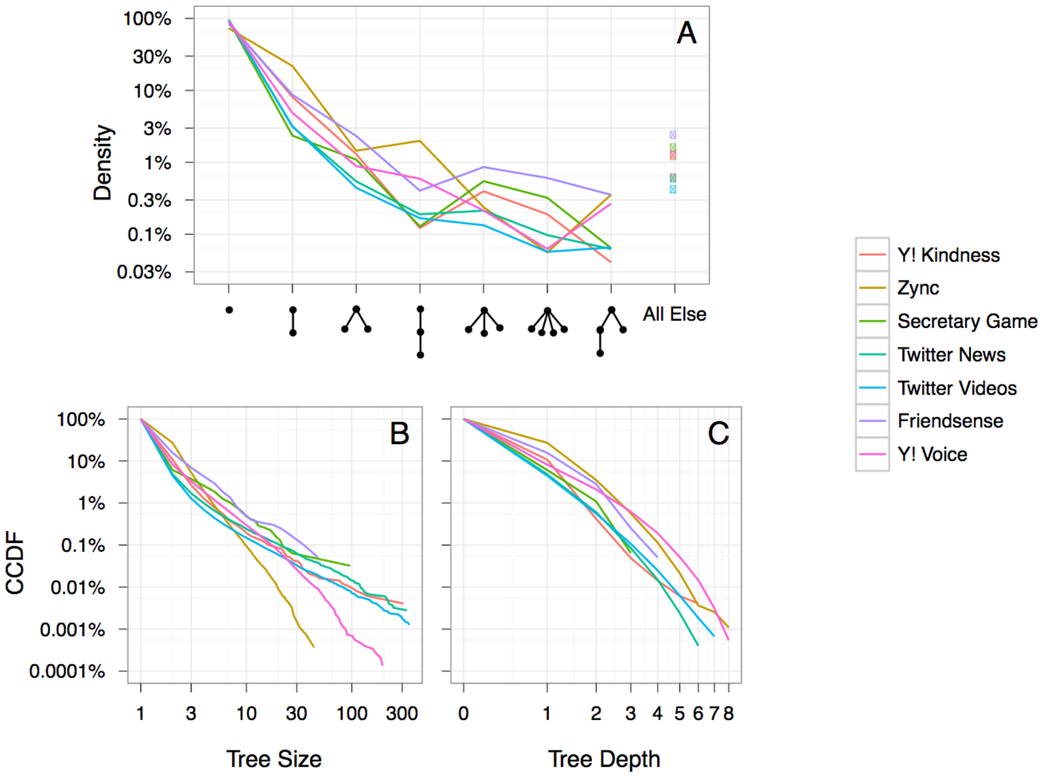

Figure 0.2: Distribution of cascade structures of different social networks for (A) cascade shapes (B) total cascade size (C) cascade depth [Goel, Watts, and Goldstein 2012]

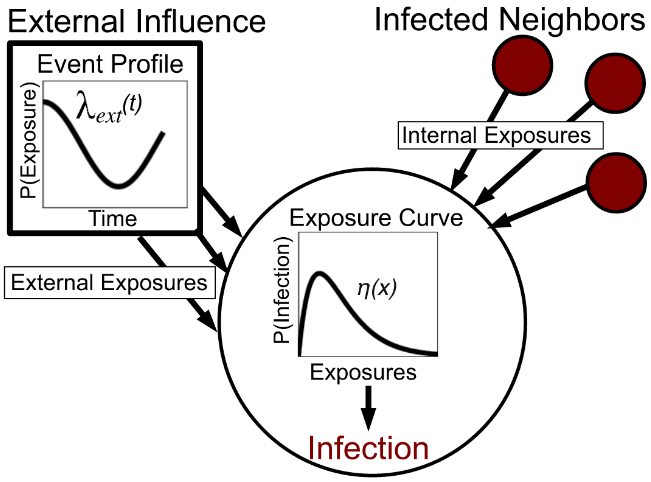

Figure 0.3: Model for external influence in information diffusion in a social network [Myers, Zhu, and Leskovec 2012]

classifying cascades based on their temporal evolution [Yang and Leskovec 2011].

Other works build models of the underlying diffusion to then predict the final outcome. In one such model, a shared piece of information is seen as a Poisson ‘arrival’ which accumulates attention but simultaneously ages [Shen et al. 2014]. In another modeling technique, each additional share contributes to the future probability of sharing [Zhao et al. 2015]. External influences can also be accounted for separately either as an additional point of attention [Rizoiu et al. 2017] or an additional exposure curve that is essentially added to the internal influence (Figure 0.3) [Myers, Zhu, and Leskovec 2012].

A further question is whether we can quantify the influence of specific. Identification of influential nodes can be based on their position in the network and can use centrality measures [Rusinowska et al. 2011] or estimated through optimization techniques [Kempe, Kleinberg, and Tardos 2003; Li et al. 2013]. Such methods have application in a broad array of fields such as transportation, urban networks, or disease propagation networks. However, a few nodes alone are not responsible for shaping opinion in a network.

How is opinion shaped? Communication and information on networks also fulfills a specific objective on the part of the content creator - to influence and convince the readers. This can take the shape of advertisers trying to market their product or celebrities marketing their brand, or news articles trying to shape the opinion of the public. We see that many individuals working together can lead to optimal outcomes, even if individually, they are suboptimal [Degroot 1974; DeMarzo, Vayanos, and Zwiebel 2003]. However, if certain nodes are more influential or refuse to change their opinion, suboptimal situations (such as disparate groups emerging) occur [Ghaderi and Srikant 2013; Bindel, Kleinberg, and Oren 2011]. This series of works use simplified theoretical models to capture macroscopic outcomes in the network.

What are the consequences on privacy? A potential consequence of information exchange for an individual can be a loss of privacy. Studies show that sparse data results in individuals being increasingly easily identified. Mobility information, for instance, needs approximately four locations to uniquely identify an individual [Montjoye et al. 2013]. Techniques exploiting several information sources can further erode privacy [Narayanan and Shmatikov 2008; Narayanan and Shmatikov 2009]. In addition to identity, other information about users of online services can be inferred from characteristic attributes [Sharma et al. 2012; Mislove et al. 2010] to link creation times [Meeder et al. 2011].

0.2 Emerging Themes in Social Networks Research

Social networks datasets are massive and have complex relationships, both between participants and between the different types of data. While this complexity can give rise to new challenges in analysis, it also results in new opportunities. We outline some emerging themes that take advantage of these inter-relationships.

0.2.1 Cross-Domain Analysis

These relationships act as connectors between different types of information, allowing us to leverage data from different sources. Methods that exploit this feature often also rely on data sparsity in general. A well-known example of this style technique is the

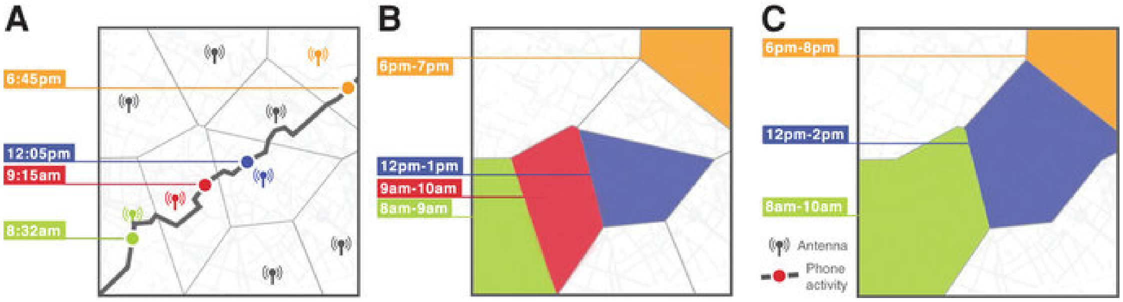

Figure 0.4: A anonymized mobile phone user by (A) location trace (B) as recorded by database (C) at lowered (and more anonymous) resolution [Montjoye et al. 2013]

de-anonymization of the Netflix Prize dataset where the (anonymized) dataset of movies and preferences were combined with user profiles and ratings from the Internet Movie DataBase to identify a large fraction of the anonymous Netflix users [Narayanan and Shmatikov 2008]. Another way to combine multiple information sources is to combine two different networks and use structural information to transfer identities from one to another[Narayanan and Shmatikov 2009; Narayanan, Shi, and Rubinstein 2011; Sharad and Danezis 2014; Pedarsani and Grossglauser 2011].

Even within the same broader dataset, different types of information can be used to great effect. For instance, spatial and temporal user information together were significantly more informative than either alone (Figure 0.4) [Montjoye et al. 2013]. The overarching question of this style of work is how to best utilize available sources of information to (a) learn a new attribute about individuals in the network (b) understand its limits to design sound mechanisms for data disclosure.

0.2.2 Network Structure and Information Flow

The complexity in interactions between participants in a network both influence and are influenced by the intertwined and evolving network structure. A key question that current research struggles with is modeling and predicting the flow of information in such a system, as well as the biases that result.

The network structure strongly affects how information is diffused online. For instance, news typically reaches a large audience not directly but through intermediaries [Cha et al. 2012; Wu et al. 2011]. A side benefit of this type of diffusion is broadening the range of opinions seen by a user [An et al. 2011]. In the context of social learning, the network structure impacts the equilibria with stochastic networks resulting in favorable asymptotic results [Bala and Goyal 1998; Acemoglu et al. 2011]. Further, the homogeneity of the user preferences facilitates learning. The existence of more 'forceful' agents in the network can result in the network as a whole not converging to a single value [Golub and Jackson 2010]. Structural features of the network (such as groups balanced in the information they provide vs the information they obtain) can prevent such outcomes [Acemoglu, Ozdaglar, and ParandehGheibi 2010].

Studying how user choice motivates link creation and network evolution gives us insight into the other side of this dynamic. In the setting where players may create new links at a fixed cost, these dynamics typically lead to extreme specialization [Galeotti and Goyal 2010]. Studies of large scale networks indicate that that individuals add edges in order to complete triangles in the network [Leskovec et al. 2008]. Overall, macroscopic properties of the network tend to be stable but locally, the networks are tend to be more unstable [Watts 2006; Kumar, Novak, and Tomkins 2010].

0.3 The *Cost* to an Individual in a Social Network

Thus, we see that the *cost* to an individual in a social network can take on many flavors: loss of privacy when they participate in online communication, decreased access to information due to suboptimal network structure, or reduced sphere of influence from inferior network position. In this thesis, we explore and develop techniques to exploit some of these features and further examine how they arise.

We tackle two of the emerging themes of study (1) How can we leverage information exchanged across multiple domains to more effectively learn about the various domains? and (2) How does the structure of the network affect the exchange of information? Note that to address these themes, we employ a diversity of methodology inspired by previous studies in social networks – quantifying and predicting features of the diffusion, reproducing certain aspects of the network, and modeling structure and communication within the network.

In the course of the next several chapters, we show that:

- Even passive users of a social network can be deanonymized with at most 10 of their clicks.
- User attention can be accurately predicted, even a day in advance, using noisy data.
- Producers of original content are a very specialized subset of the network. The properties of specialization (inequality in production) depend on spectral properties of the network.
- Further, when the network is growing, this inequality of labor is exacerbated and inequality can lead to persistent inequality.

0.3.1 Inference by Leveraging Cross-Domain Analyses

The first theme we examine is developing better inference techniques by leveraging knowledge gained from combining several types of data – a technique we refer to as 'cross domain analysis'. This style of analysis bridges the gap between multiple types of information and allows us to infer features of the users or data. Earlier studies of social networks often independently studied and measured the various facets of the networks. Only recently have researchers started taking advantage of the rich relationships between these independently collected datasets through cross domain analysis techniques. This type of analysis has become more popular with the increasing availability of large inter-connected datasets. As an active area of research, it has found use in applications such as privacy and user identification.

Privacy and Deanonymization

Social network users tend to feel secure about their privacy through the use of pseudonyms and semi-anonymous user names. In order to ascertain whether there is a false sense of security, we focus on a simple yet central problem: Can an independent first or third party (respectively hosting content or serving ads) recognize a visitor as the owner of a profile in social media? Leveraging cross-domain information from public twitter posts and relationships and click data seen only by the content host, we show that with relatively little information, even a passive twitter user is identifiable. We find that the content that a user receives on social media is highly distinguishable (in spite of the fact that most traffic is only for a few popular articles). Leveraging this observation, we develop an original identification method which identifies users with a median of at most 10 clicks.

In this Chapter1, we address the following broad questions, with the goal of harnessing publicly available information to learn private user-level information:

- How unique is a user's content? (Section 1.1)
- How can we use a user's uniqueness to deanonymize a user from their click behavior? (Section 1.2)

Predicting User Attention

Social media attention is poorly measured – the raw information is often not available publicly and few content publishers and social networks are willing to release such data. Clicking behavior itself is information that is often private and hard to obtain. Usually, only the website owner and the network itself has access to that information. In such cases, when there is scarcity of data, we study how inference methods can applied to social networks to sufficiently provide a substitute for user behavior.

We rely on a dataset from a content publisher which contains both private information of several thousand posts and publicly available information from twitter on the conversation relating to those posts. We focus on the evolution of click generation over time. A key observation from these empirical analyses is that the process of click creation is a two stage mechanism: from posting to impression (a view of a post) and impressions to clicks. Further, we find that the factors affecting click dynamics time since posting, the posting account and related graph structure, and the content quality behave almost independently of each other. Equipped with this insight, we develop a two-stage model which accurately predicted temporal clicks.

In Chapter 2, we study the following broad questions, with the goal of leverage publicly available information to better estimate private content-level information:

- How does the publicly known information about shares and clicks relate to private information available to a content publisher? How do these metrics evolve over the life-cycle of a social media post? (Section 2.1)
- Can we leverage our observations to predict private information, such as the clicks a link obtains? (Section 2.2)

0.3.2 Interaction in Information Flow and Perpetuation of Bias

Our second theme revolves around how information flows in a social network and the biases that result. One of their most dramatic consequences of social networks is the deluge of information we consult before making any decisions. Natural questions that arise in this context of news dissemination are "How is information introduced and propagated through the social network and what are the resultant biases?" and "What types of networks growth and evolution guarantees everyone to eventually benefit from information sharing?" At large scales of data, previously hidden dynamics begin to emerge as observable phenomenon that can be studied using data analysis and statistical techniques. In social network services, such as Twitter and Facebook, the content that is produced and exchanged behaves as a commodity of trade. As with any commodity, content acquisition has both an associated cost and value.

Factors affecting the valuation of content by an individual include quality, relevance to interests, speed of receiving the content, and how it relates to their neighbors. In a connected society with a sharing economy, each agent behaves so as to maximize their reward to effort ratio. These agents' choices are also affected by the choices of their neighbors in the network. There is not a strict assignment of effort and reward but rather, this effort is distributed across the network. We show that this redistribution has non-negligible effects. Hence, the behavior of the complex network is the result of the actions (algorithms) that each of these individuals follow.

Economic Models for Information Flow

While it is well-known that only a minority of participants in a network are active, we explore the setting of the introduction of *original content* into a network. Using data from multiple online sources on Twitter, we show that this addition of original content remains extremely concentrated. In fact, counter-intuitively, original content production is skewed towards less active and connected people. Since the availability of news worth reading in a social network exhibits the property of a public good, we develop a model that extends public good theory from economics which correlates with the empirically observed activity. In this model, We build a model of information sharing where reward of reading and searching for information is socialized. We show that specialization in content production emerges even when players are ex ante identical, and this occurrence is related to spectral properties of the social network and expiration time of the content when content is long-lived, specialization is unavoidable!

In Chapter 3, we address the following broad questions, with the goal of understanding how social networks function as a source of information:

- Who are the producers of original content in a social network? (Section 3.1)
- Can we develop a model to reproduce our observations? What are the implications of such a model? (Section 3.2)

Biases and Network Evolution

Building on this work, we focus on the dynamics of specialization in an evolving graph with network interactions. The empirical results from simulations on a growing network highlight the complex nature of information sharing. When the network expands and hence more information get shared, a majority of nodes suffer on the short term, seeing diminished accuracy and more individual effort. The benefits from information sharing are skewed towards who appeared earlier (hence more connected); new arrivals and other nodes with smaller degrees benefit much less. The theoretical analysis further establishes the connection between large unbalanced communities and the failure of information sharing to benefit everyone. While, in the worst case, this can result in degenerate equilibria, we show that small deviations from the worst case are enough to allow all players to contribute and gain from the network.

In Chapter 4, we address the following broad questions, with the goal of understanding how evolving networks change the dynamics of information sharing:

- How do users in an evolving graph behave when seeking to preserve privacy in the information they share? Do users gain from greater access to information through their network of friends? (Section 4.2)
- Can we replicate our observations with models? (Section 4.3)

Definitions

Social Network An online platform where individuals can share text, media or links and build relationships with other users in the platform. The underlying network can be directed or undirected depending on the platform used. E.g. Twitter, Facebook, Instagram

Share A link or url that has been posted (or reposted) by a user of social media. This is used interchangeably with ‘post’ and ‘tweet’.

Content The substance that is shared by an individual on social networks. In the context of this thesis, we focus on content that pertains to articles shared *i.e.,* the text of the content shared should contain a URL linking to a longer form article.

Follower An edge in a social media graph. In an undirected social network (such as Facebook), both users confirm the relationship and receive content that they share. In a directed social network (such as Twitter or Instagram), the follower chooses to receive all the content shared by the individual they are following. This is used interchangeably with ‘relationship’ and ‘edge’.

Impression An impression occurs each time content is fetched from its source, and is thus countable. This is typically data that is private and available only to the publisher of the content and the network itself. This is a standard metric used to measure popularity of a particular post.

Reception A potential audience member of a post. These are users who will potentially view a post. Formally $receptions = \sum_{i \in \text{users sharing a post}}$ number of followers of i

Click The selection of a link on a post. This is a conscious action online and can be used as a measure of interest in the post or topic.

Producer An individual sharing content on social media. Such individuals are responsible for producing the content consumed by others on the network.

Conversation The exchange of information among many individuals online. In a medium such as Twitter, the conversation comprises of the tweets and the retweets that follow.

Data

In this dissertation, I use the following datasets for the analyses:

NYTIMES The NYTIMES dataset contains all the Twitter posts containing a URL from the nytimes.com domain during a full week of December 2011 [May et al. 2014]. In parallel, we crawled the follower-followee relationship at the same time in order to construct the URLs that each user re- ceived. The final dataset totals 346k unique users receiving a total 22m tweets with URL (including multiplicity). Of these, there are 70k unique links. All the data in this dataset is publicly available. This dataset is used for study in chapters 1, 3, and 4.

KAIST The KAIST dataset contains the entire Twitter graph from August 2009 and consists of 8m users and 700m links [Cha et al. 2012] . Taken over the course of a month, the dataset contains 183m tweets. Of these tweets, we considered only those with urls (37m) since those are the tweets that provide an indication of sharing media on twitter. All the data in this dataset is publicly available. This dataset is used for study in chapter 1.

KAIST-NEWS In order to better study news articles, we filtered the tweets in KAIST by news domains (*e.g.,* , `cnn.com`). The classification of a domain as news was obtained from the Open Directory Project (`http://www.dmoz.org/`), a volunteer edited directly of Web links. Each link in the directory is annotated with a top level categories and multiple levels of subcategories. In our analysis, we only took into account the top level category.

We kept all the domains with a reasonable number of posts (> 2000 posts) resulting in 31 domains. We removed domains which did not seem to follow the same definition of news as others (aggregators such as e.g. `news.google.com` and `reddit.com`, weather services such as `weather.gov`, and region specific domains such as `thehindu.com`). All the data in this dataset is publicly available. This dataset is used for study in chapter 3.

DIGG Within KAIST, we focused on the domain digg.com (DIGG). This dataset consists of 216k unique URLs tweeted by 44k unique users. These users represent the population that displays some interest in the domain. Hence, the as- sociated network is derived from the 52m links connecting these individuals. All the data in this dataset is publicly available. This dataset is used for study in chapter 1.

BUZZFEED+ This dataset consists of data combined from three sources. This dataset is used for study in chapter 2.

Publisher Dataset BuzzFeed is an internet media company focussing on creation and distribution of content. They cover a wide range of topics across multiple platforms. On Twitter, they have ~ 40 active posting accounts, each targeted to a different demographic. As a publisher on Twitter, they have access to Twitter Analytics of their account, including the hourly metrics of link clicks, retweets, and impressions. These metrics gives us more granular data of the readership of an account. We leveraged these hourly analytics for BuzzFeed Twitter accounts which include links to `www.buzzfeed.com` content. We focused on original tweets, excluding retweets, in order to preserve uniformity of content source.

Our dataset includes all original tweets published by any of BuzzFeed's Twitter accounts over a 7-day period from August 3 to 17, 2016 (4K tweets). The largest account is BuzzFeed's primary, eponymous account, which has 2.8M followers. This account posts a wide variety of links to BuzzFeed web articles, typically those projected to become most viral. The next most popular account is BuzzFeedNews, with 470K followers, which posts links to traditional news stories published by BuzzFeed. The remaining accounts serve a more specific niche or content genre, and are named accordingly (e.g. BuzzFeedSports, BuzzFeedFashion).

While some of the data from this dataset is accessible by the public (the tweets posted by BuzzFeed), the Twitter Analytics data is private and is only seen by BuzzFeed and Twitter.

Retweet Dataset We used Twitter's REST API to scrape all tweets published by BuzzFeed accounts and all related public retweets over the span of the same 7 days, forming a complete public dataset of the BuzzFeed tweets. Each (re)tweet also provides the publicly available follower count of the (re)tweeting user. This part of the dataset is publicly accessible.

Clicks Dataset To further supplement our public readership data, we used bit.ly's API to gather all twitter-originating link clicks (those with `twitter.com` or `t.co` as the listed referrer domain). We collected data for the hour periods of hours $1, 2, 3, 4, 5 - 8, 9 - 12, 13 - 24$. Of the 4K original tweets published during that 7-day range (Aug 3 - 17, 2016), we considered only those that contained bit.ly URLs, and completed a 24-hour lifecycle within the date range (1.4K tweets). Note that most BuzzFeed accounts almost exclusively used one method of sharing information (either using bit.ly for almost all the links they post via the official bzfd.it shortener, or not at all). As a result, focusing on bit.ly links introduces a source-bias since accounts behave differently. However, we believe there is little intrinsic bias introduced by bit.ly itself. This part of the dataset is publicly accessible.

Broad News Domains Dataset We include comparisons of overall dynamics to other news sources in order to validate our use of `buzzfeed.com`. This dataset contains the hourly posts, receptions and `bit.ly` clicks for five major news sources, from both Twitter and bit.ly [Gabielkov et al. 2016].

Chapter 1

Social Media + Clicks = Identity

Our first study examines the potential of public information to break pseudo-anonymous identity. Pseudo-anonymous identities are common in social networks where a user might not use their real name or information but they will use consistent profiles for their social media presence. The use of such pseudo-anonymous identities on social networks gives users a sense of security about their online presence. Increasingly, there has been development of methods to break this anonymity. Often these techniques rely on some universal connecting information and exploit sparsity in data in general [Narayanan and Shmatikov 2008; Narayanan and Shmatikov 2009; Montjoye et al. 2013] to identify or infer information about users of online services, and social media in particular [Sharma et al. 2012; Meeder et al. 2011; Mislove et al. 2010].

Most of the work on social media centers on a user's *explicit* activity with regard to one or several social network providers, and occasionally on how this leaks information between or beyond them. In contrast, *implicit* activities such as clicks and reads are under-explored. They are typically much harder to study: only providers of social networks have access to individual data about them, and they rarely reveal it for privacy and commercial reasons. Studying *implicit* activities requires bridging two worlds: Content producers maintain a detailed user profile for personalization and ads using cookies, but *a priori* have no information about the user outside their domain. Social media usually have a wider view of someone's interest, but may lack detailed information about a user in a domain.

We focus on a simple yet central problem: "Can an independent first or third party (respectively hosting content or serving ads) recognizes a visitor as the owner of a profile in social media?"

In addition to research on deanonymization, our results complement previous studies of cascades and information diffusion in social media [Cha et al. 2010; Kwak et al. 2010; An et al. 2011; Rodrigues et al. 2011; Cha et al. 2012; Wu et al. 2011; Goel, Watts, and Goldstein 2012]. Indeed, measuring and predicting the success of a cascade is still a matter of controversy [Cha et al. 2010; Bakshy et al. 2011]. Validating those studies with individual data about *which* users clicked *which* links sent by *whom* requires data unavailable outside researchers at social media provider [Bosagh Zadeh et al. 2013; Bakshy et al. 2012]. This study opens an alternative way, by inferring visitors from web traffic, to study the real success of social media in generating clicks.

What sets us apart from previous work is that our work exploits basic ingredients, common to any web-domain. Hence, our results apply more generally: Whenever a user follows information from a *public* social media such as Twitter she is instantly recognizable by the website she visits unless she (1) has not visited this domain more than 4 times, (2) takes action not to be appearing as the same visitor, or (3) creates multiple identities, makes her list of connections private, or delays her visits by a non-negligible time. While each of these actions or situations are deemed possible, they significantly limit a user's web experience. In contrast with previous work, we assume *no cooperation* of any sort. We assume simply *first party tracking*: the provider can maintain a persistent identity for web visitor *only within its domain*. We assume that the domain knows only one thing: that the click was generated through a social media site (*e.g.,* from `twitter.com`).

In this chapter, we make the following contributions:

- We first unearth a critical fact: Although links shared on social media exhibit extremely skewed popularity distribution with a few receiving the most attention, the content a user receives is highly distinguishable. (§ 1.1)

- We design an original identification method which identifies users with a median of ≤ 10 clicks. This method, however, is limited in its applicability to small domains with very low click rates. To address the most challenging cases, we introduce an extension of the baseline method, using recent work on influence inference. When inference is accurate, the method promises identification in less than 4 or 5 clicks. (§ 1.2)

1.1 Uniqueness of social media users

Previous results reported that four spatio-temporal points are enough to uniquely identify 96% of the individuals in large anonymous mobility datasets[Montjoye et al. 2013]. Similarly, records from the Netflix prize revealed that most of the time the set of items rated by a user overlaps with less than half of those from the *closest* users in these data [Narayanan and Shmatikov 2008]. Similarly, we ask here: "How unique is the set of people you follow and the content you receive from them?"

How unique is the content you receive? Our first and most striking result is that users overwhelmingly receive a *unique* set of URLs. This is in spite that the majority of users receive few of them (*e.g.,* in *NYT* half of them receives less than 15 distinct URLs) and that URLs are concentrated on a few blockbuster links that are essentially received by everyone (*e.g.,* the top-15 URLs account together for 7% of all the tweets).

Figure 1.1 shows, for users who received more than n URLs, what fraction of them have a unique subset (i.e. one that no other user received). Note that, alternatively, when no more than $k \geq 1$ users receive this subset, we say the user is *k-anonymous*, and plot the fraction of such users for $k = 5$ and $k = 10$. We compare three datasets: KAIST, NYT and

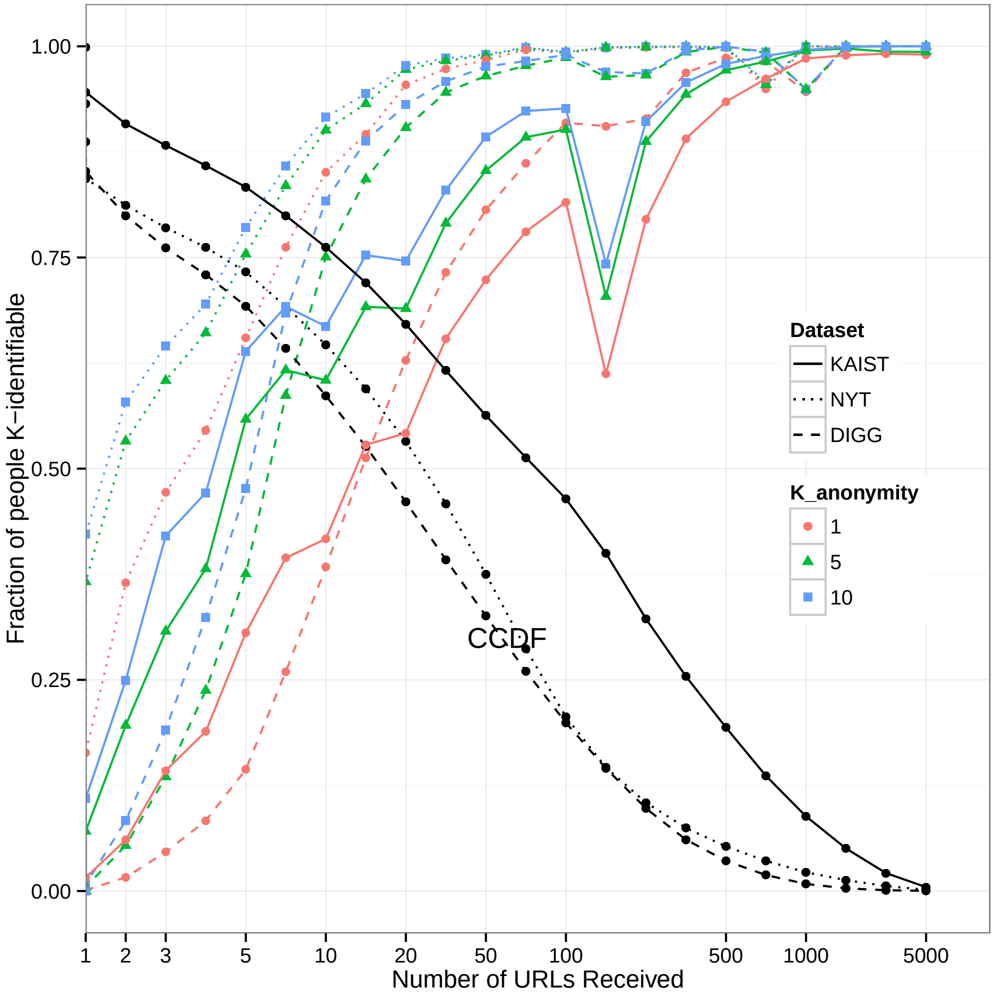

Figure 1.1: Fraction of social media users receiving more than n URLs, and those receiving a unique sets of URLs among them.

DIGG. One observes that 15 URLs appears to be a turning point, below which threshold, a user is rarely distinguished by its set of URLs.

Note that we considered a user uniquely identified only if it is the only one that receives these URLs *or a superset of them*[1]. This property is stronger and makes this result more surprising given that some users received an enormous amount of information from the `nytimes.com` domain (we had more than 10 users receiving above 5,000 URLs each).

The unicity property of your set of URLs is derived from the long-tail property of the

[1] Later this point is critical for identification as it is not in general easy to deduce that a user did *not* receive a URL.

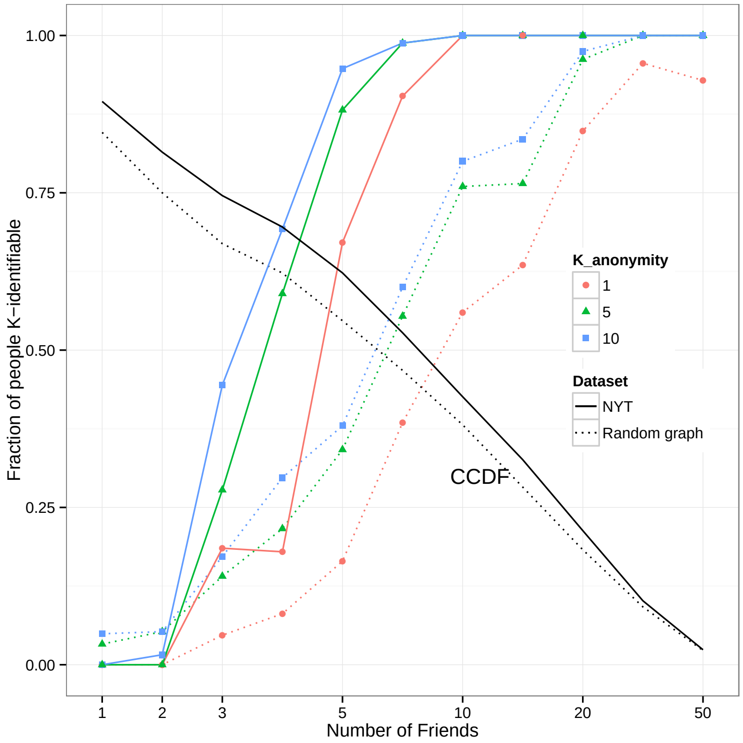

Figure 1.2: Fraction of user with at least n active friends, and those for which this set is unique.

distribution [Goel et al. 2010]. According to this property, a very large fraction of the content you receive is common, while some items will be highly specific. The occurrence of one of these items (which is likely unless your set of received URLs is very small) is sufficient to offer information that makes you distinct.

How unique are the set of people you follow? Given that the links that form your social media news feed are a direct consequence of the person you follow and their posts, it does not come as a surprise that this set of "friends" (as Twitter terminology refers to them) is unique. What is perhaps less obvious is that this set of friends distinguishes you even more than your content.

To measure that effect, we run a similar experiment on the New York Time dataset and find that knowing who contributed to your feed distinguish you with overwhelming probability, even against supersets as discussed before. This translate into almost 70% of the users being unique in that regard; those users amount to 96% of the potential traffic to `nytimes.com`. In effect, knowing a small subset of your posting friends almost always makes you a unique person.

To confirm that this effect is not an artifact of specific structures on Twitter, we run a similar experiment on a surrogate social media, reshuffling at random the edges of bipartite graph, while maintaining the overall friend and popularity distributions (Figure 1.2).

These results highlight a new promising application of sparsity methods to identify social media users, which is based on the content that they receive and the individuals participating in it. We have shown that this has a high potential, as a very large fraction of the users, and overwhelming percentage of the traffic, is created by visitors with unique patterns. It remains to be seen how these facts can come together to constitute a proper and practical re-identification method for the web.

1.2 Deanonymization Algorithm and Select Findings

We have shown that content received on social media is highly distinctive. We now study several methods leveraging this fact, to allow a web domain receiving visits from social media to identify their visitors explicitly. They are multiple motivations for a domain owner to do that: learning additional information about its audience (demographics, interests, trending topics), monitoring the content its users receive from competitors, or even personalize your results (with or without user's knowledge) based on a visitor's inferred social media profile.

Currently, there are three ways to identify your visitors: asking for *users* to sign in using the social network service, asking *social media providers* to reveal a user's identity whenever she clicks on this domain, or asking a *web-aggregator* to perform synchronization. Each is cumbersome as it poses usability concerns (*e.g.,* your visitors may leave or even lie if asked to provide a form of identification) and assumes cooperation (*e.g.,* a social network provider or web-aggregator may not want to reveal their users' identity to a domain, a domain may not want to reveal its audience to them). In contrast with previous work, we assume *no cooperation* of any sort. We aim at an identification scheme that bypasses all mechanisms mentioned above.

1.2.1 The state of web-tracking

Web content publishers or third parties generally keep track of web visitors to inform personalization, targeted advertisement and general audience analytics. There is a clear tendency to aggregate identities for a user across domains, making it ever more difficult to evade tracking [Roesner, Kohno, and Wetherall 2012]. The result is the rise of *web aggregators*, a large class of services who typically observe a user's web itinerary on a large fraction of the web [Gill et al. 2013]. The HTTP protocol allows us to aggregate identities: A single web page (*e.g.,* an article on `nytimes.com`) generates requests to multiples domains (*e.g.,* `admob.com, facebook.com, twitter.com`), each one keeping a *local* user identity. Since browser only blocks `set-cookies` in HTTP response – if they block third party at all – it means the user is recognized unless it was never called before in a first party transactions, a relatively rare case. Ad-networks allow cookies synchronization [Olejnik, Minh-Dung, and Castelluccia 2014], further extending the reach of third party tracking. Evasion techniques exist, such as stronger third-party cookies blocking or reset, but each is met with an alternative forms of tracking: 1x1 pixel image forcing third party request, malicious use of web caching to provide pseudo identifiers in the `Etag` field, Flash cookies [Ayenson et al. 2011], or use of Browser Fingerprinting and Javascript [Eckersley 2010].

Our assumptions, discussed immediately below, leverages common facts on tracking with *no third party cooperation*.

1.2.2 Problem Formulation & Assumptions

We assume simply *first party tracking*: the provider can maintain a persistent identity for web visitor *only within its domain*. This applies to content publishers (*e.g.,* `nytimes.com`).

This also captures aggregators who do not have such limitations (*e.g.,* admob.com, bluekai.com) by extending the domain considered to all those they can track.

In both cases, we assume that the domain knows only one thing: that the click was generated through a social media site (*e.g.,* from twitter.com). We note that this is commonly done today (*e.g.,* nytimes.com limits non-subscribers to 10 articles a month, but allow unlimited access from social media) in multiple ways: most commonly the HTTP referrer field[2], or by providing specific URLs to use in various social media. We do *not* assume, however, that the URL itself is indicative of which posts, tweets, or mention generated that click. It is sometimes possible to leverage that fact and make our method even more efficient, but we ignore it here. Note, however, that we assume that all clicks a user generates on the social media (*e.g.,* from twitter.com) comes from her feed and not search or special content promotion. We believe this would affect our results in the same way as attribution errors (see § 1.2 for more on that topic).

We assume that, in parallel, the domain also monitors *who* post links to its content in the social media. It seems legitimate as "active" users posting links expect it to be publicly known. Again, it's commonly done especially to promote a domain by retweeting users and celebrities mentioning its content or to follow what is said about the domain. Our last assumption is that the domain owner is able to access the graph of followers of each "active" users mentioning URLs from their domain. To simplify we first assume that this information is prior knowledge, but later on we discuss how to limit how much of that information is needed.

These assumptions are inspired from information made publicly available by Twitter. Our methods would extend to other social media with similar policies.

[2] For aggregators, we assume that the content publisher relay that information.

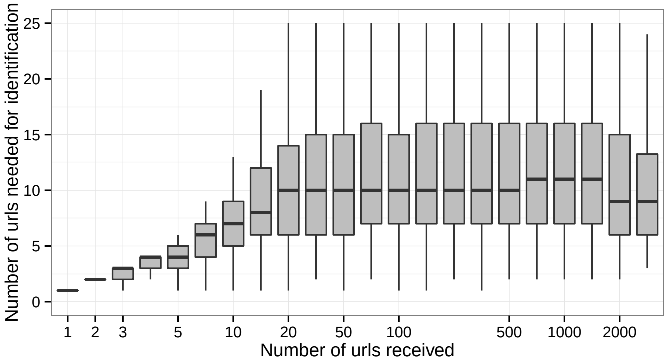

Figure 1.3: Distribution of the number of visits needed to successfully re-identify a node as a function of the size of its receiving URLs set.

1.2.3 A simple method for re-identification

Given what we have learned about content received on social media, the following scheme is promising: In the first phase, for every URL in the domain, collect the set of people who received it in the social media (*i.e.,* the union of followers of "active" users who post it). In the second phase, for each visitor, collect URLs of all her HTTP requests generated from this social media, and intersect the URLs' received sets. This method can safely conclude the identity of this visitor when this intersection contains a single node.

Our preliminary analysis suggests this method terminates, as each user often is the unique node intersecting all the URLs she receives. But this raised two questions: How many URLs from each user are needed to reach this conclusion? As a consequence, how likely is this method to complete when only a subset of the content a user receive generates a click to that domain?

Figure 1.3 presents the results of a simulation where for each user in our dataset we look at URLs included one by one in random order and stop whenever the intersecting set is a singleton. Across the whole population the median user is identified after 8 URLs, and even for large sets 10 URLs suffice on average.

In real life, however, an intersection step occurs when a user decides to clicks, and only a fraction of URLs received generate a HTTP request. In addition, some URLs may receive more clicks as they are simply more interesting. To account for that, we built the following click generation model: Many links were published using the URL shortener

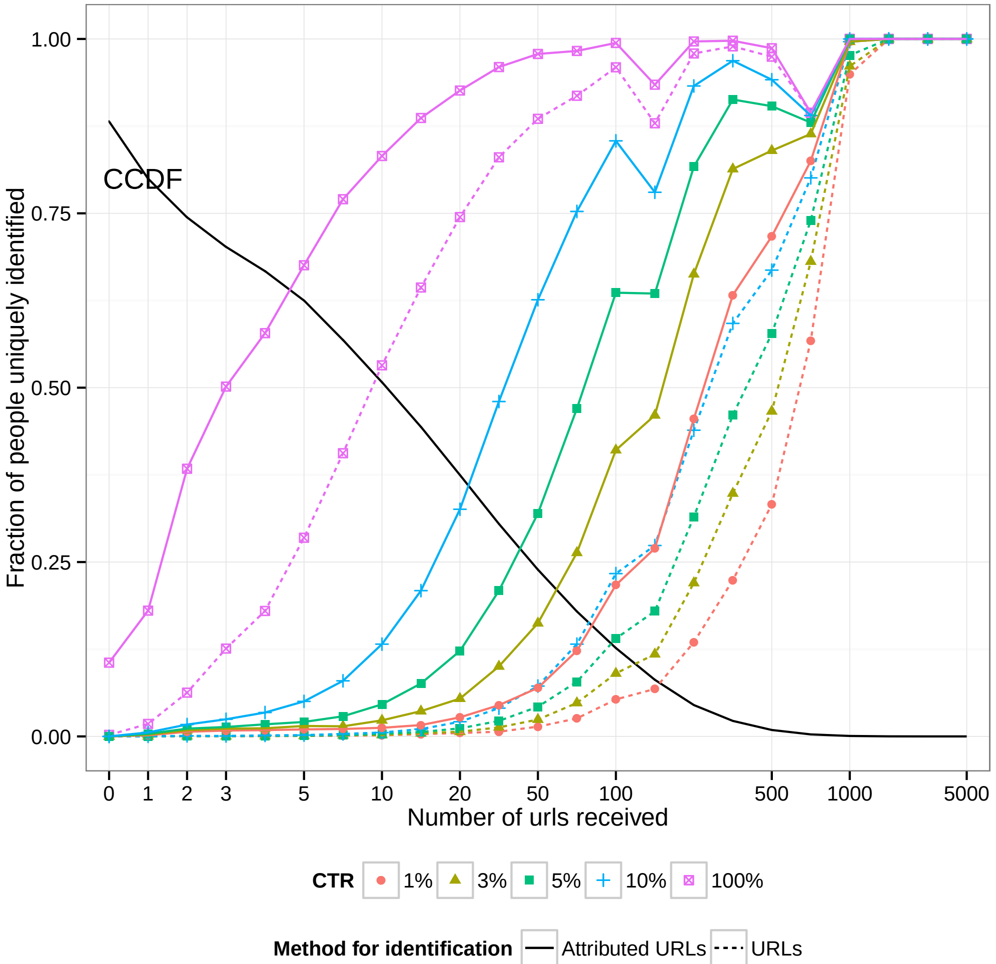

Figure 1.4: Fraction of users with at least n URLs received, and the proportion that are identified for various click generation rates and two methods.

`bit.ly`, and we use this API to obtain the number of clicks that each of those URLs generated. Dividing by the number of times this URLs is received in our dataset yields for each URL a coefficient. We scaled these coefficients by a constant so that the effective Click-Through-Rate (CTR) experienced by URLs posted on Twitter is 1%, 2%, 5% and 10% overall, chosen to represent a range of plausible hypotheses on CTR [Richardson, Dominowska, and Ragno 2007; Zhang et al. 2011]. Note that our method is approximate (*i.e.,* the measured clicks may be generated through other sources than Twitter), but it still captures heterogeneous popularity of URLs, most notably that rare URLs are less likely to generate click, under normalized conditions.

Figure 1.4 shows the fraction of users identified with the intersection method, assuming our click generation model. The qualitative trend is not surprising, the identifiability of a node depends on the number of URLs it clicks and is also inversely proportional to the click rate. If one out of twenty URLs get clicked, we can successfully identify 40% of the traffic, and for a CTR of 1%, more than 99% are users are left unidentified, since the success probably is low for anyone unless they receive at least 1,500 URLs.

1.2.4 Refining attribution with time information

Our next method is inspired by recent advances to use time in the inference of links and diffusion on social media [Gomez-Rodriguez, Leskovec, and Krause 2012; Rodriguez, Leskovec, and Schölkopf 2013]. Leveraging the fact that most clicks occur within a very short time of the URLs being posted, one can reconstruct minimal graphs to account for the visit times using convex optimization techniques. These rely on the time differences between visits of users to estimate probabilities of follower relationships existing among them. The intuition is that visits that are closer together in time are more likely to be related to each other in the social graph.

Another recent development thats helps in attribution is the adoption and use of diffusion tracking mechanisms such as BuzzFeed's pound system [Goldstein, Goel, and Watts 2015]. In such systems, each individual sharing a link has some code associated with their link which, when combined with timing information, allows us to attribute a click to an individual sharing online.

In our analyses, we use an idealized form of attribution where we assume either perfect attribution (which could be done through one of the means describe above) or attribution with errors.

We now utilize a method, Remember-Attribute-Intersect (RAI) (Algorithm 1), a three phase algorithm which uses methods of influence detection to attribute URLs to their social media source. While simulating the entire inference relies on information about click times, which is difficult to obtain and beyond the scope of this study, we conduct simulation assuming that the attribution steps succeeds with some probability in finding the source, or otherwise introduces an attribution error.

ALGORITHM 1: Intersect Algorithm of RAI

```
Data: Social Network $G(V, E)$: Node $v \in V$; URLs visited $u \in URLs(v)$
Result: $I(v)$ = identity of $v$; $f \subseteq$ Friends$(v)$ used for re-identification
Identities$(v) \leftarrow V$;
while Identities$(v) > 1$ and $\exists$ URL visit $u$ do
    I$(v) \leftarrow (\cup_{f \text{ post } u \text{ \& } v \text{ visits } u \text{ via } f}\text{Followers}(f)) \cap \text{I}(v)$;
    $u \leftarrow$ next visited URL;
end
if Length(Identities$(v)$) $\neq 1$ then // no unique identity found
    while Length(I$(v)$) $> 1$ and $\exists$ URL visit $u$ do
        I$(v) \leftarrow (\cup_{f \text{ post } u}\text{Followers}(f)) \cap \text{I}(v)$;
        $u \leftarrow$ next visited URL;
    end
end
```

Results with perfect attribution

We applied our method to *NYT* and recovered a significant fraction of the individuals. Figure 1.4 compares the performance of the two methods – the baseline method using just URLs and the modified RAI method with various CTRs. We see that there is a significant advantage in using attribution over the baseline. Even at low clicks rates of 1% and 5%, we capture individuals receiving only 100 URLs, which is a more typical user.

When we examined characteristics of the individuals used for re-identification, we find that the the set of individuals useful in identification were not significantly less popular than the others, indicating that our method does not rely on the inactive and less detectable individuals.

Effect of attribution error

All of our previous analysis assume that the attribution of URLs works perfectly. However, this stage is susceptible to error from various sources. We simulate errors as follows: for a

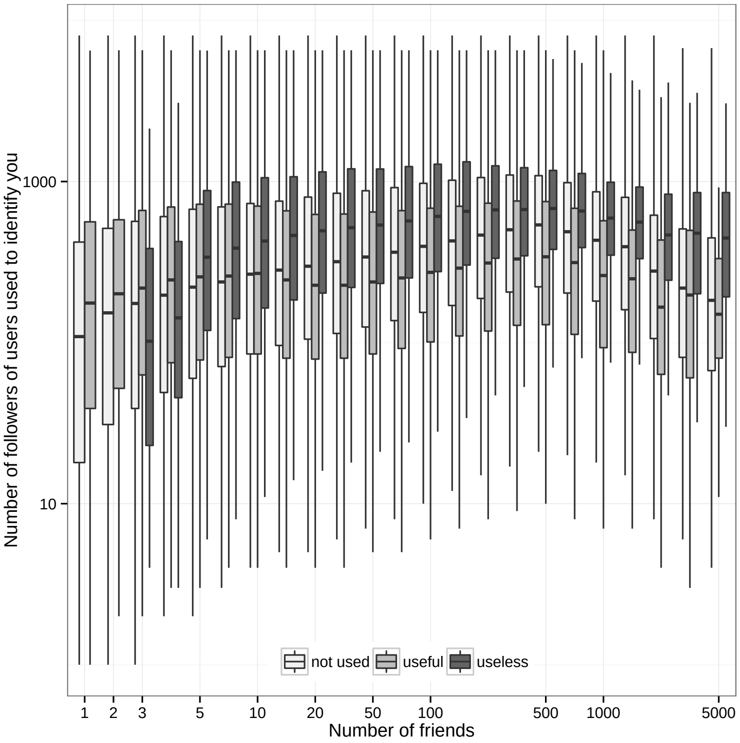

Figure 1.5: Distribution of followers for active friends used in re-identification.

certain error rate, we randomly attribute a URL incorrectly to another user who also posted that URL. Note that the set of URLs received by a web visitor is always correct under the persistence assumption. We then run a modified version of RAI. With incorrect attribution, it is possible that the intersection of followers of the attributed sources are not consistent and lead to no possible identifications. In that case, we back off, and run our initial URL re-identification method which is not susceptible to errors.

Figure 1.6 presents the results of RAI under error rates ranging from 5% to 45%, assuming a 5% uniform click through rate. The performance of RAI with errors falls somewhere between the two methods with perfect information. Indeed, for individuals receiving more than 1000 URLs, the performance is virtually unaffected by errors. In most cases, there are still gains in individuals identified with attribution which were previously not possibly by the URLs alone.

Table 1.1 shows the overall percentages of visits and people re-identified with both methods with varying CTRs and error rates. In a perfect information scenario, we can

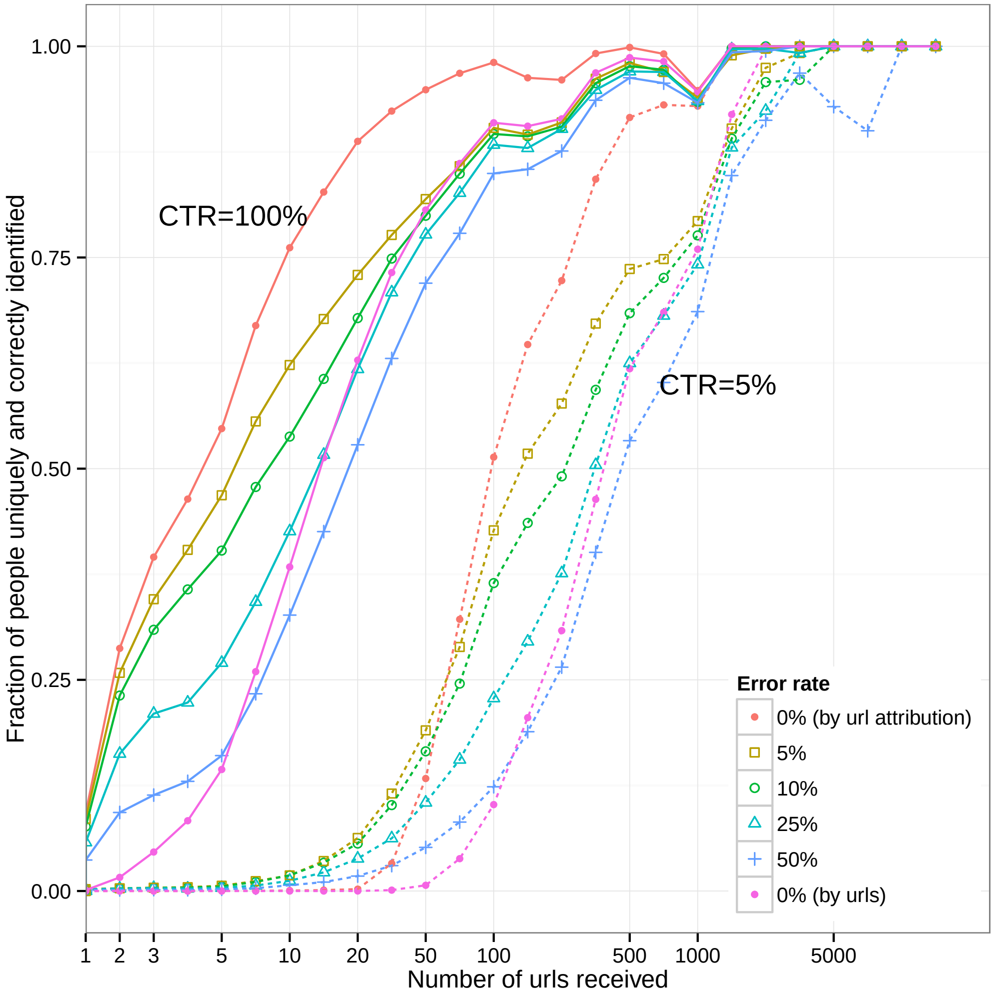

Figure 1.6: Fraction of nodes identified with attribution errors for two CTRs.

account for 97% of the visits. Even in more restrictive scenarios, with 5% CTR and significant error, we can account for about half of the visits and almost 20% of the whole population.

1.3 Conclusion

Limitations While the method we describe is applicable by many content publishers and aggregators, the method relies on certain assumptions.

Firstly, our method relies on identities being persistent over the time period being ex-

Error Rate		CTR			
		100%	30%	5%	1%
(a) 0%	% visits	97.0%	91.9%	70.3%	31.7%
	% visitors	69.1%	49.3%	19.5%	3.7%
(a) 5%	% visits	91.5%	83.1%	59.4%	
	% visitors	60.8%	43.4%	18.1%	
	(false. pos.)	(8.5%)	(8.8%)	(5.7%)	
(a) 20%	% visits	89.5%	77.9%	47.6%	
	% visitors	52.6%	34.2%	11.9%	
	(false. pos.)	(15.9%)	(18.1%)	(15.1%)	
(b) N/A	% visits	91.0%	79.2%	43.8%	11.8%
	% visitors	53.7%	30.5%	7.9%	0.8%

Table 1.1: Fraction of visits and visitors identified by (a) attributed URLs with varying error rates, and (b) unattributed URLs, for different CTRs.

amined. In Section 1.2, we covered some of the means to do so. Some of the situations in which this might not hold (and thus our method will fail) include a content publisher not tracking anonymous users by any means or a content aggregator not combining behavior from multiple sources.

Secondly, our analysis relies on observing posting behavior over a sufficient period of time. We found that a week was sufficient for a domain such as nytimes.com or digg.com. More generally our experiments indicate that one would need to collect enough data so that most users receive at least 8 posts. This might not apply in some situations (*e.g.,* very new websites, very unpopular sites, domains that are relevant only for a short period of time).

Thirdly, we require that the content publisher crawls or otherwise obtains the associated social graph for that time period. For content providers invested in an analysis of their audience, this does not prove to be a major impediment but can be resource and time intensive. Since the intersection method use simple operations, it could be run on a selective set of URLs (*e.g.,* , ignoring the most popular ones posted many times) or to re-identify only some particular set of visitors. Both of these cases may reduce the cost of this step.

Lastly, we assume that clicks are generated from users of twitter.com via their feed and not through other means (*e.g.,* search or promoted content). If the latter occurs occasionally, one might detect that as an attribution error, ignoring that URLs and applying the method on the rest

Prevention The limitations of our method also give rise to potential avenues of prevention of such an attack.

As a user, the most straightforward means to prevent this type of attribution attack is simply to use private browsing or some other means to break the condition of a persistent id. Private browsing typically does not store cookies (an easy way for a content publisher to have persistent local identity). A more extreme form would be to use public computers or to use an operating system that does not store state (such as Puppy Linux).

Since the method relies on having access to the relationships in the social network, one can also prevent this attack by relying on those networks which are predominantly private (*e.g.,* Facebook) or by ensuring that all the user's friends are private and not visible to a public crawl.

A more costly preventive mechanism is one in which the user carefully builds their social relationships such that they are indistinguishable from other individuals. However, this doesn't provide perfect protection as they would only be K-anonymous where K is the number of individuals who have identical relationships.

If the user is only particular about certain types of behavior being exposed, he could also create different accounts (and therefore different social relationships) associated with different types of consumption. For example, creating three separate profiles to access content by news media, celebrities, and personal friends.

With this technique, we show that even passive interactions with public conversation can broadly identify an individual. Even with its limitation, preventing such attacks require investment on the part of the user.

Chapter 2

Click Inference from Public and Private data

When available data is lacking, inference methods applied to social networks can prove to be especially useful. In the previous chapter, we saw this applied in the context of identity. Here, we examine the realm of understanding how links distributed on social media generate web traffic, or clicks. Progress on this essential question was essentially halted by lack of publicly available data. With the confidentiality of large-scale individual-level data on social media click habits, one can only hope to study this problem with aggregate data.

A common motivation of our work and several others is to study propagation to quantify influence online [Cha et al. 2010; Bakshy et al. 2011], how news sharing is affected by social networks [An et al. 2011; An et al. 2014], and various mechanisms and drivers behind retweeting links [Kwak et al. 2010; Boyd, Golder, and Lotan 2010]. Our work complements this line of work as it makes it possible to analyze reading habits, which was previously ignored. Most prior studies of online clicking habits are specifically targeted at online advertising. Models attempt at measuring the quality of an ad, and the relevance of personalization using its Click-Through-Rate (CTR) a metric resembling CPI in our context [Farahat and Bailey 2012; McMahan et al. 2013].

However, we show in this chapter that publicly available data is sufficient in elucidating the process of click conversion on social media. Here we leverage recent methods to simultaneously study shares and clicks on Twitter from a leading news domain using publicly available data only. Our goal is to describe and predict click dynamics for the entire lifespan of a URL posted on social media. We show that even with this scarcity of data, a dynamic prediction model that does not use this proprietary information can leverage the invariant properties of each step to its benefit and lead to accurate and fast prediction of clicks. In this section, we present the following contributions:

- We decouple the process of click creation into two separate stages: from posting to impression and impressions to clicks. This process involves measuring social media impressions – a metric that is generally not observable. We then continue and describe the temporal dynamics of both stages identified above. We find that the click dynamics are affected by time, the posting account, and the content in ways in which these factors remain seemingly independent of each other. (§2.1)
- Equipped with this insight, we develop a model which predicts temporal impressions, and then temporal clicks from those impression predictions. (§2.2)

Much of the prediction literature focuses on predicting the sharing behavior and cascade size. This direction follows an implicit assumption that content popularity by number of user shares and attributions can serve as a proxy for popularity by clicks. Some of the prediction methods are based on better modeling the underlying process by accounting for external influence [Myers, Zhu, and Leskovec 2012; Rizoiu et al. 2017] or using more complex models for the underlying diffusion [Shen et al. 2014; Zhao et al. 2015; Altman 2015]. Other methods evaluate the use of different types of features, of which the most useful features for prediction are time-based features. This time component can be differently accounted for by including a Bayesian approach [Zaman, Fox, and Bradlow 2014], reformulating the problem into several stages [Cheng et al. 2014], and classifying cascades based on their temporal evolution [Yang and Leskovec 2011]. Our work builds on this idea, further validating the use of time-based features. Another type of question asked in prediction literature is how much early success is indicative of longer range success and how much early information is needed [Szabo and Huberman 2010]. Our method further push the limits on the early information needed by relying on only the first hour of information. Further, our work takes into account the next crucial phase of the information flow – how do those shares relate to the clicks a link receives?

2.1 Understanding Click Dynamics

For online content, the click through rate (CTR) is defined as the probability that a reader clicks on the link when that it is shown to her. It has been studied and used in multiple applications including models of web surfing [Hubert, Hubert, and Mugizi 2006], ranking of search results [Chebolu and Melsted 2008], and optimization of online ads [Richardson, Dominowska, and Ragno 2007; Farahat and Bailey 2012]. With public access to impression data being limited, we introduce two new definitions for CTR:

- Clicks Per Impression (CPI): $\frac{\#\ clicks}{\#\ impressions}$

- Clicks Per Reception (CPR): $\frac{\#\ clicks}{\sum_{u \in U} \# followers(u)}$ where $U =$ the set of users tweeting or retweeting the link. Here we count one reception for each Twitter users who are potentially exposed to a tweet (i.e., who follow an account that shared the URLs).

The main difference between the two metrics is the computation of audience size. CPI is the traditionally used metric to evaluate click through rate. With CPI, we consider the audience as the number of Twitter users who have been exposed to the URL, or the number of impressions. While this is an accurate measurement of CTR, it is often hard to measure with public data. In contrast, for CPR, the audience to be the number of receptions. This method can overestimate the number of impressions and capture too much noise since number of receptions fails to account for 1) the overlap of follower sets and 2) the level of

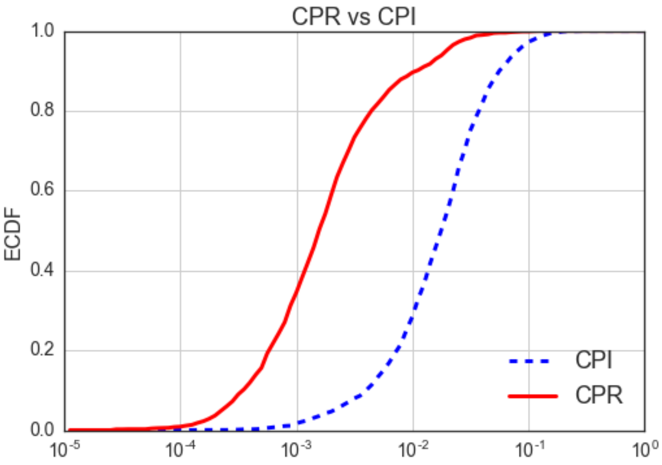

Figure 2.1: Clicks per Reception (Red) and Clicks per Impression (Blue). CPI is computed from the publisher dataset. CPR is computed entirely from public data.

activity of followers. While one can theoretically compute the number of unique receptions to account for overlap, the number of API queries involved quickly makes this prohibitively expensive. Previous work quantified this overestimation from overlap, finding it is less than 20% for 75% of reception counts [Gabielkov et al. 2016].

Figure 2.1 compares the ecdf distributions of CPI and CPR for each URL. Note that CPI is computed from our publisher dataset and CPR is computed entirely from public data: the number of clicks from bit.ly, and the number of receptions from Twitter. While the magnitudes of CPI and CPR differ by a factor, they follow the same general trend.

The Effects of Retweeting on Clicks Click rate metrics give an overall view into the performance of a URL. However they miss insight into the cause of the readership of content - its diffusion and sharing characteristics. We would expect that retweeting features bear a strong relationship to the audience and the eventual readership since the act of retweeting is the primary mechanism to generate an audience. This type of analysis is relatively new with the work of Gabielkov et al. being the first to examine this relationship [Gabielkov et al. 2016]. They found that there was a strong positive correlation between the number of retweets of a link and the number of clicks. We expand on their work by examining the relationship in content of different audience sizes with a broader set of type of content, rather than traditional news alone.

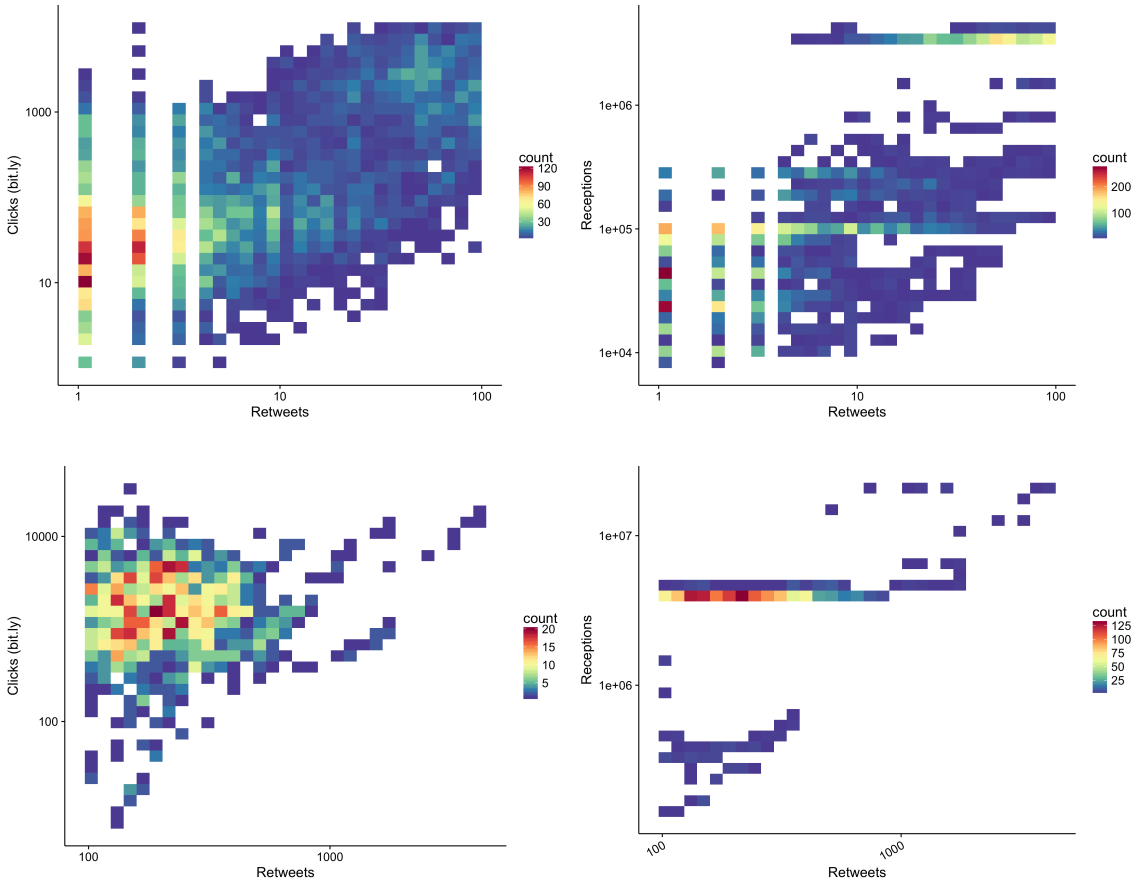

Figure 2.2: The effect of sharing on clicks and impressions, at different sharing magnitudes. (top) Retweets ≤ 100 show a strong positive correlation in both clicks and impressions while (bottom) Retweets > 100 show a positive correlation in impressions, but no correlation in clicks.

In these analyses, we use publicly attained retweets, estimated impressions, estimated clicks, and the estimated clicks per impression to analyze the relation of clicks and click rate with share rate.

We first found that the relationship between retweeting and CPI demonstrates a law of diminishing returns of clicks, as it has a slight negative correlation (Pearson's $r = -0.063$, p value$= 7.582e - 10$). However, looking into the relationship between retweets and absolute number of clicks presents another picture of the effect of endorsements on news item reach on Twitter. Here the results suggests a law of *no* returns. While this limit on reach has been previously observed in social media, those studies are based in a different

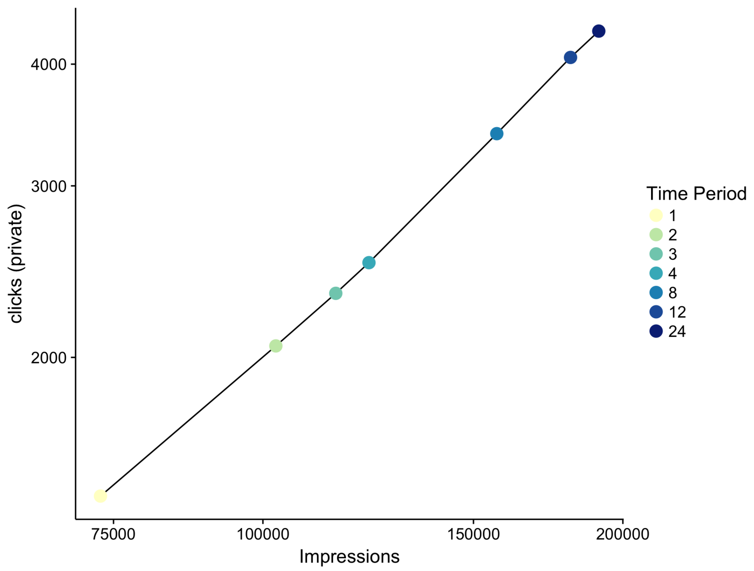

Figure 2.3: Trajectory plot for a single URL with clicks on the y-axis and impressions on the x-axis. The colored points indicating the time of the measurement.

setting, and define reach by re-shares rather than clicks [Myers, Zhu, and Leskovec 2012].

We see a threshold effect at ~ 100 retweets, above which the clicking and sharing relationship changes. When number of retweets is < 100 (Figure 2.2 (top)), clicks and shares are positively correlated (Pearson's $r = 0.55$). However, past this threshold, increasing retweets does not translate to increasing clicks (Figure 2.2 (bottom)) (Pearson's $r = 0.120$). A diminishing impression rate could explain this exhaustion of reach, but not entirely. While the growth rate of impressions diminishes, impressions do still increase with sharing with a Pearson's $r = 0.64$ (Figure 2.2 (bottom)). Unlike clicks, we don't yet observe a limit to the growth of impressions in our scope of sharing magnitude.

Given the complex relationship between posts and clicks or impressions, we cannot use a straightforward model to estimate the number of clicks (or even impressions) from easily available public posting data. Indeed, we see in the next section that when you consider the evolution of a tweet over its lifetime, that relationship is further complicated.

2.1.1 Evolution of Clicks

To better understand the relationship between posting behavior and clicks or impressions, we study the evolution of the characteristics of a link evolve. Figure 2.3 show a sample

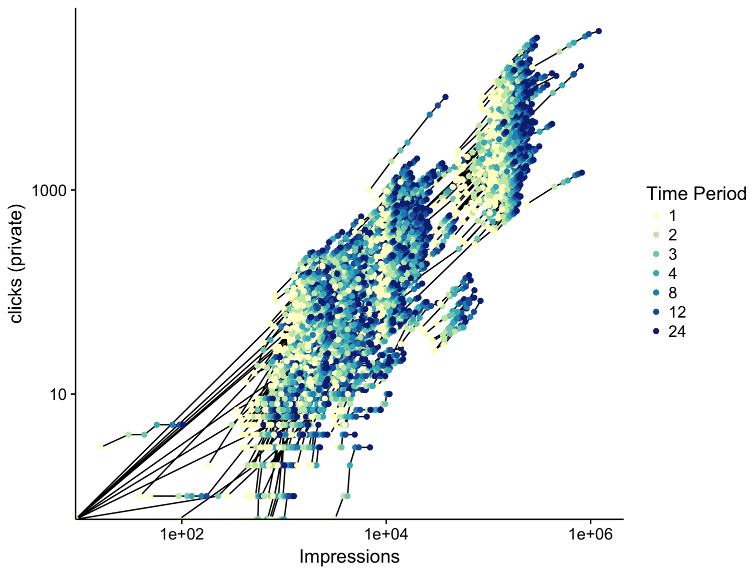

Figure 2.4: Trajectory plot for each URL with clicks on the y-axis and impressions on the x-axis. Each line represents a single URL with the colored points indicating the time of the measurement.

trajectory plot for a single URL for clarity. The black line represents a URL with each point representing a different point in time (in hours). The lighter points are earlier and the darker points are later in the URL's lifetime. The x-axis is the cumulative impressions (or receptions) in log-scale. The y-axis is the cumulative observed clicks in log-scale. While we expect to see the increasing trend (since it is cumulative), it is interesting to note the near linearity (in log-scale).

We see this bear out in the full dataset. The full trajectory plot presents a map of the the evolution of each URL (Figure 2.4). In the impressions trajectory plot (left), we observe that most of the trajectories are roughly linear and run roughly parallel to each other, indicating a close linear relationship between the two quantities.

We observe some level of clustering with URLs with similar audience size (on the x-axis). For instance, there is a large group of URLs which reach an initial audience of 10k. These URLs are shared by the BuzzFeed main account. The different clusters roughly correspond to the different BuzzFeed accounts and we can see that overall, they exhibit similar behavior.

We also notice that, even among the URLs with the same number of impressions, there is a very wide range of clicks. We see that each URL seems to have its own specific CPI,

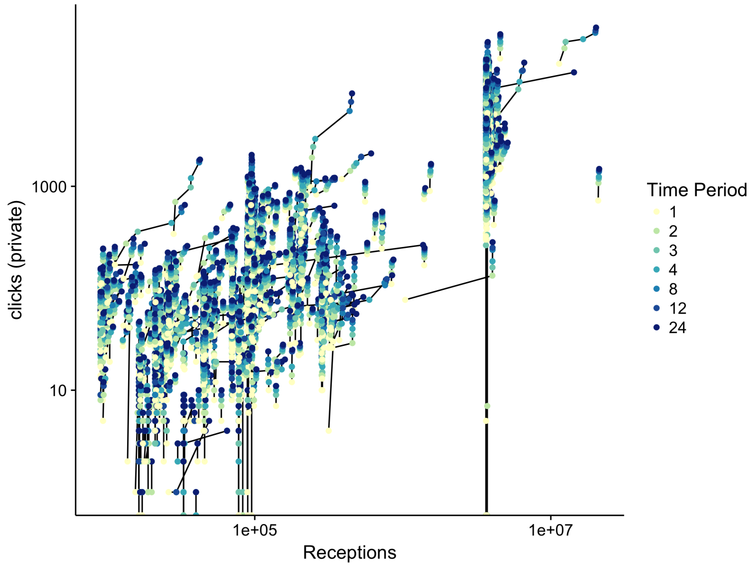

Figure 2.5: Trajectory plot for each URL with clicks on the y-axis and receptions on the x-axis. Each line represents a single URL with the colored points indicating the time of the measurement.

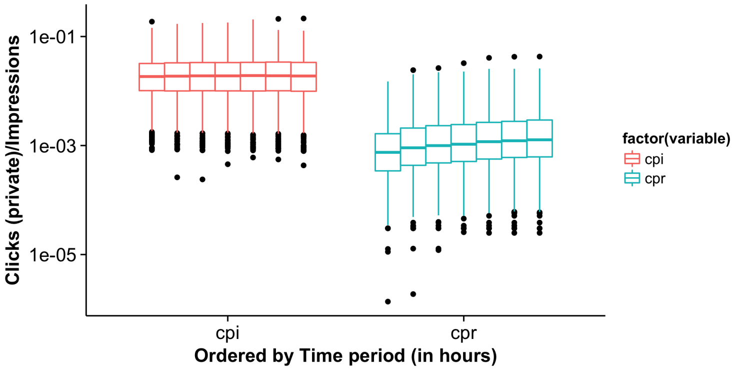

Figure 2.6: Clicks per Reception (Blue) and Clicks per Impression (Red) ordered by the time period. CPI is relatively stable over time, while CPR shows some increase.

as evidenced by the stacked trajectories. Within a single account, the shared URLs have very similar sharing characteristics (and thus, similar number of impressions). However, there is a wide range of about an order of magnitude of the number of clicks that the same number of impressions translates to. This CPI seems to be content-based, *i.e.,* some URLs will intrinsically perform better than others.

Similar analyses using receptions instead of impressions show a staggering difference 2.5. This difference stems from the way impressions and receptions are measured - impressions are measured at the time of viewing a link whereas receptions are measured at the time of posting a link.

In addition to the raw clicks and impressions, we consider the distribution of CPR and CPI as it changes with time (Figure 2.6). Each box represents the distribution of the

CTR metric in that time period. While CPR shows a definite increase with time, CPI is more stable. This stability confirms, as did Figure 2.4 (left), that clicks follow impressions independently of when the content was shared and by which source. Hence the structure of sharing in the network primarily affect the clicks produced by a Tweet solely through the number of impressions that this Tweet produces. Note that, once a source decides to post a link, impressions depend on how various followers of that source actually update their feed, whether they actively read its feed, perhaps through a list. It does not, however, depends on the content of the URL itself.

The above observations on how clicks get generated motivates us to propose a simple two stage model to extrapolate how many clicks are received. Stage 1 (described in §2.2.1) is graph dependent and describes how shared posts translates to impressions. This stage incorporates information about which users are sharing the posts and their relative times of sharing. Stage 2 (described in §2.2.2) is content dependent and establishes the translation of impressions to clicks. This model allows one to extrapolate clicks as created by impression without even requiring to measure them. Since Stage 1 depends on the source but not the content of the link, Stage 2 is the exact opposite: it depends on the content of the link but not on the source and time that create an impression. This separation allows different part of the model to be tuned using public information only.

2.2 Dynamically Predicting Clicks

We now describe how to implement a simple two stage model of click generation from social media using a dynamic estimation of impressions produced. This model is shown to improve accuracy for two general prediction problems: First, a real time click prediction, in which one attempts to deduce the amount of clicks produced by social media conversations from its associated tweets. Second, a day-ahead forecast, where the goal is to estimate the total clicks gathered at the end of the day from a single observation obtained after one hour.

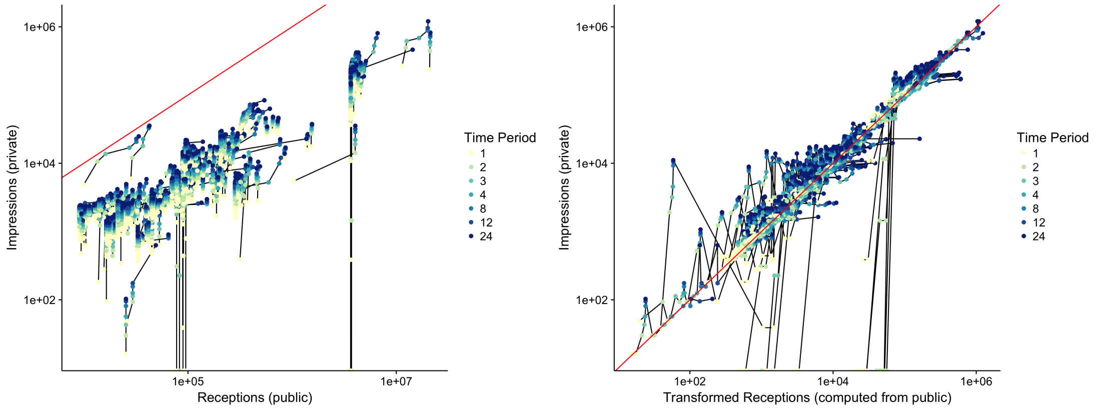

Figure 2.7: Relationship of impressions to (left) receptions and (right) predicted impressions. Each line represents a single URL and each point represents the time period. The red line is the identity line for comparison.

2.2.1 A Memory-less Model of Impressions

Impressions produced by links on social media are rarely publicly known. As seen before, receptions which are publicly known can function as a proxy, but only when considering total counts over long periods of time. This is because they exhibit very different dynamics. (Figure 2.7 (left)), owing to the time lag between when a link is posted (when receptions are counted) and viewed (when impressions increases). In fact most of the receptions of a given link occur within the first hour, confirming the trend that while shares usually occur early in the life-cycle of a tweet, impressions and, therefore, clicks occur later in its life ([Gabielkov et al. 2016]).

We propose a simple memory-less model in which a fraction q of the receptions are 'activated' in each of the T time periods. Following that, all activated impressions are removed, and the process repeats on the remaining fraction $(1 - q)$ in the next time period, and so on until none are left. Of all activated receptions in a given time slot (which corresponds for instance, to the potential viewer logging to Twitter), we assume a fraction s will create an impression.

Let x_t be the observed number of receptions in time $t \in [1, T]$ and $\tilde{y}_t$ be the predicted

Table 2.1: Mean and median of scaling parameter s and geometric parameter q.

Training	**mean(s)**	**median(s)**	**mean (q)**	**median (q)**
Hour 1	0.08	0.07	0.32	0.32
Hour 1 to Hour 4	0.46	0.06	0.32	0.28

cumulative number of impressions in time t. Then we have

$$\tilde{y}_t = s \cdot \sum_{\tau=1}^{t} x_\tau \cdot (1 - q^{t-\tau+1}).$$

Loss function For each article, we learn the model parameters s and q by minimizing the square error between the predicted number of clicks ($\tilde{y}_t \forall t \in [1, T]$) and the actual number of clicks ($y_t \forall t \in [1, T]$). The optimization used is

$$\min_{s,q} J = \frac{1}{2} \sum_{\tau=1}^{T} (y_\tau - \tilde{y}_\tau)^2.$$

Parameter Optimization We used the L-BFGS-B optimization method [Liu and Nocedal 1989] from Python's SciPy toolkit [Jones, Oliphant, Peterson, et al. 2001–] for the minimization. We learn the parameters individually for each article shared. The means and medians of the parameter values (split by the training data used) are given by Table 2.1.

We compare these predicted impressions with the actual impressions (Figure 2.7 (right)). Here, each line represents a single URL shared and so we see the accumulation of impressions with time. We see that the transformation according to the model accomplishes two things: (1) scaling down the over-estimated receptions and (2) correcting for the time bias in the reception measurement.

2.2.2 Predicting Clicks from Impressions

The memory-less model in §2.2.1 predicts the degree of *exposure* or potential audience of an article. We observe that the ratio of clicks per impression remains stable over time (Figure 2.6). Thus, we can use a simple multiplicative factor to estimate the clicks from

Table 2.2: Mean absolute percentage error (MAPE) and adjusted MAPE of the prediction by the memory-less model.

Training	MAPE	MAPE (adjusted)
Hour 1	17.29 %	16.20 %
Hour 1 to Hour 4	12.68 %	11.97%

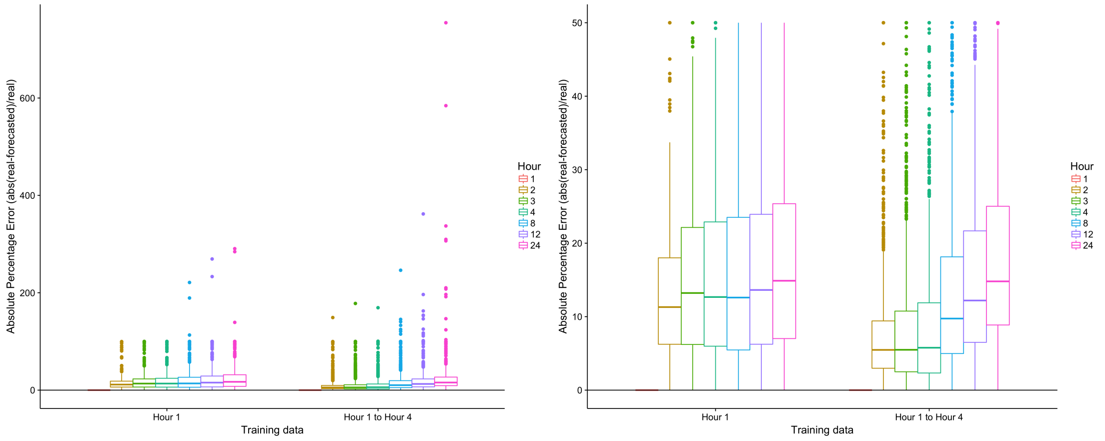

Figure 2.8: Distribution of adjusted absolute percentage error (MAPE) for each time period (right) and truncated to better see the distributions (left).

the predicted impressions at hour t ($\tilde{y}_t$). This factor is computed at the first time period as the clicks-to-impressions ratio at time $t = 1$ *i.e.*, $CPI_{t=1} = \frac{z'_1}{\tilde{y}_1}$ where z'_t is the observed number of (bit.ly) clicks at time t. Let $\tilde{z}_t$ be the predicted cumulative number of clicks at time t. Then

$$\tilde{z}_t = CPI_{t=1} \cdot \tilde{y}_t.$$

In order to predict clicks, we combine the two phases – the geometric memory-less model and the multiplicative prediction. In our evaluation, we use two sets of training data (1) the first hour of shares and clicks (2) the first four hours of shares and clicks. Table 2.2 indicates the mean absolute percentage error (MAPE) excluding those articles with zero clicks (which have an infinite MAPE). We also compute an adjusted MAPE value with every prediction and actual value being incremented by 1 to capture the errors for those articles with zero clicks. The distribution of the absolute percentage errors for each time period shows that while later time periods have slightly more errors, the increase is not

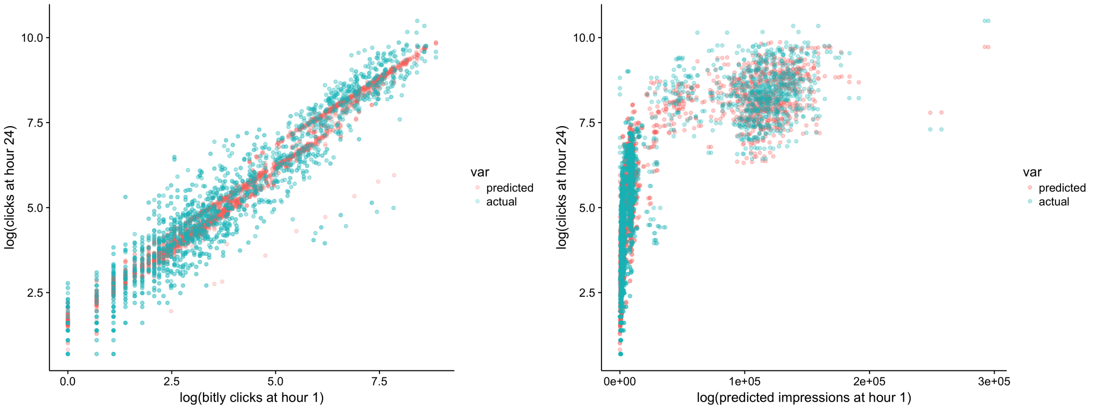

Figure 2.9: Predicted ($\tilde{z}_T$) and actual (z_T) total clicks vs (left) bit.ly clicks at hour 1 (z'_1) and (right) predicted impressions at hour 1 ($\tilde{y}_1$)

dramatic (Figure 2.8). Further we see that most of the predictions are actually within a narrow range of the correct value (less than 10%).

2.2.3 Predicting Future Clicks

In the previous method, we used hourly `bit.ly` clicks and hourly reception counts, both public forms of data, to estimate hourly clicks and hourly impressions, as validated by our private data. The method in fact, uses share information from the *current* time period to predict the number of clicks the link will receive in the same time period. A more predictive model would be to predict *future* click information. We extend our model, using early data about clicks and our impression predictions to predict future clicks.

We created a linear model for predicting the total number of (private data) clicks on a URL ($\tilde{z}_T$), from just the observed number of (bit.ly) clicks at time 1 (z'_1) and our predicted impressions at time 1 ($\tilde{y}_1$). The model is defined as:

$$\log(\tilde{z}_T) = \beta_0 + \beta_1 \log(z'_1) + \beta_2 \log(\tilde{y}_1).$$

Using 10-fold cross validation, we arrived at an MAPE of 12.4% (adjusted for zero values, MAPE=9.23%) (computed in log-space). This can be seen in Figure 2.9.

In fact, using the first hour of clicks *alone* proves predictive of future success ($MAPE = 13.3\%$, adjusted $MAPE = 10\%$). We find modest improvements (not shown) by considering the fourth hour of clicks. The addition of account-based features (e.g. posting frequency, median total CPR) also provided only trivial improvement, giving us further clue that they don't provide much additional impact on the dynamics of the clicks aside from what hour 1 clicks and impressions already capture.

2.2.4 Comparison with Other Prediction Methods

Our model is one of the few to predict clicks rather than shares or posts, making a direct comparison challenging. In the following sections, we tackle this in the following ways:

- Adapting share prediction models to predict future clicks: By modifying other methods to predict clicks, we can directly compare our day-ahead click prediction with their models.
- Comparison of our method across different BuzzFeed accounts: The different accounts in BuzzFeed have very different behavior from news to lifestyle related posts. By training on one set to predict others, we can verify that our methods results are not dependent on a specific type of behavior but rather can generalize to other datasets.
- Comparison to hourly prediction methods: New diffusion prediction models predict popularity at every time period. We can use these as direct comparisons with our hourly prediction methods.

Comparison with Adjusted Prediction Methods

One means by which we address this is to adapt previous methods to predict click information rather than shares. In Table 2.3 we compare the results from our predictions with several other commonly used ones. In the model of Szabo et al. ([Szabo and Huberman 2010]), final share counts are predicted using the early information about the shares. In our adapted comparison, the click counts are predicted using the first hour of click information

Method	Feature(s)	MAPE	R^2
Day-Ahead Prediction	clicks in hour 1, impressions in hour 1	34%	0.903
Szabo et al. [Szabo and Huberman 2010]	clicks in hour 1	40%	0.76
Pinto et al. [Pinto, Almeida, and Gonçalves 2013]	clicks in hour 1, hour 2, hour 3, and hour 4	147%	0.73
Szabo et al. [Szabo and Huberman 2010] with click transform	shares in hour 1	78.8%	0.4
Pinto et al. [Pinto, Almeida, and Gonçalves 2013] with click transform	shares in hour 1, hour 2, hour 3, and hour 4	81.3%	0.41

Table 2.3: Prediction results for several methods.

in log-space. Pinto et al. ([Pinto, Almeida, and Gonçalves 2013]) leverage retweet counts from subsequent hours as additional features of a linear model. We tested this idea, using hour 1 and hour 4 clicks as the feature set. In both cases, we use 20-fold cross validation. We see that using our estimated impressions provides a significant boost (even over additional time periods of click behavior).

Another means by which we test our method is to combine principles from previous work as well as our current work. While we see that earlier methods might not be the most effective at predicting clicks, they have been shown to work well in predicting shares. The underlying reasoning for the usage of shares as a measure of popularity is that they are highly correlated to clicks. If this is the case, then there should be some constant fraction of shares that are converted to clicks. We computed this constant fraction from the publicly observed clicks and retweets at the end of the first hour of a clicks lifetime. Table 2.3 shows the results of this transformation. The results of this is dependent on the amount of information used to compute the clicks to shares ratio. For instance, using an entire day of data to compute the ratio results in much improved predictions.

We can also compare our click prediction against the click values derived from bit.ly. These bit.ly derived click values are publicly available but sometimes does not accurately reflect the actual observed clicks. Our model has smaller residuals than untouched bit.ly clicks do, across all time periods (Figure 2.10).

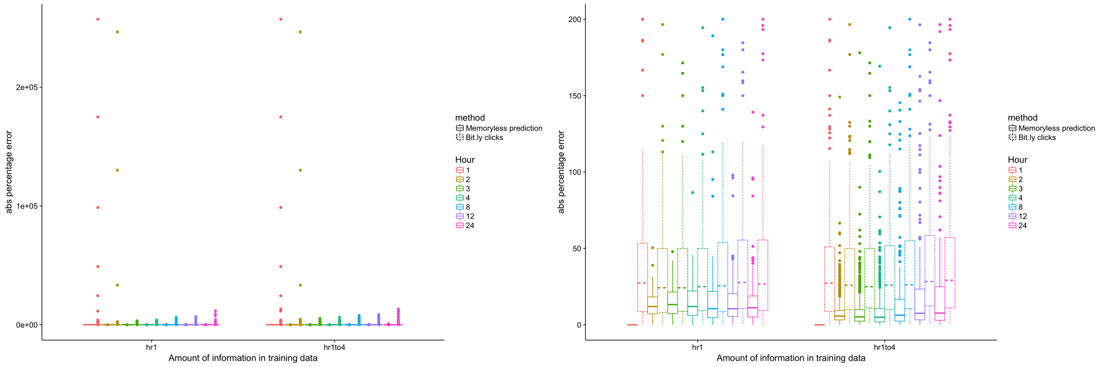

Figure 2.10: Distribution of adjusted absolute percentage error (MAPE) for our method (solid) compared to bit.ly (dotted) for each time period (right) and truncated to better see the distributions (left).

Compared to many other models, our model only relies on the first hour of data. In particular, it only relies on easy to obtain public information. As such, it provides greater utility in forecasting applications with time constraints.

Cross-Account Prediction

Both sharing and clicking behavior is different across different domains. This is quite visible in Figure 2.4 where each of the clusters correspond to different BuzzFeed accounts. In order to validate the generality of our approach, we compare our predictions on the various accounts. This also accounts for effects of content type *e.g.,* news vs quizzes. We hold out all the links and information for a particular Twitter account and train on the remaining accounts.

We do this for the 5 accounts with the most tweets in this particular dataset: the main BuzzFeed account, a quiz-based account, and three international edition accounts. We see results very similar to that of the mixed account datasets, in Figure 2.11. Our model's MAPE was, on average, around 30% (average R^2 = 0.25) across the accounts tested. Further, we compared this approach using other prediction models. The model of of Szabo et al. had an average MAPE of 50% (average R^2 = -0.58). The Pinto et al. model yielded a MAPE average of 340% (average R^2 = -2.14).

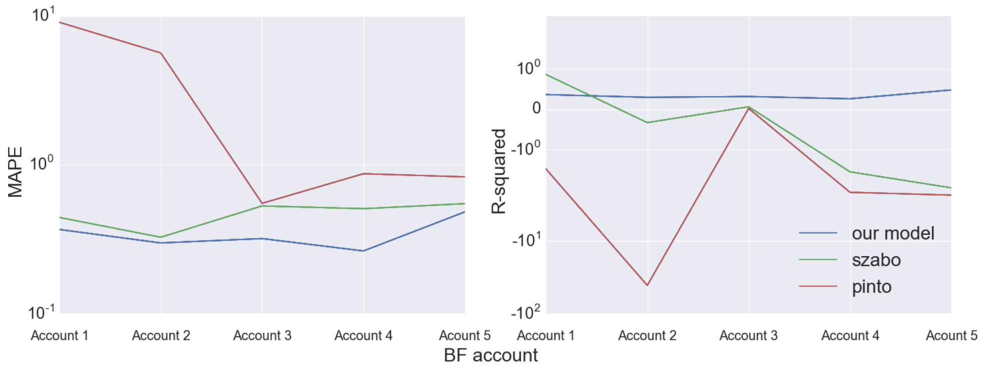

Figure 2.11: Predicting (day-ahead) clicks on held-out accounts. We compare the model's cross-account predictive performance to Szabo's and Pinto's, using Median Absolute Percent Error (Left) and R^2 (Right).

Comparison with Hourly Prediction Models

Recently, research has been focused on popularity prediction methods which rely on modeling the underlying diffusion [Shen et al. 2014; Zhao et al. 2015; Rizoiu et al. 2017]. These primarily rely on timing information about the diffusion. Further, several of these model exogenous and endogenous diffusion separately to better account for different processes that result from each.

At a high level, methods to forecast clicks typically assume some initial impact of a share, which then has a decaying impact over time. One paper on which we focus our comparison (HIP) uses a Hawkes Internsity Process to model the decay and accumulate impact from multiple shares [Rizoiu et al. 2017]. In this model, the forecasted number of views is the sum of the influence of current shares and the sum of influences of deprecated previous shares. They use a number of parameters to capture features like the type of content, network of diffusion, and sensitivity to promotion. The feature sets used are the number of shares of the link/video at each point in time. HIP predicts view counts of a YouTube video, a metric that corresponds to clicks for a URL link.

There are several differences between our geometric model and the HIP model. Our model is simpler, with only two parameters used (s and q) compared to the 5-6 used by

Table 2.4: Mean absolute percentage error (MAPE) and adjusted MAPE of the prediction by HIP and by our memory-less model.

Method	**Training**	**MAPE**	**MAPE (adjusted)**
Geometric model	Hour 1	17.29 %	16.20 %
HIP	Hour 1	102.2 %	99.5 %
Geometric model	Hour 1 to Hour 4	12.68 %	11.97%
HIP	Hour 1 to Hour 4	91.7 %	89.7 %

HIP. Our model also uses the additional information that comes from impressions. This is valuable because it not only captures how much a link is shared but also the potential influence of each share. Ours is memoryless and we only need to keep track of how many individuals are in the potential audience, not when they joined the audience pool. However, the main difference is the model itself. Our model uses a polynomial depreciation of influence which results in the influence of a share having a longer effect compared to the exponential depreciation used in HIP.

We implemented their model, modifying it to account for the varying granularity in our dataset. We used two feature sets in the comparison - the first hour and the first four hours of the lifespan of the post. Specifically, for the first hour of a post, we used the number of shares in that first hour. For our method, we also include the number of impressions in the first hour. Similarly, when using the first four hours, the feature sets were the the number of shares (and impressions) for hour 1, hour 2, hour 3 and hour4 of the lifespan of the tweet. We trained the models on the predicted number of clicks after the first hour (or the fourth hour). Note that this method has fewer data points because the (non-linear) optimization failed to converge.

In Table 2.4, we see that our method significantly outperforms HIP. One possibility for why this might be the case is that our method uses the the additional information of the receptions. To test this idea, we used their model with an input of the number of receptions (instead of shares). If this is the key difference, then, one would expect their model to match or outperform ours. We find that the MAPE is 102.2% for 1 hour of training data and 24.0% for 4 hours of training data (adjusted MAPE is 106.3% and 32.2% respectively).

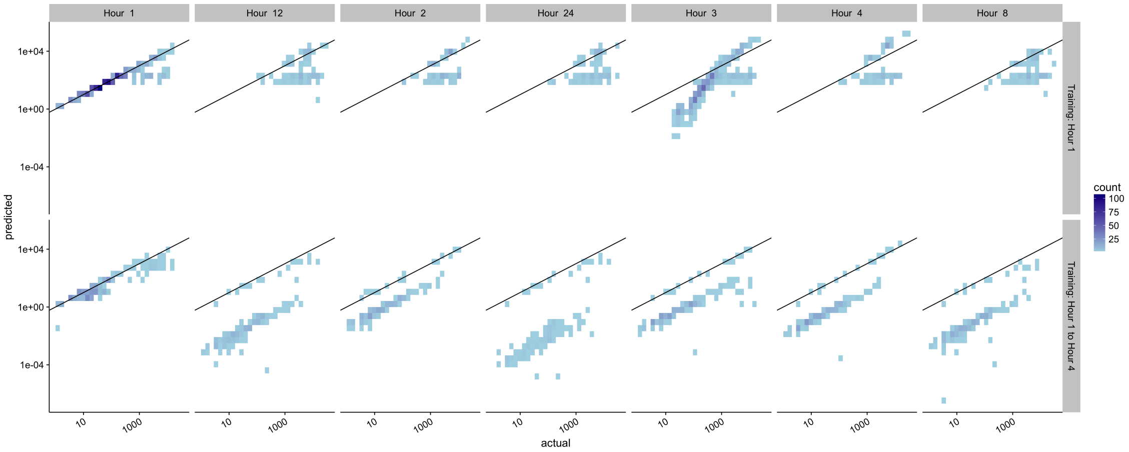

Figure 2.12: Scatter plot of the actual (x-axis) vs predicted (y-axis) values from the HIP model for training data of (top) 1 hour of data and (bottom) 4 hours of data. Each column shows the results of prediction at a particular time period.

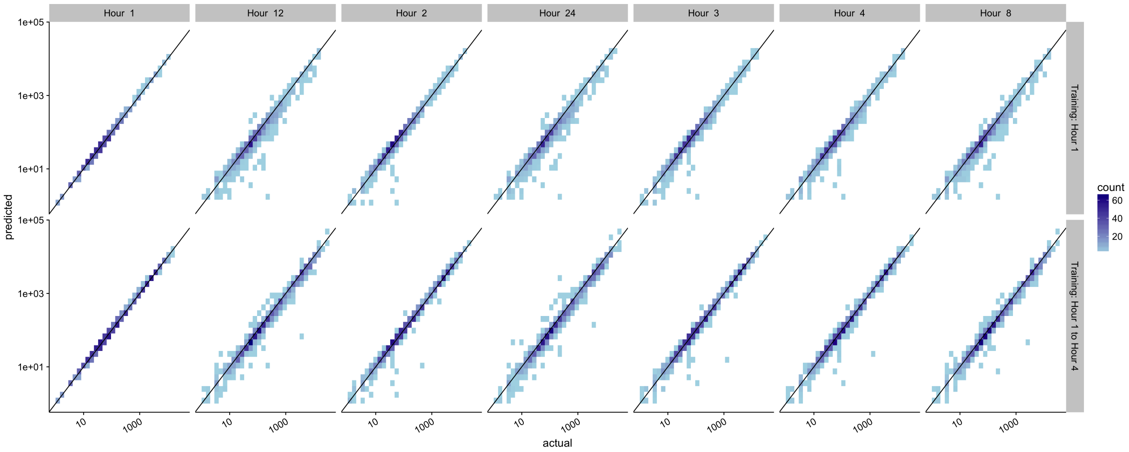

Figure 2.13: Scatter plot of the actual (x-axis) vs predicted (y-axis) values from our memory-less model for training data of (top) 1 hour of data and (bottom) 4 hours of data. Each column shows the results of prediction at a particular time period..

From the modeling perspective, there are differences as well. While fundamentally both assume a decaying impact of a share, they do so in different ways. Our memory-less exponential decrease is more short-lived while the polynomial decay in HIP assumes a longer-term impact. HIP habitually underestimates the number of clicks, even with 4 hours of data (Figure 2.12.) By comparison, our method closely aligns with the actual values (Figure 2.13).

2.3 Conclusion

In this chapter, we documented techniques to use multiple public data sources to predict private information. We validated the use of click-rate metrics derived from public information. We found that the time, account, and content affect the dynamics of a link but they all act independently of each other. In our analyses, we show that the relationship between shares and clicks is not straightforward. Rather, it is a two-stage process with a key intermediate piece of information, impressions. We developed a model based on this two-stage process to predict the temporal dynamics of a post using publicly available data. Our model reliably predicts the performance of a link based on very early information. Overall, we hope that our model and methodology will help foster better understanding of sharing and audience dynamics.

Since this method relies on data that is public (the shared posts as well as the number of followers of the poster), it would not apply in private settings such as Facebook. Such cases are yet another avenue of research, where the prediction model could relies on observed clicks/visitors to estimate the future popularity.

In Chapter 1 and 2, we see that the use of multiple sources of data can extend the power to learn about both the users of a network and the information shared in the network. With very minimal information about the shared information, we can infer its longer term performance. In the following chapters, we transition from studying the result of information production to its initial generation. We look at how this information is generated and how network structure might have an impact into its diffusion.

Chapter 3

Specialization in Static Networks

In social network services, such as Twitter and Facebook, the primary commodity produced and exchanged is content and information. While, arguably, much of this process is solely driven by personal gain, these social conversations play an increasingly larger role in today's economy. This is unsurprising since decades of empirical studies, predating any online conversation, have shown how individuals rely on their peers or contacts to acquire information before making a choice.

Our goal is to understand how individual choices govern how *original* information is produced and acquired in today's social networks. We focus on the domain of identification of news content worth reading, where social connections are massively used. Social networks benefit users by making the result of this effort available to more people. Previous studies highlight that only a minority of participants add information to those networks, as opposed to simply listening or passing it on (via, e.g., retweets, likes). Many important open questions remain: In a given network, which users have an incentive to produce more original content? Previous studies have shown that influencers are not easy to differentiate from ordinary users. Can we predict the outcome of such a mechanism, where some users specialize? Are there types of content or networks that favor the formation of an elite?

In this section, we show the following contributions:

1. We analyze data from multiple online sources exchanged through Twitter, highlighting the production of original content remains extremely concentrated. Barring institutional accounts, the majority of the original content comes from users with mid-range popularity rather than just the just well known people. In fact, counter-intuitively, original content production is skewed towards less active and connected people. (Section 3.1).
2. Since the availability of news worth reading in a social network exhibits the property of a public good, we propose a simple model that extend public good theory to accommodate investment made by individual players towards a perishable good. This model allows us to answer how specialization occurs in knowledge sharing, even where players are *ex ante* identical in structure and behavior *i.e.,* players are identical before they begin interacting with others and even after interacting, continue to behave in the same way as others. We first prove that a unique Nash Equilibrium exists for sufficiently short-lived content, under a condition related to spectral properties of the social network. However, we prove that when content is long-lived, specialization is unavoidable, even with identical players on a symmetric graph. Given the presence of multiple equilibria and sensitivity to initial conditions, predictions are complex. (Section 3.2).

3.0.1 Related Work

To model the content production choices of individuals, we draw upon literature regarding the analysis of the private provision of public goods, or investments made by players that more generally affect the outcome of others, which originally emerged to inform public policy. Its most celebrated result, the *neutrality principle* [Bergstrom, Blume, and Varian 1986], states that the investment produced by a group is entirely carried by most wealthy individuals, and is insensitive to income redistribution. This, however, holds only for a *global* public good in which all players are equally affected by others, and recently was shown not to generalize beyond regular graphs [Allouch 2015]. The general network case was studied more recently [Ballester, Calvó-Armengol, and Zenou 2006; Bramoullé and Kranton 2006; Bramoullé et al. 2014].Even for that simple case, predictions vastly differ: On the one hand, a study of small effects [Bramoullé et al. 2014] proves that the system converges to a unique equilibrium in which all participate.

On the other hand, more general cases prove that specialization is unavoidable, and that multiple equilibria can be attained [Bramoullé and Kranton 2006]. Our analysis extends those results by providing the first non-linear dynamics for which a similar dichotomy can be proved; in particular, it proves that a simple model of perishable public goods leads to either of these behaviors depending on the product lifespan. The main novelty of our approach is to model information as a public good with decaying value over time *i.e.,* they are perishable goods. As a public good, the utility of information to a user comes from her own contributions as well as those of her neighbors. This new approach allows theory and practice to qualitatively align, in spite of simplistic modeling of user behavior.

Our work also relates to studies of online diffusion of information which have previously established the importance of content produced by mass media in online diffusion. They highlight in particular that news typically reaches a large audience not directly but through a set of influencers or connectors [Cha et al. 2012; Wu et al. 2011]. This result confirms the classical hypothesis of a two-step information flow [Katz 1957], and was shown to have additional benefits, such as broadening the range of opinions seen by a user [An et al. 2011]. However, the dynamics of participation and influence remains elusive. For instance, relying on number of followers to judge an influencer can be misleading [Cha et al. 2010; Bakshy et al. 2011] and predicting who is successful at an individual level was shown to be generally unreliable [Bakshy et al. 2011].

Our work takes a different starting point: We follow evidence that a large fraction of diffusion cascades occur close to a seed node [Goel, Watts, and Goldstein 2012]. Hence we focus on identifying those who contribute in adding *original* content in the network, and how this relates to temporal characteristics of the content being exchanged. Previous studies of temporal properties of diffusion typically focused on leveraging that those are short-lived [Cha et al. 2009; Rodriguez, Balduzzi, and Schölkopf 2011], or on using patterns in the time series for better classification [Kwon and Cha 2014; Kamath et al. 2013; Yang and Leskovec 2011].

These elites in information acquisition have been studied in very different contexts such as social learning [Bala and Goyal 1998; Acemoglu et al. 2011] and opinion formation [Golub and Jackson 2010; Acemoglu, Ozdaglar, and ParandehGheibi 2010], and even user generated content [Easley and Ghosh 2013; Ghosh and McAfee 2011]. Those results are different in spirit from ours as they typically focus on aggregation of multiple contributions on the same specific topic, either within a social networks or in the presence of a kernel of experts. For that reason, they typically assume specific types of information or interactions. Our model focuses on a simpler model in which information can be produced under some exerted effort, but is free to reproduce within a given network. The work motivated similarly to ours considers a similar process in an endogenous network where players may create new links at a fixed cost [Galeotti and Goyal 2010]. It was shown that these dynamics typically lead to extreme specialization, even among *ex ante* identical players. However, heterogeneous systems can't be analyzed in the same manner, and networks produced are typically very schematic (bi-partite). Our work proves that specialization emerges in an exogenous network, even without the reinforcing process of strategic link formation.

3.1 The existence of specialization

Unsurprisingly, in social media like Twitter, a small fraction of users are responsible for a large part of the activity. Many behaviors comprise online activity (*e.g.,* sharing something new, retweeting a link, commenting). We concentrate on understanding how knowledge

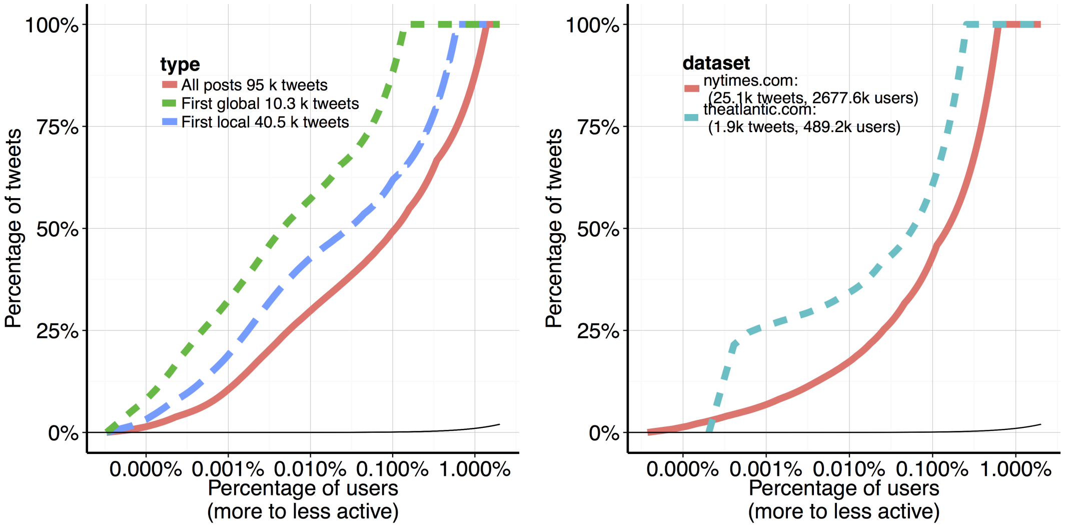

Figure 3.1: Lorenz curve (*i.e.,* cumulative share of the top x% nodes in the audience seen as a function of x) comparing (right) production of tweets and original content for `cnn.com` from KAIST-NEWS and (left) “first local tweets in two different domains.

sharing occurs. With that end in mind, we focus our studies on posts that share a link to some news article (the source of knowledge).

Imbalanced content creation

We quantify this concentration of activity by using the Lorenz curve [Lorenz 1905]. With the KAIST-NEWS dataset, we plot the cumulative share of the top x % of users as a function of x in Figure 3.1(right). Since some domains only cater to niche groups, the fraction x here is measured relative to the domain’s audience size (*i.e.,* anyone who received or sent at least one such URL).

A quick glance at the plot confirms that the size of passive and active audience differ by orders of magnitude (*e.g.,* as seen here and in other domains, 99% do not tweet a single URL. Equivalently, 1% of the audience produces almost all the new tweets in the network).

In addition to examining how users post in general (red solid line), we also look at how they acquire original information for the network. We, hence, looked at users who were the first on twitter to post a URL link (“global first” represented by the short green dotted line) and users who were the first in their local network, i.e. they did not receive the URL from anyone they followed before they sent the URL ("local first" represented by the long blue dashed line). Note that in each of these cases, the overall audience remains the same - those who have received the link either directly or indirectly from an originator. Here, in the left figure, 0.1% of the `cnn.com` audience produces half of all tweets. But the same number of people produce 60% of the globally original content and almost 90% of the locally original content. Perhaps unsurprisingly, while only a small minority of nodes re-post articles, it is an even smaller minority that introduces original content in the network.

Specialization is the phenomenon of users taking extreme positions - in our case, some users expend a lot of effort while others are on the other extreme of expending almost no effort. To help quantify this phenomenon, we introduce the 90%-volume originators measure defined as the fraction of the audience that together produce 90% of the volume.

Characterizing content originators

It has been shown (see, *e.g.,* [May et al. 2014]) that a user's tweeting activity is strongly correlated with their in- and out-degree. Intuitively, an active online presence is required to gather many followers. Having many followers encourages a return connection by other users. One hypothesis of a simple hierarchy of social media emerges: the content producers responsible for new content creation, the power users and intermediaries who drive the traffic and the passive consumers. As we see here, reality is at odds with this expectation when it comes to production of original content.

Figure 3.2 (left) presents, for users binned according to their activity on the x-axis, the distribution of the fraction of local first content they produce with median and various percentiles. To help interpretation, we represent qualitatively with a thin solid line the number of users in each bin, where the first bin contains approximately 129k users. On the right we observe the effects of a few heavy nodes: there are in total 90 users posting more than 400 URLs in a month, who are primarily either institutional accounts or professional

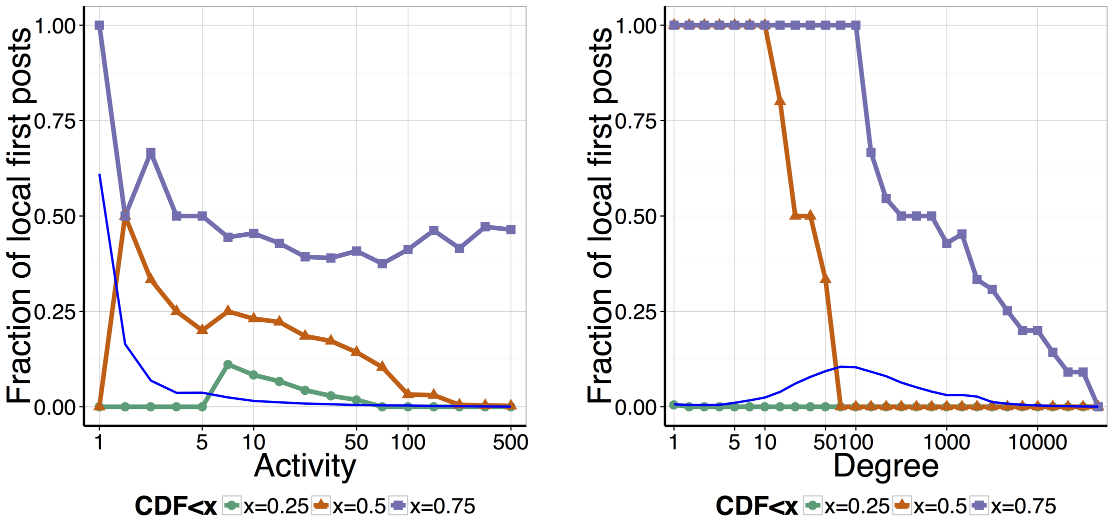

Figure 3.2: Fraction of locally original activity, presented as percentiles among users population binned according to (top) activity and (bottom) number of accounts they follow.

journalists and are almost always original. However, those are exceptions: among the active users, originators are generally a minority typically the 25% most original chosen across all activity levels. On the contrary, this trend proves that a URL is most likely to be locally original when it is posted by less active users. Equivalently, if the authors of that tweet post approximately 50 URLs in a month, it is likely to be one she has previously received. Another concurring observation, shown in Figure 3.2 (right), presents the same distribution where users are binned on the x-axis according to the number of people they follow. The trend here is even more pronounced as users belonging to the less connected half are much more likely to produce original information.

While, at first, this trend appears relatively surprising, the theory of public goods offers a simple explanation that we leverage later: that the effort exerted by others creates a disincentive for a well connected player to acquire new information. It seems in particular that 50% of users with larger than average degree rely entirely on the information they receive for their posts.

Effect of Time

Finally, we study the factors quantitatively affecting specialization. To take an example, first, we show in Figure 3.1(left) a comparison between the Lorenz curves for two news media domains: New York Times and The Atlantic. These are different in multiple ways: The New York Times is a daily newspaper with a very large readership while the Atlantic is a monthly magazine with a smaller readership. When comparing Lorenz curves, the Atlantic is more specialized than the New York Times with 0.4% of the audience accounting for 75% of `theatlantic.com` tweets while 0.8% of the audience accounts for 75% of `nytimes.com` tweets. This indicates that audiences of different sources specialize in different ways.

One metric for characterizing the difference is a notion of how long an article is expected to be relevant. We define the *shelf life* of an article to be the amount of time for which it is relevant *i.e.,* it continues to be shared among users. For every media, we measure its average *shelf life* by using the number of unique URLs produced over a month. This captures the fact that, since all media compete for attention within the same online network, one producing ten times more content expects the content to be renewed ten times faster.

domain	# unique	# users	# users	# posts	expiration time
	URLs	(who receive or post)	posting		estimate (min)
bbc.co.uk	19600	3252997	6248	113693	2.20
businessweek.com	777	622615	927	9405	55.60
cnn.com	10255	3026569	4458	94965	4.21
csmonitor.com	337	460161	492	3561	128.19
economist.com	232	802242	922	3714	186.21
forbes.com	1934	921576	1198	12375	22.34
foxnews.com	1529	510935	2845	19383	28.25
ft.com	6750	1373373	2647	28497	6.4
guardian.co.uk	5612	2106241	2294	45911	7.70
huffingtonpost.com	3742	1443562	2492	36974	11.54
mirror.co.uk	1306	638255	708	4863	33.08
news.yahoo.com	7684	1467238	5227	65734	5.62
newsweek.com	517	783171	679	3465	83.56
newyorker.com	299	754866	656	2444	144.48
npr.org	447	1220573	1066	12100	96.64
nytimes.com	5917	2677563	4085	111674	7.30
online.wsj.com	5077	1394111	2075	37581	8.51
reuters.com	16634	1435299	2621	61955	2.60
salon.com	803	1082391	745	4501	53.80
slate.com	518	676407	897	3097	83.40
theatlantic.com	5917	489222	804	4670	7.30
theonion.com	795	1427288	1238	11969	54.34
time.com	2293	530981	4370	14299	18.84
usatoday.com	3281	1141070	1570	20912	13.17
usnews.com	1089	580657	373	4222	39.67
vanityfair.com	162	598879	743	2261	266.67
washingtonpost.com	2886	1755915	2051	35554	14.97
wired.com	1751	1325465	2307	17640	24.67

Table 3.1: Expiration times of different news sources

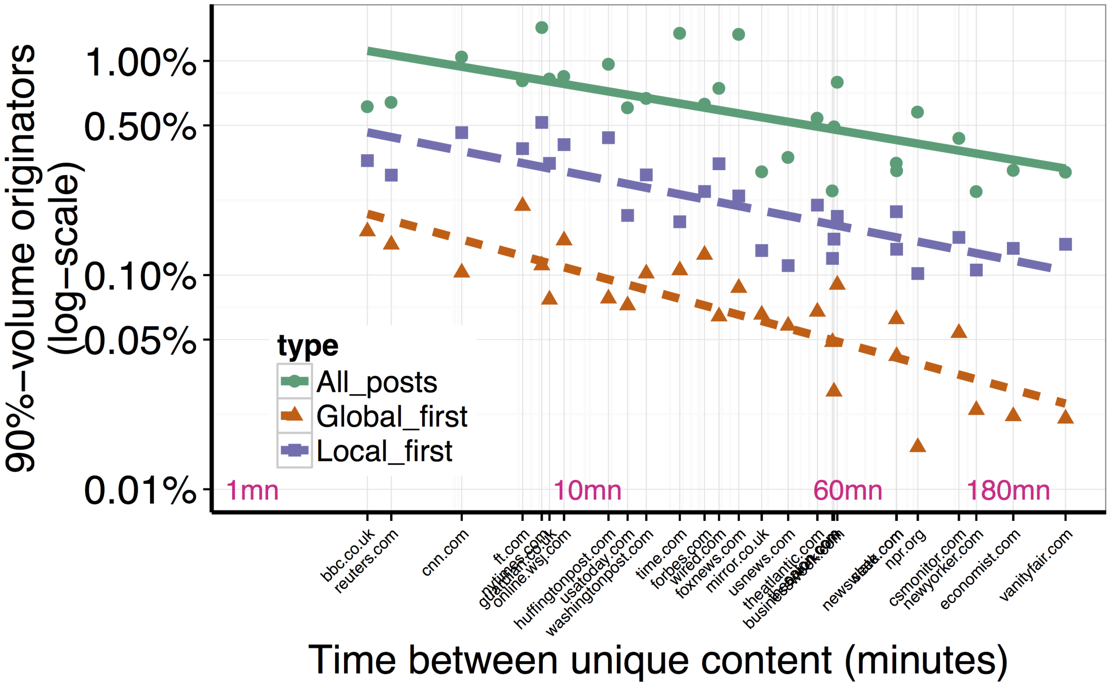

Figure 3.3: Concentration of sharing compared to the shelf-life for each media source. Each point is the fraction of the audience responsible for 90% of the tweet volume of the media source.

Our main observation is as follows: the degree of specialization is related to the temporal dynamics of the content, with remarkable regularity. For every media, we measure its average *shelf life* by using the number of unique URLs produced over a month. We define the shelf life of an article to be the amount of time for which it is relevant *i.e.,* it continues to be shared among users. Figure 3.3 shows the 90%-volume originator (*i.e.,* the percentage of the audience producing 90% of tweet volumes) for 31 media sources. There is a fairly large range of shelf life from approximately 2 minutes to over 2 hours. However, we consistently observe that domains with long shelf times tend involve a smaller fraction of the population to produce most of the content.

Note that these results hold for various definitions of content creation from introducing new content to the network to participating in the conversation around a shared article. In a medium like twitter, this conversation (the original tweet as well as all associated retweets

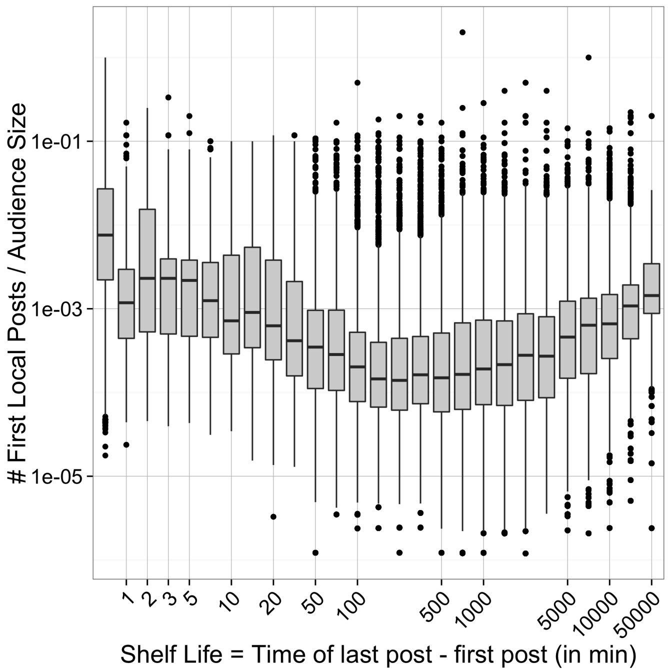

Figure 3.4: Concentration of sharing compared to the diffusion-life for each article. Each article's diffusion life is the total active time (in minutes) of the article.

and links) behave as the conversation compared to mediums like blogs where the comments serve the same purpose.

We also examined the effect of different measures of shelf lives in Figure 3.4. We calculate the diffusion life as the length of time that the article is shared (time of last post - time of first post). The y-axis is a measure of concentration, fraction of locally first posts of the total number of people receiving the article. We normalized by the number of users posting the article, in order to better account for larger cascades. Other measures of concentration, such as the fraction of first local posts by the total number of posts of an article, also exhibit similar trends, albeit in a more muted fashion. We continue to see the trend of articles with longer shelf lives tend to be more concentrated in sharing.

We have made several observations: (1) The presence of specialization where a small number of individuals are responsible for most of the original content produced on Twitter. (2) These individuals who produce most of the original content are not, as expected at first glance, the most well connected or the highest degree nodes. Rather, they are average-degree nodes in the network. (3) There is a correlation between the shelf life of an article, the time for which it is relevant, and the degree of specialization. In the following section, we present an idealized model which retains the flavor of the problem of information search.

3.2 Why does specialization occur?

While information diffusion on social media is complex and topic dependent, our goal in this section is to provide a simple model with which previous observations of information acquisition can be predicted. We leverage the economic theory of public goods – goods that are non-rivalrous where use by one individual does not reduce availability to others. In fact, in many public goods models, the ownership of the good by on individual has an impact on the utility of his neighbors. Further, we consider news as a perishable good, i.e. a good that needs to be used within a short period of time and bought again (such as milk or produce). While news does not spoil in the same sense as produce does, the value of news does decrease with time due to updated information and later events occurring. In both cases, since the product is short-lived and the demand is persistent, there is a time dynamic to renew it.

A Public Good Approach to Original Content Production

A individual, i, derives value from the content produced. Whether this is through her own efforts or her neighbors' efforts does not change the underlying value. In the general case, we assume this content has some value some convex function $f(y_i)$ where y_i represents the effort made by i. The user also has some convex cost, $c(y_i)$, associated with searching and finding the information.

There is a social component to the interaction: users make the results of their work available to neighbors in a social network graph. We denote the adjacency matrix of the social network as $G = (V, E)$ and it can either be undirected (*e.g.,* Facebook) or directed (*e.g.,* Twitter). The neighbors of i is denoted by $N(i)$. Without loss of generality, we assume that the effort of a user only affects its direct neighbors. The general case simply requires redefining neighboring relations to include future descendants.

The value to the user is thus a function of the available results of all the efforts ($y_i + y_{-i}$ where y_{-i} = $\sum_{j \in N(i)} y_j$). The total utility of a user is then defined as

$$U(y_i, y_{-i}) = f(y_i + y_{-i}) + c(y_i).$$

The above general utility is true of any networked public good. In the setting of information aqcuisition, the utility of information is represented as being in an informed state. In this state, a user has an additional unit of return compared to being uninformed. Upon a discovery, a user remains in the informed state for a time τ equal to the shelf time of this item. We assume τ is a constant.

As content online is vast and not easy to navigate, we assume that player i seeks knowledge at a given rate. This results in content being discovered by her at random times with an intensity y_i, forming a Poisson process of discovery times. The effort of that user to individually achieve a discovery rate y_i has a convex cost $c(y_i)$. This captures the fact that as more effort is exerted, or time is invested, worthwhile information becomes rare and harder to find.

Let us denote y_{-i} = $\sum_{j \in N(i)} y_j$ as the rate of content discovery that a user i in the network receives at no cost from her neighbors. Then, including her own effort cost $c(y_i)$, the average utility received per unit of time can be written as:

$$U(y_i, y_{-i}) = 1 - e^{-\tau(y_i + y_{-i})} - c(y_i) \ .$$

At time $t = T$, the probability to have received one content item within $]T - \tau; T]$ is the probability that a Poisson process of rate $(y_i + y_{-i})$ creates no point in that interval.

Note here, that discovering multiple content simultaneously creates no additional benefit to the user since the user is already in the informed state. Note also that having content items of various shelf-lives would result in the same dynamics as long as those durations are chosen independently of the discovery process. Finally, while most of the properties of the model we show generalizes to general convex cost, we are primarily interested in polynomial cost $(c : y_i \mapsto \frac{\theta}{\alpha+1} y_i^{\alpha+1}), \alpha > 0$. We can think of θ as the reference time period. A reward of 1 is equivalent to the effort spent to produce content once every θ time. In this work, we assume, in general, that the cost is normalized such that $\theta = 1\text{hr}$. This means that the reward exactly compensates for the search effort incurred to produce original content every hour. More general models, especially ones with heterogeneous costs and a matrix of benefits transfer between users, are likely to perfect realism of this model, but we leave them for future work.

Best Response We first analyze a single individual response of a player to her neighbors' efforts. Even with non-linear dynamics is non-linear, we can represent this best response action in a simple closed form.

Theorem 1. *For a node, i, of $G = (V, E)$, the best response to i's neighbors' efforts, y_{-i}, is given by*
$\phi(y_{-i}, \tau) = \frac{\alpha}{\tau} W(\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha} e^{-\frac{\tau y_{-i}}{\alpha}})$*, where W is the Lambert function defined on $[0; \infty[$ as the inverse of the function $x \mapsto x \exp(x)$.*

Proof. For an individual, i, their best response to their neighbors efforts occurs when i's utility is maximized w.r.t. the amount of effort i invests, y_i.

$$\max_{y_i} U(y_i, y_{-i}) \text{ s.t. } y_i \geq 0 \text{ , \textit{i.e.}, , } \frac{\partial U(y_i, y_{-i})}{\partial y_i} = 0 \text{ .}$$

This yields $\tau e^{-\tau(y_i + y_{-i})} - y_i^{\alpha} = 0$.
Hence $\tau y_i = \alpha W(\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha} e^{-\frac{\tau y_{-i}}{\alpha}})$ where W denotes the Lambert function, which proves the result. □

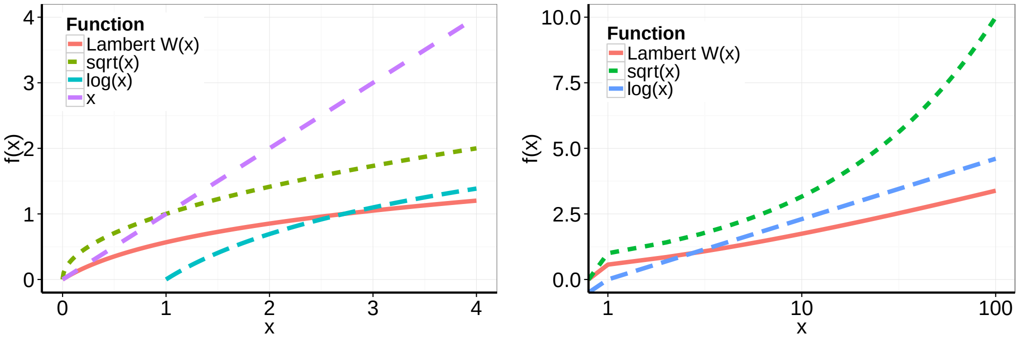

Figure 3.5: Comparison of the Lambert function ($W(z)$ where $z = W(z)e^{W(z)}$) to the common function of x, $\log(x)$, and $\sqrt{x}$ in the range (left) [0,4] and (right) [1,100].

The Lambert function W is a positive increasing function, that is asymptotically equivalent to the identity near 0 and comes within a negligible distance of the function $x \mapsto \ln(x) - \ln\ln(x)$ as x becomes large. The last two decades has found numerous applications of this function to differential equation, combinatorics, theoretical physics and others. Its computation, both through formal calculus and numerical approximation can be done fast.

Our closed form implies the bound for any y : $0 = \lim_{x\to\infty} \phi(x) \leq \phi(y) \leq \phi(0) = \frac{\alpha}{\tau} W(\tau^{\frac{\alpha+1}{\alpha}})$.

Nash Equilibrium We initially focus on analyzing the Nash equilibrium in symmetric graphs.

Definition 1. *A graph G is symmetric if, given any two pairs of edges (u_1, v_1) and (u_2, v_2) of G, there is an automorphism $f : V(G) \to V(G)$ such that $f(u_1) = u_2$ and $f(v_1) = v_2$.*

In a symmetric graph, in a unique Nash Equilibrium, all nodes exert the same amount of effort. Observe that if this were not the case, a transformation of the graph results in another equilibrium.

Lemma 2. *For a D-regular graph, a symmetric Nash Equilibrium always exists and is* *given by*

$$y_i = \frac{\alpha}{\tau(1+D)} W(\tau^{\frac{\alpha+1}{\alpha}} \frac{(1+D)}{\alpha}), \forall i.$$

Proof. In a symmetric equilibrium, $y_i = y, \forall i \in G$. Also, for a node i, $y_{-i} = Dy$.

$$\begin{aligned} \text{At equilibrium} \quad y &= \frac{\alpha}{\tau} W(\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha} e^{-\frac{\tau D y}{\alpha}}) \\ \frac{\tau}{\alpha} y e^{\frac{\tau}{\alpha} y} &= \frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha} e^{-\frac{\tau D y}{\alpha}} \\ y \frac{\tau}{\alpha}(1+D) &= W(\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha}(1+D)) \\ y &= \frac{\alpha}{\tau(1+D)} W(\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha}(1+D)) \end{aligned}$$

□

The case of symmetric graphs is interesting because, as we show in Section 3.2.1, this symmetric equilibrium need not always be a unique or stable equilibrium.

3.2.1 Model Validation

Real world graphs are, of course, more complex than the above symmetric graph models. We validate our model on a subset of the NYT graph (a random sample of 10% of the edges). We use an iterative update method (Algorithm 2) to find the Nash equilibrium numerically. In these simulations, we used a range of shelf-life times ranging from short ($\tau = 1$) to long ($\tau = 1000$).

Matching our observations from the KAIST-NEWS dataset, users with larger degree have less “information seeking activity”. This is reflected in a smaller amount of effort spent in the Nash Equilibrium. Figure 3.6 (left) shows the correlation of the Nash Equilibrium effort with out-degree of a node ($\tau = 0.5$ on a sample of 0.1% of the NYT graph). Here, we see a very strong relationship between the degree and the amount of effort ex-

ALGORITHM 2: Finding the Nash equilibrium for our system by iterative update, in the instance when a simple closed form does not exist.

```
Input: Graph $G = (V, E)$, Expiration-time $\tau$, initialization value (optional)
Output: The amount of effort $\mathbf{y}^* = y_1^*, ..y_I^*, ..y_N^*$
y = 0; repeat
    for each node $i \in V$ do
        $y_{-i} = \sum_{j \in N(i)} y_j$;
        bestResponse $= \frac{1}{\tau} W(\tau^2 e^{-\tau y_{-i}})$;
        $\delta = abs(y_i -$bestResponse$)$;
        $y_i =$bestResponse;
    end
until $\delta > 0$;
```

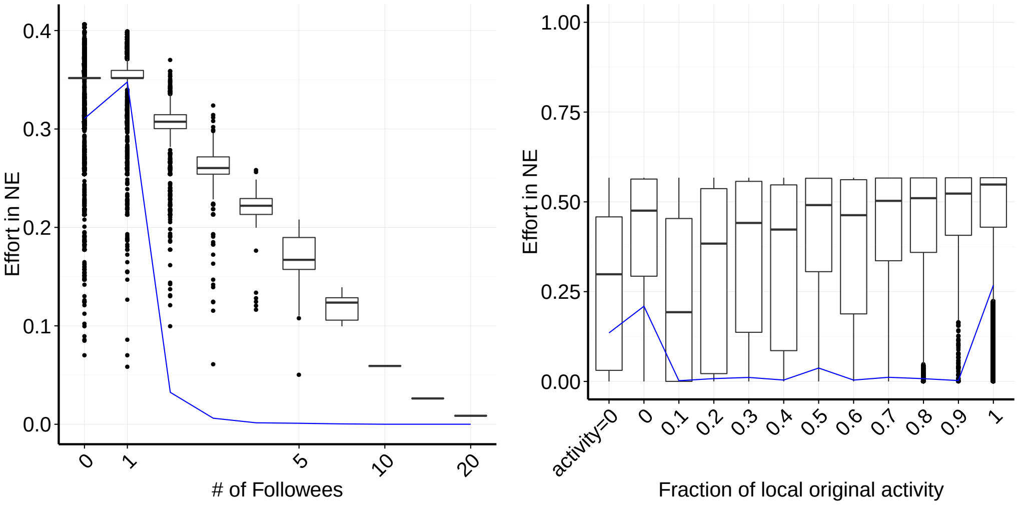

Figure 3.6: The Nash Equilibrium (as a function of (left) node degree and (right) fraction of first local activity) in a sample of the NYT graph

pended in the Nash Equilibrium. Thus, our model yields predictive power for relation of connection and investment in information search

We then observe that the elite in the modeled equilibrium share similar structure to those observed empirically (Figure 3.7). A small subset of individuals are responsible for a large fraction of the effort spent – mimicking the behavior of individuals with original content.

Lastly, we examine how the effort in the Nash equilibrium of our model correlates to the fraction of local original activity vs total activity observed in the NYTimes dataset

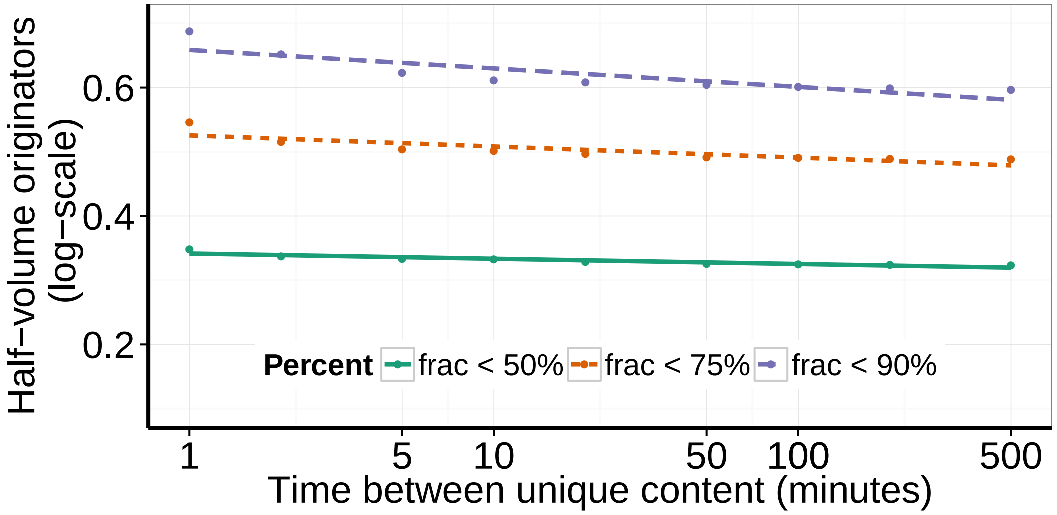

Figure 3.7: Proportion of population responsible for 50%, 75% and 90% of the effort in the Nash Equilibrium in sample of NYT graph.

(Figure 3.6 right). Ideally, we would expect to see perfect correlation since the effort in our model captures exactly this, the effort you spent to bring new content to your neighbors. We see that individuals who in the real world had no effort (the left most group) expend low effort in the Nash Equilibrium. Those who posted at least one article expended more effort and the amount of effort steadily rises.

Conditions for a Unique Nash Equilibrium Different classes of goods exhibit different types of behavior. In economic theory, one of these classifications are that of a *normal good* is a good for which demand increases with increased wealth. Mathematically, if $\gamma : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ is a differentiable function representing the income elasticity of demand (the responsiveness of the demand to a change in the income), then the good is normal iff the derivative satisfies $0 < \gamma' < 1$. A *network normal good* carries that idea to a networked case where there is a income elasticity of demand function for each player i in the network. The consumption γ_i is defined in terms of the wealth of i (set externally), w_i, and i's "social income", the income from neighbors of i, y_{-i}. A network normal good satisfies the condition: $1 + \frac{1}{\lambda_{\min}} < \gamma_i'(w_i + y_{-i}) < 1$ [Allouch 2015]. We can also express these conditions in terms of the best response $\phi(y_{-i}) = \gamma_i(w_i + y_{-i}) - w_i$ as follows.

Fact. In the above notation, a good is network normal iff for every player i, $\frac{1}{\lambda_{\min}} < \phi'(y_{-i}) < 0$.

In our model, there can exist multiple equilibria for the effort that individuals expend. Using network normality conditions, we now give a condition involving the expiration time parameter, τ under which the Nash equilibrium for the system will be unique.

Lemma 3. $\frac{\partial \phi}{\partial y} = -\frac{W(\tau^2 e^{-\tau y})}{1+W(\tau^2 e^{-\tau y})}$

Proof.

$$\begin{aligned}
\frac{\partial \phi}{\partial y} &= \frac{\partial\big(\frac{1}{\tau}W(\tau^2 e^{-\tau y})\big)}{\partial y} \\
&= \frac{1}{\tau} \cdot \tau^2 \cdot (-\tau) \cdot e^{-\tau y} \cdot W'(\tau^2 e^{-\tau y}) \\
&= -\tau^2 e^{-\tau y} \cdot \frac{W(\tau^2 e^{-\tau y})}{\tau^2 e^{-\tau y}(1 + W(\tau^2 e^{-\tau y}))} \\
&= -\frac{W(\tau^2 e^{-\tau y})}{1 + W(\tau^2 e^{-\tau y})}
\end{aligned}$$

□

Theorem 4 (Short-Lived Content Exhibits Less Specialization). *Let $\lambda_{\min}$ be the minimum eigenvalue of the adjacency matrix of the network, $G = (V, E)$, and let τ be the expiration time parameter of the system. Then, a unique Nash Equilibrium exists if*

$$\tau < \hat{\tau} =^{def} \big(\frac{\alpha}{-\lambda_{\min} - 1}\big)^{\frac{\alpha}{\alpha+1}} e^{\frac{\alpha}{(\alpha+1)(-\lambda_{\min}-1)}} .$$

Proof. We will prove the theorem by using the previously established connection between network normality of the system and the existence of a unique Nash equilibrium [Allouch 2015; Bramoullé and Kranton 2006; Bramoullé et al. 2014; Bergstrom, Blume, and Varian 1986]. Hence we only need to show that the network normal conditions hold under the assumptions of the theorem.

We will show that the condition holds for every player, i. For ease of notation, let $\phi = \phi_i$ and $x = y_{-i}$.

Observe that since W is an increasing function, we have $\phi'(x)$ is a non-decreasing function. Hence the derivative only takes values in $[\phi'(0), \lim_{x\to\infty} \phi'(x)] = [-\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha} W'(\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha}), 0]$. Now, the network normality condition simplifies to verifying

$$\frac{1}{\lambda_{\min}(G)} < -\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha} W'(\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha}) < 0.$$

Simplifying the first inequality, we get:

$$\tau < (\frac{\alpha}{-\lambda_{\min} - 1})^{\frac{\alpha}{\alpha+1}} e^{\frac{\alpha}{(\alpha+1)(-\lambda_{\min}-1)}} = \hat{\tau}$$

Thus, the network normality conditions holds and a unique Nash equilibrium exists for any $\tau < \hat{\tau}$. □

Here, we show the conditions necessary for a unique Nash Equilibrium to exist for various graph families. Let $\lambda_{\min}$ be the minimum eigenvalue of the adjacency matrix of the network, $G = (V, E)$, and let τ be the expiration time parameter of the system. Then, a unique Nash Equilibrium exists if

$$\tau < \hat{\tau} \stackrel{\text{def}}{=} (\frac{1}{-\lambda_{\min} - 1})^{\frac{\alpha}{\alpha+1}} e^{\frac{\alpha}{(\alpha+1)(-\lambda_{\min}-1)}} .$$

Lemma 5. *A complete graph always has a unique Nash equilibrium*

Proof. In a complete graph, $\lambda_{\min} = -1$. Thus, for any value of τ, there exists a unique Nash equilibrium. □

Lemma 6. *In a star graph with $n-1$ leaf nodes, a unique Nash equilibrium for $\tau < \hat{\tau} = (\frac{1}{\sqrt{n-1}-1})^{\frac{1}{2}} e^{\frac{1}{2(\sqrt{n-1}-1)}}$,* *Proof.* In a star graph of size n, $\lambda_{\min} = -\sqrt{n}$ ([Brouwer and Haemers 2012]). $\therefore W(\tau^2) < \frac{1}{\sqrt{n}-1} \implies \tau^2 < \frac{1}{\sqrt{n}-1} e^{\frac{1}{\sqrt{n}-1}}$ □

Lemma 7. *An even cycle graph of size n has a unique Nash equilibrium for $\tau < \hat{\tau} = \sqrt{e}$.*

Proof. An even cycle has $\lambda_{\min} = -2$ [Brouwer and Haemers 2012]. $\therefore W(\tau^2) < 1 \implies \tau = \sqrt{e}$. □

Lemma 8. *An odd cycle graph of size n has a unique Nash equilibrium for $\tau < \hat{\tau} = \frac{n}{(n^2-\pi^2)^{\frac{1}{2}}} e^{\frac{n^2}{2(n^2-\pi^2)}}$.*

Proof.

$$
\begin{aligned}
\lambda &= 2\cos\frac{2\pi j}{n} \quad \text{([Brouwer and Haemers 2012])} && (j = 0, 1, ..., n-1) \\
\lambda_{\min} &= 2 \cdot \cos{(\pi - \frac{\pi}{n})} && (\text{for } j = \frac{n-1}{2}) \\
\lambda_{\min} &= -2 \cdot \cos\frac{\pi}{n} && (\because \cos(\pi - \theta) = -\cos(\theta)) \\
\lambda_{\min} &= -2 \cdot \sum_{n=0}^{\infty} (-1)^n \frac{\frac{\pi}{n}^{2n}}{(2n)!} && (\text{Taylor expansion}) \\
\lambda_{\min} &\approx -2 \cdot (1 - \frac{\pi^2}{2!n^2} + \frac{\pi^4}{4!n^4}) \\
\lambda_{\min} &\approx -2 + \frac{\pi^2}{n^2}
\end{aligned}
$$

Substituting the value for $\lambda_{\min}$,

$$
\begin{aligned}
W(\tau^2) &< \frac{1}{1 - \frac{\pi^2}{n^2}} \\
\tau^2 &< \frac{n^2}{n^2 - \pi^2} e^{\frac{n^2}{n^2-\pi^2}} \\
\tau &< \frac{n}{(n^2 - \pi^2)^{\frac{1}{2}}} e^{\frac{n^2}{2(n^2-\pi^2)}}
\end{aligned}
$$

□

Lemma 9. *An Erdös-Renyi graph with constant p has a unique Nash equilibrium for $\tau < \hat{\tau} = (\frac{1}{2\sqrt{np}-1})^{\frac{1}{2}} e^{\frac{1}{2(2\sqrt{np}-1)}}$.*

Proof. For a Erdös-Renyi graph, with constant p ([Füredi and Komlós 1981])

$$\begin{aligned}
\lambda_{\min} &= -c\sqrt{n} \\
\max|\lambda_{\min}| &= 2\sigma\sqrt{n} + O(n^{\frac{1}{3}}\log n) \qquad\qquad \text{where } \sigma = \sqrt{p} \\
\lambda_{\min} &> -2\sqrt{np}
\end{aligned}$$

Substituting the value for $\lambda_{\min}$,

$$\begin{aligned}
W(\tau^2) &< \frac{1}{2\sqrt{np}-1} \\
\tau^2 &< \frac{1}{2\sqrt{np}-1} e^{\frac{1}{2\sqrt{np}-1}}
\end{aligned}$$

□

Lemma 10. *A complete bipartite graph of size n has a unique Nash equilibrium for $\tau < \hat{\tau} = (\frac{2}{n-2})^{\frac{1}{2}} e^{\frac{1}{n-2}}$.*

Proof. The minimum eigenvalue for a complete bipartite graph is given by $\lambda_{\min} = -\frac{n}{2}$. Thus $W(\tau^2) < \frac{2}{n-2} \implies \tau < (\frac{2}{n-2})^{\frac{1}{2}} e^{\frac{1}{n-2}}$ □

The quantity $\hat{\tau}$ of G specifies the condition under which a unique Nash equilibrium exists. Table 3.2 details the value of $\hat{\tau}$ for various regular graphs ([Ramachandran and Chaintreau 2015b]).

Our observations on simple regular graphs give us an understanding of the behavior of the Nash Equilibrium in different types of settings. We see that for shorter lived information (content with smaller τ), the process of sharing is relatively straightforward. In most graphs, for small $\tau < \hat{\tau}$, there exists a unique equilibrium. In symmetric graphs, this equilibrium is symmetric. In non-regular graphs, the equilibrium response is inversely related

Table 3.2: Conditions for unique Nash Equilibrium ($\tau < \hat{\tau}$) for graphs with n nodes ($\alpha = 1$)

Graph	$\lambda_{\min}$	$\hat{\tau}$
Complete	-1	$\forall\tau\ (\infty)$
Cycle (Even)	-2	$\sqrt{e}$
Cycle (Odd)	$-2 + \frac{\pi^2}{n^2}$	$\frac{n}{(n^2-\pi^2)^{\frac{1}{2}}} e^{\frac{n^2}{2(n^2-\pi^2)}}$
Erdös-Renyi	$-2\sqrt{np}$	$(\frac{1}{2\sqrt{np}-1})^{\frac{1}{2}} e^{\frac{1}{2(2\sqrt{np}-1)}}$
Star	$-\sqrt{n-1}$	$(\frac{1}{\sqrt{n-1}-1})^{\frac{1}{2}} e^{\frac{1}{2(\sqrt{n-1}-1)}}$
Complete Bipartite	$-\frac{n}{2}$	$(\frac{2}{n-2})^{\frac{1}{2}} e^{\frac{1}{n-2}}$

to the degree of a node since higher degree nodes can rely on good quality content through their many neighbors. Conversely, lower degree nodes tend to expend more effort since they have few neighbors that they can free ride on.

In general, more balanced graphs (with larger $\lambda_{\min}$) have less sensitivity to the ephemeral nature of information *i.e.,* the conditions for a unique equilibrium encompass a larger range of shelf life values. In more segregated graphs (with smaller $\lambda_{\min}$), the efforts of a few people can be enough for the graph as a whole and the equilibrium is less balanced in nature.

Understanding the dependencies of the equilibrium in real world graphs is a little more challenging. Since these are not d-regular graphs, we do not expect symmetric equilibria to occur. In the case of the real world NYTimes graph, $\lambda_{\min} \approx -70$ (computed with python's sparse matrix package). Considering that the size of the NYTimes graph is $n = 346k$ users, this case more closely resembles a balanced graph, like an Erdös-Renyi graph. For $\alpha = 1$, a case where there is a relatively low cost of finding information, $\hat{\tau} \approx 0.12$ of the reference time period. For $\theta = 1$hr (*i.e.,* ., assuming readers' utility for content roughly compensate an effort to search every hour for new information), $\hat{\tau} \approx 7$min which is close to the empirically estimated shelf life of $\tau = 7.30$ min.

Tuning Shelf Life to Maximize Original Information A media source would want to encourage users to spend more time on their site. Thus, they might be interested in tuning their parameter to maximize user effort. In a disconnected setting, each person is responsible for finding and consuming their own content. In this case, $y_{-i} = 0$ and the best response simplifies to $\phi(0) = \frac{\alpha}{\tau} W(\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha})$. At the value $\tau = \tau^*$, an individual is incentivitized to expend maximal effort.

Claim 11. *For an isolated node, i, the effort is maximized at $\tau^* = e^{\frac{1}{\alpha+1}}$.*

Proof. The τ that corresponds to the maximum effort satisfies $\frac{\partial \phi}{\partial \tau} = 0$. Further, since i is isolated, $y_{-i} = 0$. Hence,

$$\begin{aligned}
\frac{\partial \phi}{\partial \tau} &= \frac{\partial \frac{1}{\tau} W(\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha})}{\partial \tau} = 0 \\
\frac{\alpha}{\tau} \cdot \frac{1}{\alpha} \cdot \frac{\alpha+1}{\alpha} \tau^{\frac{1}{\alpha}} \cdot W'(\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha}) + W(\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha}) \cdot (-1) \frac{\alpha}{\tau^2} &= 0 \\
\text{Simplifying,} \quad W(\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha}) &= \frac{1}{\alpha} \\
\tau &= e^{\frac{1}{\alpha+1}}
\end{aligned}$$

It is easy to verify that this critical point is a maxima. □

In the case of symmetric graphs, there is always a symmetric equilibrium (Lemma 2). We can calculate, for symmetric graphs, the τ^* that maximizes the amount of effort by any node in a symmetric equilibrium.

Claim 12. *For an symmetric graph of degree D, the effort in a symmetric equilibrium, y_i, is maximized at $\tau^* = \frac{e}{(1+D)^\alpha}^{\frac{1}{\alpha+1}}$*

Proof. Note that $y_{-i} = \frac{\alpha}{\tau(1+D)} W(\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha}(1+D))$ since i has degree D and the equilibrium is symmetric. Again, the τ that corresponds to the maximum effort satisfies $\frac{\partial \phi}{\partial \tau} = 0$. By evaluating these expressions, we get

$$W(\frac{\tau^{\frac{\alpha+1}{\alpha}}}{\alpha}(1+D)) = 1 \implies \tau = \frac{e}{(1+D)^\alpha}^{\frac{1}{\alpha+1}}$$

□

Specialization and Symmetry We use simulations to examine how these theoretical results translate to various graph families. For each graph family, we look at graphs of sizes ranging from $n = 4$ to $n = 400$ and edge density from $p = 0.0001$ to $p = 0.5$ (for Erdös-Renyi graphs). We then run an iterative algorithm that updates the best response until convergence [Ramachandran and Chaintreau 2015b] . The point of convergence (when it converges) is the Nash equilibrium. In the cases that we examined, the best responses converged to an equilibrium within 20 steps (though our algorithm does not guarantee convergence).

Considering, first, the case of symmetric graphs (figure 3.8), each line in the graph is the effort made by a particular node. Note that since many nodes have the same effort across different regimes of τ, those lines overlapping each other and are hence not visible. In both the bipartite and cycle graph, in the specialized equilibrium, half the nodes overlap and expend most of the effort and the remaining half free-ride on those nodes. We see that, with shorter shelf-lives, individuals are more self-reliant. Conversely, longer shelf lives result in individuals relying on others efforts. Both cycle graphs and complete bipartite graphs exhibit the property that when content is long-term, the equilibria becomes more specialized with some individuals doing the majority of the work and others doing almost no work. Bipartite graphs split into their two partitions where those in one partition do all the work while those in the other do none.

The story is more complex in the case on asymmetric graphs (figure 3.9). In each of the cases, we see a specialized equilibrium emerge. We consider the case of a star graph and an Erdös-Renyi graph, which gives us simple cases without the effect of heterogeneity. We also looked at a 10% subset of a real world graph. In the case of the star graph, the single central node does almost no work while all of his neighbors overlap and have much higher effort.

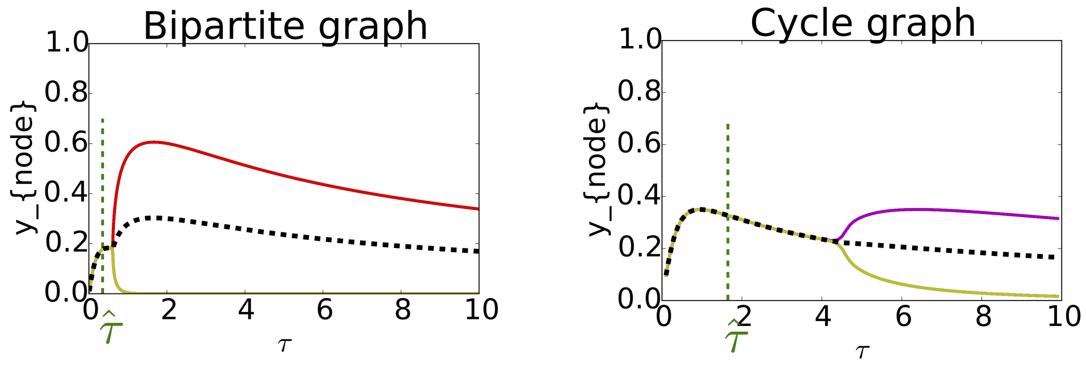

Figure 3.8: Differing effort levels in the Nash Equilibrium (y-axis) with different τ (x-axis) in symmetric graphs. Each node (of $n = 20$ nodes) is represented by a line in the figure. The unique equilibrium ($\tau < \hat{\tau}$) is always symmetric. (left) Complete bipartite graph (right) Cycle graph.

We see that specialization can occur as a result of the degree distribution (as in asymmetric graphs). However, this also occurs in symmetric graphs, when all nodes have the same degree. From lemma 2, we know that a symmetric Nash equilibrium exists, but we observe that the system converges to a specialized Nash equilibrium. In the following section, we show that symmetric equilibria are not stable for large τ.

3.2.2 Theoretical Proof of Specialization

When a unique Nash equilibrium exists, we understand the convergent network configuration. However, when there are multiple equilibria, it is not clear which of these configurations are realized — for instance, some of these Nash equilibria can be unstable and, hence, never realized in practice. Here, we use the same definition of stability as in [Bramoullé and Kranton 2006; Bramoullé et al. 2014]. A Nash equilibrium is stable if a small change in the strategy of one player leads to a situation where two conditions hold: (i) the player who did not change has no better strategy in the new circumstance (ii) the player who did change is now playing with a strictly worse strategy.

Empirically, we observe that for longer-term content, the equilibrium for a cycle graph

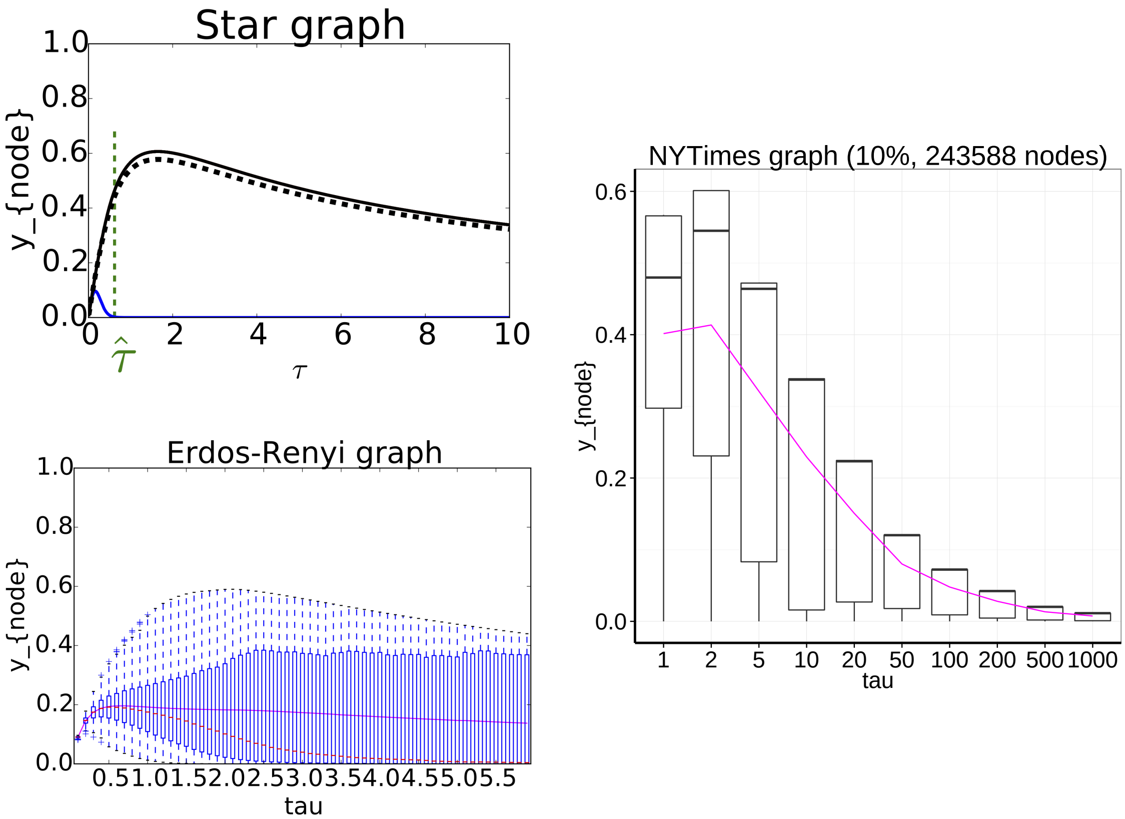

Figure 3.9: Differing effort levels with different τ in asymmetric graphs. Each bar represents the distribution of the amount of effort by all the nodes. The pink line is the average effort of all the nodes. (top-left) Star graph. (bottom-left) Erdös-Renyi graph ($n = 1000, p = 0.01$). (right) Randomly sampled NYTimes graph with 243k nodes.

and a bipartite graph are specialized (figure 3.8), *in spite* of them being symmetric graphs. This indicates that the stability of the Nash equilibrium has some dependency on τ.

Theorem 13 (Specialization for Longer Shelf-Life)**.** *There exists an shelf-life τ, such that, for any symmetric graph G of degree $D \geq 3$, the symmetric equilibrium is not stable.*

Proof. The proof follows the outline of the Proof of Theorem 2 in [Bramoullé and Kranton 2006]. It has two steps. The first step is a simple observation: If $\vec{y} < \vec{y'}$, then $\phi \circ \phi(\vec{y}) < \phi \circ \phi(\vec{y'})$. This follows because the response function $\phi(y)$ is a decreasing function of y.

The second step is to show that under some small perturbation $\vec{\epsilon} > 0$, we have $\phi \circ \phi(\vec{y} + \vec{\epsilon}) > \vec{y} + \vec{\epsilon}$(here the vector inequality $\vec{x} > 0$ corresponds to coordinate wise inequality $x_i >$

$0\,\forall i$). In other words, with any small change from the equilibrium, the best response moves further away (strictly) from the equilibrium. This shows that the equilibrium is not stable in the sense of [Bramoullé and Kranton 2006; Bramoullé et al. 2014]. For simplicity's sake, we consider only a quadratic cost function.

Let $\tilde{\mathbf{y}}$ be the symmetric equilibrium in the symmetric graph of degree D. Then, $\tilde{y}_i = \tilde{y}, \forall i$. Note that $\tilde{y} = \phi(\tilde{y})$ because it is an equilibrium. Here, we perturb all the responses by some $\epsilon > 0$

$$\begin{aligned}\phi(\vec{y} + \vec{\epsilon}) &= \phi(\vec{y}) + \nabla\phi \cdot \vec{\epsilon}\\ \phi_i(\vec{y} + \vec{\epsilon}) &= \phi_i(\vec{y}) + D\frac{\partial \phi_i}{\partial y_j}\epsilon \quad \text{for some } j \in N(i)\end{aligned}$$

since $\frac{\partial \phi_i}{\partial y_j} = 0 \quad$ if $j \notin N(i)$ and equal otherwise. Similarly,

$$\begin{aligned}\phi_i \circ \phi(\vec{y} + \vec{\epsilon}) &= \phi_i([\ldots, \phi_j(\vec{y} + \vec{\epsilon}), \ldots])\\ &= y_i + D^2\big(\frac{\partial \phi_i}{\partial y_j}\big)^2\epsilon \qquad \text{any } j \in N(i)\end{aligned}$$

To show that the symmetric equilibrium is not stable, we need

$$\begin{aligned}y_i + D^2\big(\frac{-W(\tau^2 e^{-\tau\tilde{y}})}{1 + W(\tau^2 e^{-\tau\tilde{y}})}\big)^2\epsilon &> y_i + \epsilon\\ W(\tau^2 e^{-\tau\tilde{y}}) &> \frac{1}{D-1}\end{aligned}$$

In other words, we want $\tau^2 e^{-\tau\tilde{y}} > \frac{1}{D-1}e^{\frac{1}{D-1}}$. Substituting for $\tilde{y}$ (lemma 2) and simplifying, we get that the symmetric Nash equilibrium is not stable when

$$2\ln\tau - \frac{1}{(D+1)}W(\tau^2(D+1)) > -\ln(D-1) + \frac{1}{D-1}.$$

Setting τ to be a constant (*e.g.,* $\tau = 10$), one only needs to verify that the following holds: $W(D+1) < (D+1)(\ln(D-1) + 2\ln\tau) - \frac{D+1}{D-1}$, which is true for $D > 3$. $\square$

3.3 Conclusion

Knowledge sharing has been greatly facilitated by social network services. Increasingly, it affects businesses, political debates and public services. Yet, after years of measurements, the structure of online diffusion remains complex and was shown to vary across media and topics. Our results identify, how the shelf life of information affects its diffusion. This leads to various types of specialization that can all be described in the unifying theory of public good.

While we empirically observe a remarkable match to the theoretical predictions on a qualitative level, we would like to point out that the current model of public good we introduce is highly idealized, especially as it assumes homogeneous cost of information acquisition. Proving that specialization occurs even in such symmetric cases is, in a sense, a worst-case result. In reality, several other factors contribute to users exerting higher effort in information acquisition including enjoyment [Feick and Price 1987], which typically varies across users depending on topics. However, our results generalize to heterogeneous perishable public goods to predict, for instance, that a single equilibrium exists whenever shelf life is sufficiently small. The qualitative effect of shelf life should also remain since our empirical observations prove it, even in a large number of very different mass media sources. We do, however, observe some amount of variance within this trend and accounting for other previously identified factors to predict span of content diffusion more accurately seems a promising direction.

Limitations & Applications Our model considers the case of user decisions based on one feature of content (the shelf life) but there are other that they might deem relevant (*e.g.,* topic, length of article, source) resulting in heterogeneous utility functions. This would result in more complex equilibria but the broad public goods results of more connected network structures resulting in better distribution of effort should still apply (assuming that these utility functions are convex).

In our public goods modeling, as with any other economics model, we assume rational behavior. An interesting setting to study would be to understand when a fraction of the agents are potentially irrational. In this case, we would conjecture that the overall equilibria results would depend on how the irrational agents are connected to the rest of the graph. For instance, if they essentially form a clique, then their behavior would not impact the majority of the users.

Another unpredictable player in social networks is bots that behave in seemingly irrational ways. In this case we could potentially model them as either having a different utility function or as exhibiting irrationality.

From the perspective of a network designer or social media platform, our results indicate potential avenues to encourage more participation from its users via the network structure. Public goods results indicate that certain types of networks are better suited for increased user engagement. Facebook, for instance, could give suggestions of people to follow based on this principle, potentially resulting in better connected and less partisan networks.

In most online networks, the networks themselves are not static but rather grow and change over time. In the next chapter, we study how this evolving structure affects information production and user choice.

Chapter 4

Biases in Network Evolution

From the previous chapter, we know that the existence of specialization is dependent on the structure of the graph – graphs which are more hierarchical tend to also have more specialized equilibria. In this chapter, we focus on understanding the dynamics of specialization in an evolving graph with network interactions. We use our model to extend the idea of 'wisdom of the crowd.'

One of their most dramatic consequences of the scale of information sharing on social media is the deluge of information we consult before any of our life's decisions: Word of mouth, electronically delivered, affect where we apply for jobs, who we support for political office, and our most mundane choices over a dinner plan or our next online purchase. Behind each choice lies a belief in the wisdom of the crowd observed in numerous instances: from Galton's original experiment on bull weight-judging [Galton 1907], to recent online applications like collaborative encyclopedia [Kittur et al. 2007], question answering [Wang et al. 2013] or prediction games [Goldstein, McAfee, and Suri 2014]. Formally, the wisdom of the crowd is said to emerge when an expanding social network connects each of us to an increasing number of contacts - or equivalently to a growing collection of information - and it enables everyone to come to an estimation that has quasi perfect precision.

In this chapter, we analyze for the first time how the structure of a social network affects the benefit of information sharing between users who are parsimonious in the information they share. Our goal is to understand the following question: "Which types of network's growth and evolution guarantees everyone to eventually benefit from information sharing?"

We now present the following contributions:

- We introduce a simple model where each participant of a social network attempts to estimate a value with best possible accuracy, using her own effort as well as sharing information with her contacts.
- In order to study the effect of topologies on the above general case, we simulate how nodes' estimation accuracy vary in our model when contact lists follow some real word evolution. Our empirical results highlight the complex interaction of information sharing: First, when the network expands and hence more information get shared, we find that a majority of nodes *suffer* on the short term, seeing diminished accuracy and more individual effort. Second, nodes tend to compensate their losses and they benefit from network expansion overall, but this typically require network to double or quadruple in size before a majority benefit. Thirdly, as expected, the benefits from information sharing at anytime are very uneven. Even when the network size is multiplied by 100, only a small minority see substantial gain. Those are invariably nodes who appeared earlier and are more connected, new arrivals and other nodes with smaller degrees benefit much less. (Section 4.2).
- Our theoretical analysis further demonstrates the connection between large unbalanced hierarchies and the failure of information sharing to benefit everyone. We prove that a large class of social networks exhibit an even more advantageous result: A stronger version of the above vision which we call the "wisdom of parsimonious crowds". However, that result is sensitive to network evolution and fails to emerge in many models of expanding social networks, including, as we prove, those with large segregated hierarchies. (Section 4.3).

4.1 Privacy Model

4.1.1 Model Overview

Let us introduce a simple generic *collaborative estimation* task. We assume N nodes aim at assessing the same objective mean value of a variable from a set of samples or evaluation that each of them possesses. We make the usual assumption that the sample of each participant is a noisy observation and that samples from different nodes are independent variables; their common mean is precisely the value that each participant aims at estimating.

Participants typically communicate with each other once their observation is made. We denote by $N(i)$ the list of i's contacts, which may represent friends, or alternatively members of various clubs and social groups in which information relevant to the estimation are shared. Information is then shared in a social graph $G(N, E)$ where edges are symmetric, which follows the social etiquette that during information sharing everyone share their experience within the relevant group. As in many previous works, we will be interested by sequence of expanding graphs, which grow to expose each user to an ever growing amount of information.

Now comes the most specific aspect of this estimation model. Instead of assuming that each individual receive a sample drawn from a fixed exogenous noise model, we assume that participants are *parsimonious*. Motivated by various situations below, we assume that i can individually produce an estimate with quality $\lambda_i > 0$ for a cost following a non-decreasing and convex function $c(\lambda_i)$. That estimate is then shared with all of i's contacts during information sharing. Ultimately, i is able to aggregate all estimates from either her or someone in her contact list $N(i)$ to obtain a more accurate estimation, with overall quality $\zeta_i \geq \lambda_i$. To model her incentive towards more accurate estimation, we assume that i pays afterwards a non-increasing *estimation cost* $G(\zeta_i)$.

Motivating scenarios

Beyond the obvious application of individuals wishing to evaluate the quality of a product through word of mouth, the aggregation of noisy estimates across individuals have multiple applications including content moderation [Ghosh, Kale, and McAfee 2011] and personalization [Takács et al. 2009; Isaacman et al. 2011]. For all those applications, improved accuracy ultimately enhances a user's experience, creating a natural incentive for participants to input accurate information. But we assume that alone is typically not sufficient to guarantee a minimum quality for every λ_i. The model of parsimonious agents we introduced draws inspiration from crowd-sourcing models [Ghosh and McAfee 2012; Ghosh and McAfee 2011] applying to a new social information sharing setting.

Why would one consider estimation with parsimonious users that may provide very low quality λ_i? First, the estimation we aim to study may genuinely be difficult and require significant effort. Second, it could be that this estimation is only one task among many to be done in a small amount of time. Participants may then answer very fast or very inaccurately and yet try to make informed estimates overall. Finally, it may not be that the individual task of estimating is costly *per se*, but that disclosing this exact value causes privacy concerns, as modeled in a growing body of research [Ioannidis and Loiseau 2013; Chessa, Grossklags, and Loiseau 2015; Ramachandran and Chaintreau 2015a]. In this situation, a participating individual may decide to provide a lower quality λ_i than the one it possesses, with the the hope that it does not affect the overall estimate too much while retaining privacy.

A special case of interest

We will assume without loss of generality that the quality λ_i of the estimate provided by i is defined as the inverse of that estimate's variance σ_i. If we further assume that the estimate is Gaussian for every node, this implies that combining two estimates of quality λ_1 and λ_2, is equivalent to obtaining a single estimate with quality $\lambda_1 + \lambda_2$. It follows that in the above model we have.

$$\zeta_i = \lambda_i + \sum_{j \in N(i)} \lambda_j \ ,$$

obtained by combining all estimates that node i have produced or seen. All qualitative results of our model hold however functions c and G defined above are chosen, but for tractability it helps to consider a specific case. Without loss of generality we can assume that the estimation error is the variance of the estimate (hence that $G(\zeta_i) = \frac{1}{\zeta_i}$). The choice of the function c is more arbitrary, so to span a large class of convex function, we will assume that $c(\lambda_i) = \frac{C^2}{\alpha+1}\lambda_i^{\alpha+1}$ where $\alpha > 1$ and $C \in \mathbb{R}$. Other choices where G, c are convex and twice differentiable would make the analysis more complex but not significantly different.

4.1.2 Best Response and Goal

Each agent seeks to minimize her overall cost: $J_i(\lambda_i, \zeta_i) = c(\lambda_i) + G(\zeta_i)$.

For an individual, i, their best response occurs when cost is minimized w.r.t. the privacy level λ_i chosen,

$$\phi(\lambda) : \min_{\lambda_i} J_i(\lambda) \text{ s.t. } \lambda_i \geq 0$$

. Hence $c'(\lambda_i) = -G'(\zeta_i)$; Since both are convex functions, it is easy to see that the second derivative is positive and this is hence, a minimum point.

Lemma 14. *In the estimation problem, the best response of an agent i, to its neighbors' effort is given by $\zeta_i^2\lambda_i^\alpha = \frac{1}{C^2}$.*

Proof. For an agent, i, her best response is given as the solution of $c'(\lambda_i) = -G'(\zeta_i)$. Simplifying in the case of the estimation problem: $\frac{d\frac{C^2}{\alpha+1}\lambda_i^{\alpha+1}}{d\lambda_i} = -\frac{d\frac{1}{\zeta_i}}{d(\zeta_i)} \implies C^2\lambda_i^\alpha = \frac{1}{\zeta_i^2}$ □

4.1.3 What is *Wise* Crowd?

We say in this model that a crowd is *wise* if as the network grows all individuals eventually have arbitrarily precise estimate (*i.e.,* $\zeta_i \to \infty$). A crowd is private/parsimonious if all individuals eventually reveal information about their value with arbitrarily small precision (*i.e.,* $\lambda_i \to 0$), or equivalently exerts a vanishing effort. A crowd is privately-wise (or parsimoniously wise) if all individuals are both wise and private (or parsimonious). Ideally as there are more users in the system, the increased access to information compensates for the decreased amount of individual sharing. We would like to understand the conditions under which a sequence of increasing graphs implies that all individuals are wise and private. In the rest of this chapter, we will refer to that results as the wisdom of the private crowds.

Theoretical Examples

In the case of a complete graph, previous work has shown that the crowd is always privately-wise [Chessa, Grossklags, and Loiseau 2015]. On the other hand, a trivial, degenerate case when wisdom of the private crowds fail is when the degree of some nodes in the graph is bounded. Another trivial case is that of a d-regular graph.

Claim 15. *For a d-regular graph, a symmetric Nash Equilibrium always exists and is given by $\lambda^* = \big(\frac{1}{C^2 \cdot (d+1)^2}\big)^{\frac{1}{\alpha+2}}$. Moreover, if the d is increasing in the size of the graph, wisdom of the private crowd always exists.*

Proof. Let $\lambda_i = \lambda^*, \forall i \in N$.

$$C^2 \cdot \zeta_i^2 \lambda_i^\alpha = C^2 \cdot ((d+1)\lambda^*)^2 \lambda^{*\alpha} = 1$$

$$C^2 \cdot (d+1)^2 \lambda^{*\alpha+2} = 1 \implies \lambda^* = \big(\frac{1}{C \cdot (d+1)}\big)^{\frac{2}{\alpha+2}}$$

We see that if d is an increasing function of the size of the graph, $\lambda^* \to 0$. □

However, this property does not always generalize to more complex graphs. Indeed, we show that for a bipartite graph with cubic cost, the crowd usually fails to work together and in fact, small deviations from symmetry lead to suboptimal outcomes (Section 4.3).

4.2 Evolving Social Graphs

4.2.1 Data

In order to study how individuals' sharing efforts progress with the growth of the social graph, we used a dataset which included information about the evolution of the graph. We used a previously collected dataset based on users who post `nytimes.com` articles to `twitter.com` ([May et al. 2014]). This dataset also includes users and relationships up to 2 degrees away from the original posters. The final dataset had 346k users and 13 million edges connecting them. We allowed the graph to grow by adding new nodes based on the following orderings: (1) by creation time of the node on twitter, (2) by activity of the node posting a nytimes.com article (Note that not all the nodes in the graph were actively posting so the latter graph is smaller than the first.), and (3) in random order.

Wisdom of the Private Crowd in an Evolving Real World Graph Our primary question is whether the crowds become more efficient with increasing graph size. In order to test this, we evolved the graph from the NYTIMES dataset by starting with an initial graph of 1.25% of the original size and then adding nodes in the three orderings described, growing it by 25% in each step. At each stage, the nodes played the 'game' based on our privacy model, and we computed the Nash Equilibrium using an iterative algorithm that updates the best response until convergence [Ramachandran and Chaintreau 2015a]. The point of convergence (when it converges) is the Nash equilibrium. In the cases that we examined, the best responses converged to an equilibrium within 20 steps (though our algorithm does not guarantee convergence).

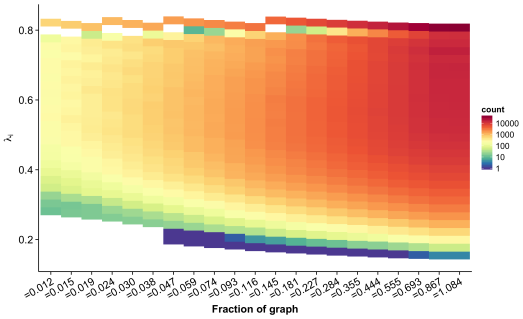

Figure 4.1: Heatmap of distribution of λ_i for increasing large fractions of the graph.

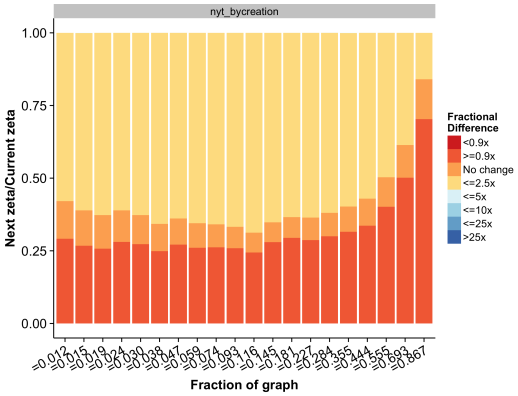

Figure 4.2: Fractional change in ζ_i from one stage to the next. We see that with larger graphs, there are, in fact, an increasing fraction of nodes who lose.

We see that with more nodes in the network, there are more and more nodes putting in less effort (λ_i), *i.e.,* they are taking advantage of their crowd (Figure 4.1). However, this isn't the case for everyone and there are some nodes whose effort doesn't decrease. This is evident from the bimodal-like distribution of the λ_i values with one peak at a high λ_i value and the other at an increasing low value. This same pattern is observed when examining the total information that a node sees (ζ_i). Thus, in a network more complex than a complete or a d-regular graph, the wisdom of the private crowd property does not trivially exist.

We identify those nodes which are 'winners', 'losers' and 'constant' based on their difference in effort from the graph in one stage to the next. We would expect that with more nodes in the network, they leverage each others' efforts and gain more information while not having to put in as much effort. We plot, at each stage, the fraction of nodes that gain, lose or stay the same (Figure 4.2). Surprisingly, we see that this is not the case and, in fact, there are *more* losers with larger graphs.

One way to better understand the degree of loss compared to the gains is to consider a measure of inequality among all the nodes. The Gini index, a measure of statistical dispersion, is one such measure. The index is based on the Lorenz curve which plots

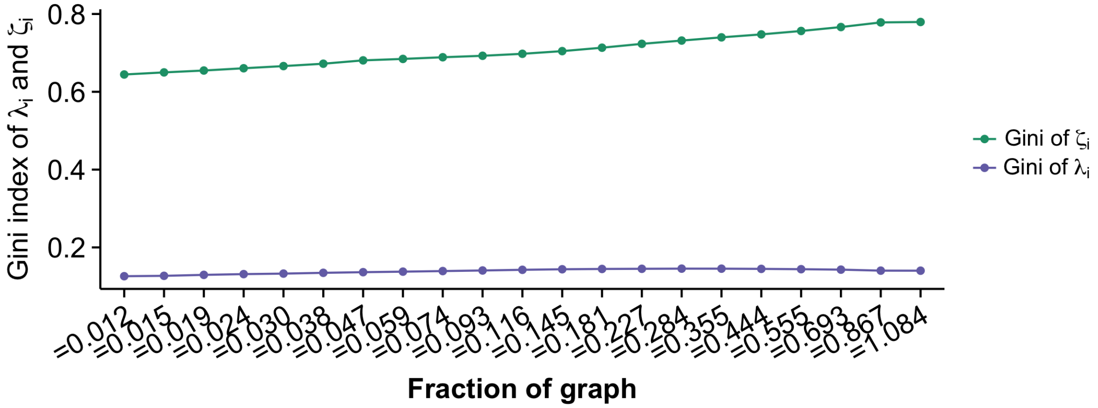

Figure 4.3: Gini index measured of the distributions of λ_i and ζ_i. The Gini index is a measure of inequality – lower values indicate less inequality.

the percentage of the effort (or total observed precision) made by the bottom x% of the population. The Gini coefficient is the ratio of the area between the line of equality and the Lorenz curve and the total area under the line of inequality. It is typically used as a measure of income inequality. A Gini index of 0 indicates perfect equality. We plot the Gini index for the distribution of λ_i's and ζ_i's for each graph size (Figure 4.3). Here, we can clearly see that with increasing graph size, the inequality increases (whether considering inequality of λ or ζ).

Who are the Winners and Losers? While we can clearly see that the degree of inequality increases with larger graph size, it's less clear which nodes are the ones who gain or lose. To better understand this, we split nodes by their relative twitter age, *i.e.,* the number of stages for which they are present in the network. We plot the relative change in the total effort that a node sees (ζ_i) by this relative age, focusing on the final stage of growth from 87% of the network to the whole network (Figure 4.4). We observe that for, other than the very new nodes, a roughly similar proportion of the nodes lose (depicted in red) from the penultimate stage to the last one. However, we also note that fraction of nodes that *gain* decrease with age (depicted in teal) *i.e.,* older nodes stand to gain more than newer nodes

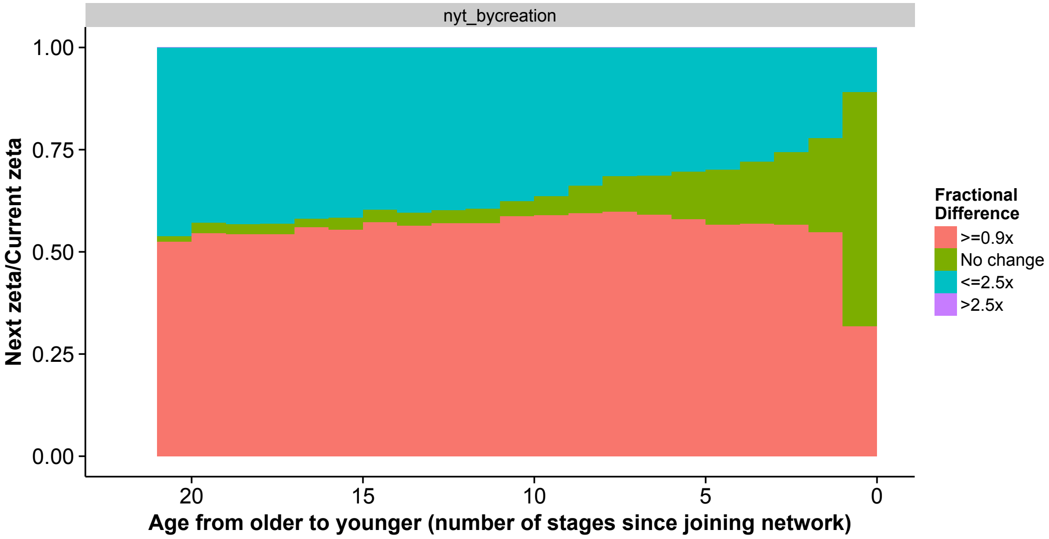

Figure 4.4: Fractional change in ζ_i from the penultimate stage to the last stage, split by the relative twitter age on the x-axis.

when the network grows. We observe that while there are more nodes which are worse off, the majority of those are newer nodes to the network. Nodes that have been in the network from the earlier stages have a better chance of eventually gaining, sometimes even by a large magnitude.

On short time scales, the majority of nodes lose. However, when we look at longer time spans, we find that nodes tend to compensate for their losses and overall benefit from the larger networks. Figure 4.5 shows the change in a node's total precision (ζ_i) from each stage to the final graph. Essentially, we ask the question, how much does a node gain/lose in the end stage compared to the current one? We again see that a node's age in the network affects whether they gain or lose. Note that while the last few stages seem like a small part of the evolution, they actually represent the growth from 75k nodes to 340k nodes. The young nodes which lose as a result represent a significant fraction of the network.

We see that larger gains only come after significant expansion in the network – the network has to double or more in size before it affects most nodes. In fact, we see that it is only after the network grows to more than 15% of its its total size that we see significant gains. The small minority who have made substantial gains are not only older but they are

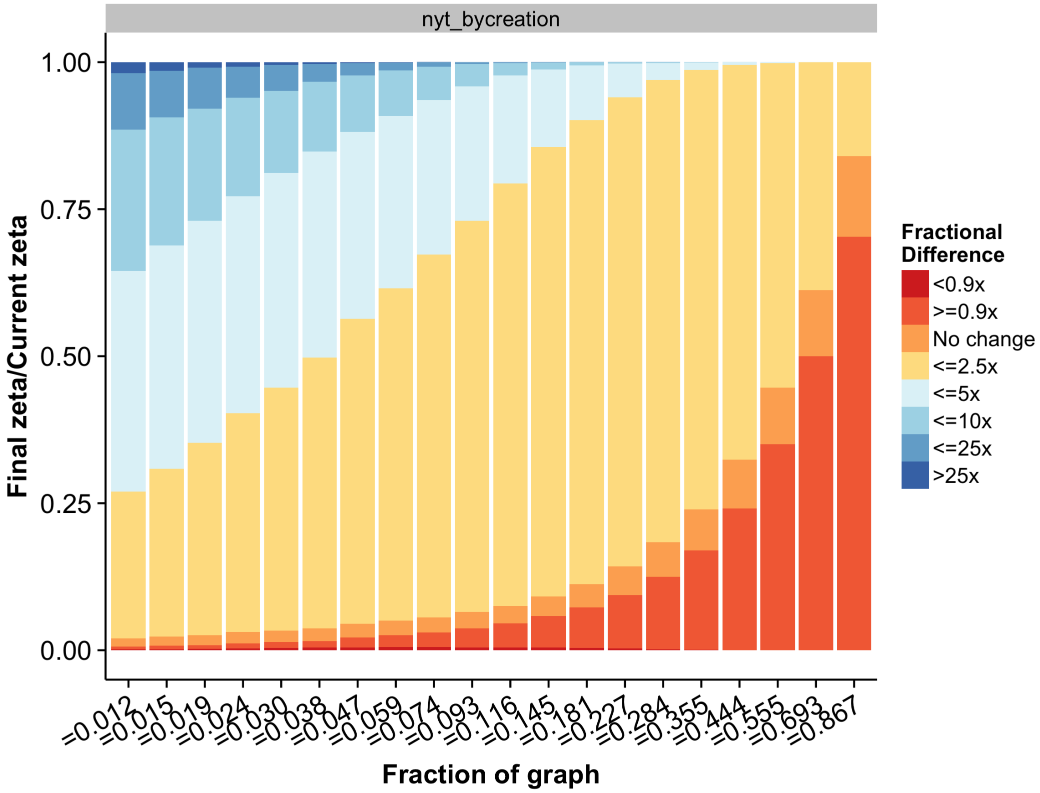

Figure 4.5: Fractional change from current stage to the final network.

also better connected. In figure 4.6 we see that the nodes that gain are increasingly better connected in larger networks and can, thus, take advantage of that.

The structure of the network itself is a key factor in deciding who are the winners and losers. We develop a null hypothesis model where the graph grows randomly, rather than by the nodes' creation time. We compare this null hypothesis with a graph evolving by when a node joins the network (Figure 4.2) and when it becomes active in the network (Figure 4.7). In both the null model and the evolution by activity, the characteristics of the winners and losers are less skewed and age gives less of an advantage. The natural evolution clearly results in some hierarchy that results in the observed imbalance of effort. In the next section, we examine the consequences of such a hierarchy by considering some simple models.

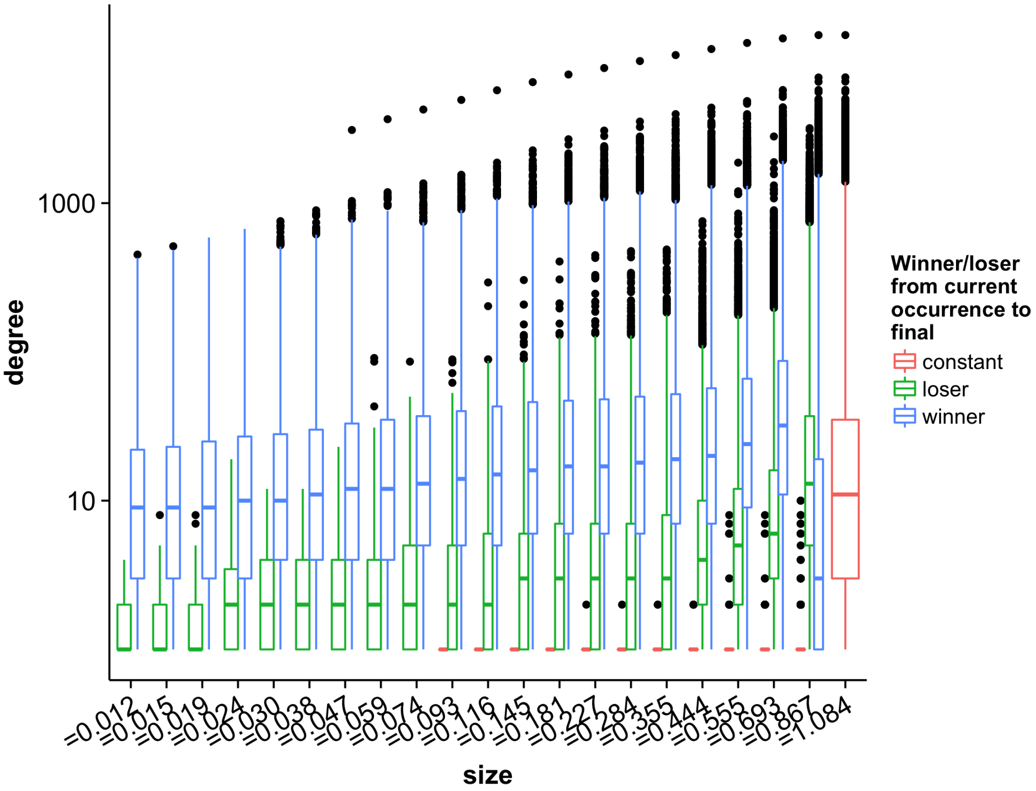

Figure 4.6: Difference in degree distributions for winners vs losers. The classification of winner vs loser is computed from the current stage to the final stage.

4.3 Network Structure and Inequality

In Chapter 3, we saw that the spectral features of the graph, specifically the minimum eigenvalue, had an impact on the outcome of the equilibrium. Certain graphs were more susceptible to specialized and unequal outcomes. In this section, we examine how the size of the graph (and thus structural features) of the graph affects the individual users and whether they gain or lose overall by being part of a bigger or more connected graph.

4.3.1 Network Normality

Lemma 16. *In the estimation problem with a general convex and increasing privacy cost, the network normality condition can be expressed as* $\frac{c''(\lambda_i)}{c'(\lambda_i)} \cdot \frac{1}{\sqrt{c'(\lambda_i)}} \geq 2 \cdot (-\mu_{min} - 1)$.

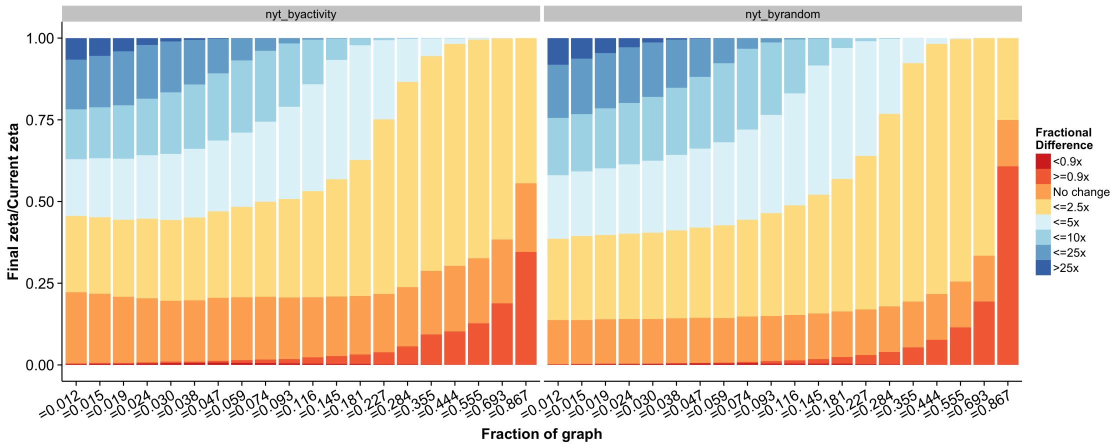

Figure 4.7: Fractional change from current stage till the final stage for the graph evolving by activity (right) and randomly (left).

Proof. The network normality condition is $\frac{1}{\mu_{min}} \leq \frac{\partial\lambda_i}{\partial\lambda_{-i}}$ [Bramoullé et al. 2014] .

Differentiating the best response wrt λ_{-i} :

$$\frac{d^2c(\lambda_i)}{d\lambda_i^2}\cdot\frac{d\lambda_i}{d\lambda_{-i}} = -\frac{d^2G(\zeta_i)}{d(\zeta_i)2}\cdot\frac{d(\zeta_i)}{d\lambda_{-i}}$$

$$\frac{d\lambda_i}{d\lambda_{-i}} = \frac{-G''(\zeta_i)}{c''(\lambda_i)+G''(\zeta_i)}$$

Incorporating this with the Network Normality condition (4.3.1),

$$\begin{aligned}\frac{1}{\mu_{min}} &\leq \frac{\partial\lambda_i}{\partial\lambda_{-i}}\\ &\leq \frac{-G''(\zeta_i)}{c''(\lambda_i)+G''(\zeta_i)}\\ \frac{c''(\lambda_i)}{G''(\zeta_i)} &\geq -\mu_{min}-1\\ \frac{c''(\lambda_i)}{c'(\lambda_i)}\cdot\frac{1}{\sqrt{c'(\lambda_i)}} &\geq 2\cdot(-\mu_{min}-1)\end{aligned}$$

□

Lemma 17. *In the estimation problem with polynomial privacy cost, the network normality* *condition can be expressed as* $\lambda_i \leq \big(\frac{\alpha}{2C\cdot(-\mu_{min}-1)}\big)^{\frac{1}{(1+\frac{\alpha}{2})}}$.

Proof. Substituting a polynomial cost in Lemma 16,

$$\frac{C^2\alpha\lambda^{\alpha-1}}{C^2\lambda^{\alpha}} \cdot \frac{1}{\sqrt{C^2\lambda^{\alpha}}} \geq 2\cdot(-\mu_{min}-1)$$

$$\lambda_i \leq \big(\frac{\alpha}{2C\cdot(-\mu_{min}-1)}\big)^{\frac{1}{(1+\frac{\alpha}{2})}}$$

□

4.3.2 Bipartite Graph

In networked public goods, the bipartite graph proves to be especially interesting. As we will see in §4.3.3, graphs with less negative minimum eigenvalues (μ_{min}) values tend exhibit wisdom of the private crowd. A complete $K_{m.n}$ graph ($\mu_{min} = -\sqrt{mn}$) is the worst-case.

Theorem 18. *Let* $G(L, R, E)$ *be a regular bipartite graph with* $|L| = m, |R| = n$*, with left-degree* γn *and right-degree* γm*. Let* $\rho(n) = \frac{m}{n}$ *denote the imbalance. When* $\alpha = 2$ *(cubic privacy cost), the following holds:*

- *If* $\lim_{n\to\infty}\rho(n) = 1$*, then wisdom of the private crowds exists,* i.e., $\forall i \in L \cup R$, $\lim_{n\to\infty}\lambda_i = 0$ *and* $\lim_{n\to\infty}\zeta_i = \infty$*. Further, if* $\lim_{n\to\infty}(m-n)\gamma > 0$*, nodes in* L *and* R *accumulate information at increasing, but different rates.*
- *If* $\lim_{n\to\infty}\rho(n) = \rho \neq 1$*, wisdom of the private crowds does not exist.*

Proof. Let $\gamma_m = \beta\gamma_n$. Expressing γ_m and γ_n in terms of λ_m and λ_n and solving, $\zeta_n^2 = \frac{1}{C}\cdot\frac{\gamma^2 mn-1}{\gamma m\beta-1}$ and $\zeta_m^2 = \frac{1}{C}\cdot\beta^2\cdot\frac{\gamma^2 mn-1}{\gamma m\beta-1}$, where $\beta = \frac{-\gamma(m-n)\pm\sqrt{\gamma^2(m-n)^2+4}}{2}$.

For large $\gamma(m-n)$, $\lim_{n\to\infty}\beta \approx \frac{1}{\gamma(m-n)}$.

Case 1 (Imbalanced Partitions): $\lim_{n\to\infty} \phi(n) = \phi$**;** $\phi > 1$

$$
\begin{aligned}
\zeta_n^2 &\approx \frac{1}{C} \cdot (\gamma^2 mn - 1)(\phi(n) - 1) \\
\lim_{n\to\infty} \zeta_n^2 &= \infty \\
\text{And, } \zeta_m^2 &\approx \beta^2 \frac{1}{C} \cdot (\gamma^2 mn - 1)(\phi(n) - 1) \\
&= \frac{1}{C} \cdot \frac{\phi}{(\phi - 1)} \\
\lim_{n\to\infty} \zeta_m^2 &= \Theta(1)
\end{aligned}
$$

$\therefore$ The crowd is not fully wise. The nodes in M are not wise while the nodes in N are wise.

Case 2 (Balanced Partitions): $\lim_{n\to\infty} \phi(n) = 1$ This case is further divided based on how $\gamma(m - n)$ grows. Let the degree of each node be d.

Case 2.1: $\lim_{n\to\infty} \gamma(m - n) = 0$

$\lim_{n\to\infty} \beta = 1$. $\therefore$ $\lambda_m = \lambda_n$. $\lambda_{m \text{ or } n} = \frac{1}{\sqrt{C(d+1)}}$ and $\lim_{d\to\infty} \lambda_{m \text{ or } n} = 0$. Thus, the crowd is always wise.

Case 2.2: $\lim_{n\to\infty} \gamma(m - n) = \Theta(1)$

$\lim_{n\to\infty} \beta = \Theta(1) \Rightarrow \lambda_m = \Theta(1) \cdot \lambda_n$. Thus $\lambda_n = \frac{1}{\sqrt{C(1+d\Theta(1))}}$ and $\lambda_m = \frac{\Theta(1)}{\sqrt{C(1+d\Theta(1))}}$. For larger degrees, $\lim_{d\to\infty} \lambda_m = 0$ and $\lim_{d\to\infty} \lambda_n = 0$. The crowd is 'wise' but at different rates (within a constant factor).

Case 2.3: $\lim_{n\to\infty} \gamma(m - n) = \infty$

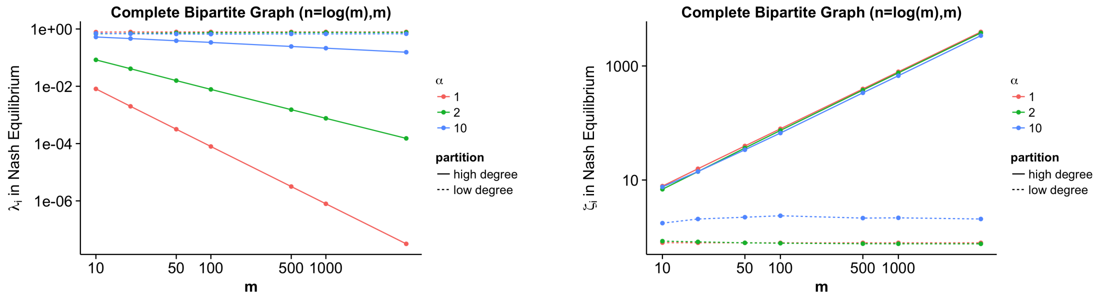

Figure 4.8: Privacy choices ((left) λ_i and (right) ζ_i) in Nash equilibrium for a bipartite graph with differential rates of growth in each partition ($C = 0.1$). The higher degree nodes quickly become wise while the lower degree ones do not improve substantially with size.

$$
\begin{aligned}
\zeta_n^2 &\approx \frac{1}{C} \cdot (\gamma^2 n^2 \phi(n) - 1)(\phi(n) - 1) && \text{(from case 1)} \\
&\approx \frac{1}{C} \cdot \gamma^2 n^2 (\phi(n) - 1) && (\because -1 \text{ is negligible}) \\
\lim_{n\to\infty} \zeta_n^2 &= \infty \\
\zeta_m^2 &\approx \frac{1}{C} \cdot \frac{\gamma^2 n^2 \phi(n)}{\gamma^2 n^2 (\phi(n) - 1)} && \text{(from case 1)} \\
&= \frac{1}{C} \cdot \frac{\phi(n)}{(\phi(n) - 1)} \\
\lim_{n\to\infty} \zeta_m^2 &= \infty
\end{aligned}
$$

The crowd is wise but not equally so; nodes in N becomes wiser faster than nodes in M. □

We see that, in a bipartite graph, even a small deviation from symmetric partitions results in the better connected partition increasingly benefiting as the cost of the other partition. To see some examples for how this dynamic plays out, we numerically compute the Nash equilibrium for increasing sizes of bipartite graphs (Figure 4.8). One partition grows linearly, the other logarithmically. The higher degree nodes quickly become wise while the lower degree ones do not improve substantially with size. Thus, we see that

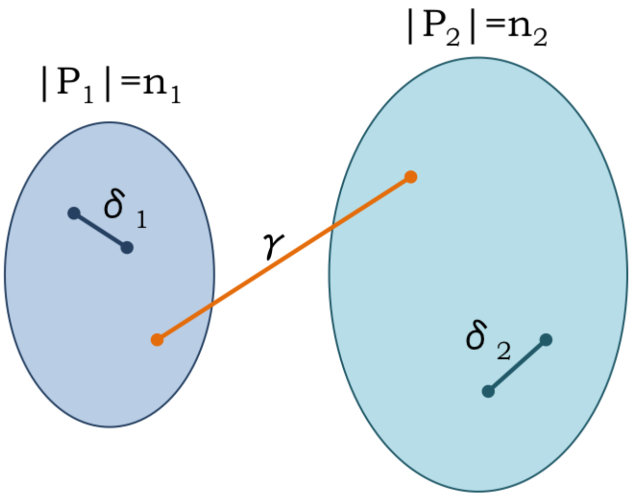

Figure 4.9: Graphical depiction of regular arbitrary clustered graph along with relevant parameters.

wisdom of the private crowd only exists for the nodes that grow much faster.

4.3.3 Clustered Graph

Building on our previous case of the bipartite graph, we consider a model where the two partitions are connected *i.e.,* a clustered graph. Specifically, we consider the case of a regular arbitrary graph with two clusters (Figure 4.9).

Definition: Consider a family of graphs $\mathcal{G}_{n_1,n_2,\delta_1,\delta_2,\gamma}$ with 2 clusters, P_1 and P_2. Let $|P_1| = n_1$ and $|P_2| = n_2$. $\forall i \in P_1$, the node i is connected to $\delta_1 n_1$ nodes in P_1 and γn_2 nodes in P_2. The nodes in P_2 are similarly connected to $\delta_2 n_2$ nodes in P_2 and γn_1 nodes in P_1. For convenience in notation, let $n_1 = n$. Let $\phi(n) = \frac{n_2}{n_1}$ be the size imbalance between the two clusters.

Best response: All the nodes within the same cluster have the same λ since they have the same degree (and extended neighborhood). As in the case of the bipartite graph, we express ζ_1, ζ_2 in terms of λ_1, λ_2 and simplify.

Lemma 19. *Let* $\zeta_2 = \beta\zeta_1$. *Then,* $\zeta_1^2 = \frac{1}{C} \cdot \frac{\gamma^2 n_1 n_2 - (1+\delta_1 n_1)(1+\delta_2 n_2)}{\gamma n_2 \beta - (1+\delta_2 n_2)}$ *and* $\zeta_2^2 = \beta^2 \cdot \frac{1}{C} \cdot \frac{\gamma^2 n_1 n_2 - (1+\delta_1 n_1)(1+\delta_2 n_2)}{\gamma n_2 \beta - (1+\delta_2 n_2)}$ *where* $\beta = \frac{-\gamma(n_2 - n_1)}{2(1+\delta_1 n_1)}(1 - \sqrt{1 + \frac{4(1+\delta_1 n_1)(1+\delta_2 n_2)}{\gamma^2 (n_2 - n_1)^2}})$.

We find that β can be further simplified based on the relationships of $\delta_1, \delta_2, \gamma$ and ϕ. We define two characteristics of the graph based on whether the clusters are internally tightly connected (homophily) or are more connected with each other (heterophily).

Definition: A graph $G \in \mathcal{G}_{n_1,n_2,\delta_1,\delta_2,\gamma}$ is *heterophilic* when $\frac{4\delta_1\delta_2\phi}{\gamma^2(\phi-1)^2} \to 0$ and *homophilic* when $\frac{4\delta_1\delta_2\phi}{\gamma^2(\phi-1)^2} \to \infty$.

Lemma 20. *When a graph $G \in \mathcal{G}_{n_1,n_2,\delta_1,\delta_2,\gamma}$ is heterophilic, $\beta \approx \frac{(1+\delta_2 n_2)}{\gamma(n_2-n_1)}$. When a graph $G \in \mathcal{G}_{n_1,n_2,\delta_1,\delta_2,\gamma}$ is homophilic, $\beta \approx \sqrt{\frac{1+\delta_2 n_2}{1+\delta_1 n_1}}$.* [1]

Proof. In a heterophilic graph,

$$\begin{aligned}
\beta &= \frac{-\gamma(n_2-n_1)}{2(1+\delta_1 n_1)}(1-\sqrt{1+\frac{4(1+\delta_1 n_1)(1+\delta_2 n_2)}{\gamma^2(n_2-n_1)^2}})\\
&\approx \frac{-\gamma(n_2-n_1)}{2(1+\delta_1 n_1)}(1-1-\frac{1}{2}\cdot\frac{4(1+\delta_1 n_1)(1+\delta_2 n_2)}{\gamma^2(n_2-n_1)^2})\\
&= \frac{(1+\delta_2 n_2)}{\gamma(n_2-n_1)}
\end{aligned}$$

In a homophilic graph,

$$\begin{aligned}
\beta &= \frac{-\gamma(n_2-n_1)}{2(1+\delta_1 n_1)}(1-\sqrt{1+\frac{4(1+\delta_1 n_1)(1+\delta_2 n_2)}{\gamma^2(n_2-n_1)^2}})\\
&\approx \frac{\gamma(n_2-n_1)}{2(1+\delta_1 n_1)}(\frac{2\sqrt{(1+\delta_1 n_1)(1+\delta_2 n_2)}}{\gamma(n_2-n_1)}) = \sqrt{\frac{1+\delta_2 n_2}{1+\delta_1 n_1}}
\end{aligned}$$

□

Theorem 21. *Let $G \in \mathcal{G}_{n_1,n_2,\delta_1,\delta_2,\gamma}$ be a heterophilic graph. When $\alpha = 2$ (cubic privacy cost), the following holds:*

- *If $\phi = \lim_{n\to\infty}\phi(n) > 1$ and $\delta_2 = 0$, wisdom of the private crowds does not exist.*
- *If $\phi = \lim_{n\to\infty}\phi(n) > 1$ and $\lim_{n\to\infty}\delta_2 n = \infty$, wisdom of the private crowds exists but nodes in P_1 and P_2 accumulate information at different rates.*

[1] We use the notation $f(n) \approx g(n)$ to denote $f(n) = g(n) \pm o(g(n))$

- *If* $\lim_{n\to\infty}\phi(n) = 1$ *and* $\lim_{n\to\infty}(\phi(n)-1)n = \infty$*, wisdom of the private crowds exists.*
- *If* $\lim_{n\to\infty}\phi(n) = 1$ *and* $\lim_{n\to\infty}(\phi(n)-1)n = 0$*, wisdom of the private crowds does not exists.*

Proof. **Case 1 (Clusters are imbalanced):** $\phi(n_1) = \phi > 1$

Case 1.1: $\delta_2 n = 0$: In this case, $\beta \approx \frac{1}{\gamma(n_2-n_1)}$.

$$\begin{aligned}
\therefore,\ \zeta_1^2 &= \frac{1}{C}\cdot\frac{\gamma^2 n_1 n_2 - (1+\delta_1 n_1)(1+\delta_2 n_2)}{\gamma n_2\beta - (1+\delta_2 n_2)} \\
&= \frac{1}{C}\cdot\frac{\gamma^2\phi n^2 - (1+\delta_1 n)}{\gamma\phi n\frac{1}{\gamma n(\phi-1)} - 1} \\
&= \frac{1}{C}\cdot(\phi-1)(\gamma^2\phi n^2 - 1 - \delta_1 n) \\
\lim_{n\to\infty}\zeta_1^2 &= \infty \\
\text{and, } \zeta_2^2 &= \beta^2\zeta_1^2 \\
&= \frac{1}{C}\cdot\frac{1}{\gamma^2 n^2(\phi-1)^2}\cdot(\phi-1)(\gamma^2\phi n^2 - 1 - \delta_1 n) \\
&= \frac{1}{C}\cdot(\frac{\phi}{\phi-1} - \frac{1}{\gamma^2 n^2(\phi-1)} - \frac{\delta_1}{\gamma^2 n(\phi-1)}) \\
\lim_{n\to\infty}\zeta_2^2 &= \Theta(1)
\end{aligned}$$

Thus, the crowd is partially wise and the partition which is not internally connected doesn't improve significantly with more nodes.

Case 1.2: $\delta_2 n \to \infty$:

$$\zeta_1^2 \approx \frac{1}{C} \cdot \frac{\gamma^2 n_1 n_2 - (1+\delta_1 n_1)(1+\delta_2 n_2)}{\gamma n_2 \cdot \frac{(1+\delta_2 n_2)}{\gamma(n_2-n_1)} - (1+\delta_2 n_2)} \text{ (Ł20)}$$

$$\approx \frac{1}{C} \cdot \frac{(\gamma^2\phi - \frac{1}{n^2} - \frac{\delta_1}{n} - \frac{\delta_2\phi}{n} - \delta_1\delta_2\phi)(\phi - 1)}{\frac{1}{n^2} + \frac{\delta_2\phi}{n}}$$

$$\lim_{n\to\infty} \zeta_i^2 = \infty$$

$$\text{and, } \zeta_2^2 = (\frac{(1+\delta_2 n_2)}{\gamma(n_2-n_1)})^2 \cdot \frac{1}{C} \cdot \frac{\gamma^2 n_1 n_2 - (1+\delta_1 n_1)(1+\delta_2 n_2)}{\gamma n_2 \frac{(1+\delta_2 n_2)}{\gamma(n_2-n_1)} - (1+\delta_2 n_2)}$$

$$= \frac{(1+\delta_2\phi n)}{\gamma^2(\phi-1)} \cdot \frac{1}{C} \cdot (\gamma^2\phi - (\frac{1}{n} + \delta_1)(\frac{1}{n} + \delta_2\phi))$$

$$\lim_{n\to\infty} \zeta_2^2 = \infty$$

The crowd is wise but unfair with one partition gaining more information than the other.

Case 2 (Clusters are balanced): $\phi(n) \to 1$

$$\zeta_1^2 = \frac{1}{C} \cdot \frac{\gamma^2 n_1 n_2 - (1+\delta_1 n_1)(1+\delta_2 n_2)}{\gamma n_2 \beta - (1+\delta_2 n_2)} \; (Lemma 20)$$

$$\approx \frac{1}{C} \cdot \frac{(\gamma^2\phi n^2 - 1 - \delta_1 n - \delta_2\phi n - \delta_1\delta_2\phi n^2)(\phi-1)}{1 + \delta_2\phi n}$$

$$\approx \frac{1}{C} \cdot \frac{(\gamma^2\phi - \frac{1}{n^2} - \frac{\delta_1}{n} - \frac{\delta_2\phi}{n} - \delta_1\delta_2\phi)(\phi - 1)}{\frac{1}{n^2} + \frac{\delta_2\phi}{n}}$$

$$\lim_{n\to\infty} \zeta_1^2 = \begin{cases} \infty, \text{if } \lim_{n\to\infty}(\phi(n)-1)n = \infty \\ 0, \text{if } \lim_{n\to\infty}(\phi(n)-1)n = 0 \end{cases}$$

Similar to case 1.2,

$$\zeta_2^2 = (\frac{(1+\delta_2 n_2)}{\gamma(n_2-n_1)})^2 \cdot \frac{1}{C} \cdot \frac{\gamma^2 n_1 n_2 - (1+\delta_1 n_1)(1+\delta_2 n_2)}{\gamma n_2 \frac{(1+\delta_2 n_2)}{\gamma(n_2-n_1)} - (1+\delta_2 n_2)}$$

$$= \frac{(1+\delta_2\phi n)}{\gamma^2(\phi-1)} \cdot \frac{1}{C} \cdot (\gamma^2\phi - (\frac{1}{n} + \delta_1)(\frac{1}{n} + \delta_2\phi))$$

$$\lim_{n\to\infty} \zeta_2^2 = \infty$$

□

In the (above) heterophilic case, the internal connections of a cluster can compensate for the external connections (or lack of) but only if the cluster is sufficiently connected. The wisdom of the private crowd property only holds in those cases where there is sufficient compensation.

Theorem 22. *Let $G \in \mathcal{G}_{n_1,n_2,\delta_1,\delta_2,\gamma}$ be a homophilic graph. When $\alpha = 2$, wisdom of the crowd always exists.*

Proof. We consider different cases based on ϕ.

Case 1 (Clusters are imbalanced): $\phi > 1$

$$\begin{aligned}\zeta_1^2 &= \frac{1}{C} \cdot \frac{\gamma^2 n_1 n_2 - (1+\delta_1 n_1)(1+\delta_2 n_2)}{\gamma n_2 \sqrt{\frac{1+\delta_2 n_2}{1+\delta_1 n_1}} - (1+\delta_2 n_2)} \; (Lemma 20) \\ &= \frac{1}{C} \frac{1+\delta_2 \phi n}{1+\delta_1 n} \cdot \frac{\gamma^2 \phi n^2 - (1+\delta_1 n)(1+\delta_2 \phi n)}{\gamma \phi n \sqrt{\frac{1+\delta_2 \phi n}{1+\delta_1 n}} - (1+\delta_2 \phi n)}\end{aligned}$$

$$\lim_{n\to\infty} \zeta_1^2 = \infty$$

Similarly, $\lim_{n\to\infty} \zeta_2^2 = \infty$.

Case 2 (Clusters are perfectly balanced): $\phi = 1$

$\beta \approx \sqrt{\frac{1+\delta_2 n_2}{1+\delta_1 n_1}} \approx \sqrt{\frac{\delta_2}{\delta_1}}$, for large n. Since β is a constant, ζ_1 and ζ_2 are within a constant fraction. Thus both clusters are wise, though depending on the values of δ_1 and δ_2, one will consistently be better off than the other.

Case 3 (Clusters are balanced): $\phi \to 1$ As in case 1, both $\lim_{n\to\infty} \zeta_1^2 = \infty$ and

$\lim_{n \to \infty} \zeta_2^2 = \infty$.

$$\begin{aligned}
\zeta_1^2 &= \frac{1}{C} \cdot \frac{\gamma^2 \phi n^2 - (1+\delta_1 n)(1+\delta_2 \phi n)}{\gamma \phi n \sqrt{\frac{1+\delta_2 \phi n}{1+\delta_1 n}} - (1+\delta_2 \phi n)} \\
&\to \infty && \text{(as in imbalanced case)} \\
\zeta_2^2 &= \frac{1}{C} \frac{1+\delta_2 \phi n}{1+\delta_1 n} \cdot \frac{\gamma^2 \phi n^2 - (1+\delta_1 n)(1+\delta_2 \phi n)}{\gamma \phi n \sqrt{\frac{1+\delta_2 \phi n}{1+\delta_1 n}} - (1+\delta_2 \phi n)} \\
&\to \infty && \text{(as in imbalanced case)}
\end{aligned}$$

□

The homophilic more closely resembles a complete graph (or rather multiple complete graphs) and, like the complete graph, always has a wise and private crowd. When we consider the transition from the extreme hierarchy of the bipartite graph to the less extreme clustered graph, we see that even a small amount of connectedness within a cluster can be enough to allow all nodes to be wise. In the next section, we examine a condition based on spectral properties of the graph that are sufficient for *any* graph to have a wise, private crowd.

4.3.4 A General Condition for Wisdom of the Private Crowd

Many graphs, however, do not fall into the extreme conditions seen in the bipartite graph. Consider a series of graphs of increasing size: since the Nash equilibrium dynamics do not depend on the way the graph was built but, rather, the overall structure, we can independently consider each graph in the series without involving the intermediate stages. As graphs get larger, the amount of information that is available increases. A wise crowd is one that takes advantage of this, *i.e.,* even when the amount of individual effort decreases, the amount of information received by individuals increases. A sufficient condition wisdom of the private crowds is for the graph to be network normal, a condition that states that when privacy costs are sufficiently low compared to estimation costs, a unique Nash equilibrium exists [Allouch 2013; Allouch 2015; Bramoullé and Kranton 2006; Bramoullé et al. 2014].

When considering the best response of a node in the context of network normality, we find that the amount of work that any node does is bounded and is inversely proportional to the minimum eigenvalue, μ_{min}. For many graph families, $|\mu_{min}|$ increases with the size of the graph (*e.g.,* $\mu_{min} = -\frac{n}{2}$ for a complete bipartite graph of n nodes). Thus we see that the condition of network normality is sufficient for the crowd to be privately wise.

Theorem 23. *For a series of graphs $G_1, G_2, .., G_t, ..$, where $G_1 \subset G_2 \subset G_3..$, if the graphs are network normal, $|\mu_{min}|$ is an increasing function of $|G_t|$ and $\lim_{|G_t| \to \infty} |\mu_{min}| = \infty$, the crowd is privately wise.*

Proof. For network normality, $\frac{1}{\mu_{min}} < \frac{\partial \lambda_i}{\partial \lambda_{-i}} < 0$ [Allouch 2013; Allouch 2015; Bramoullé and Kranton 2006; Bramoullé et al. 2014].

$$\begin{aligned}
\frac{1}{\mu_{min}} &< \frac{\partial \lambda_i}{\partial \lambda_{-i}} \\
&< \frac{-G''(\zeta_i)}{c''(\lambda_i) + G''(\zeta_i)} \\
\frac{c''(\lambda_i)}{G''(\zeta_i)} &> -\mu_{min} - 1 \\
\frac{c''(\lambda_i)}{c'(\lambda_i)} \cdot \frac{1}{\sqrt{c'(\lambda_i)}} &> 2 \cdot (-\mu_{min} - 1) \\
\lambda_i &< \Big(\frac{\alpha}{2C \cdot (-\mu_{min} - 1)}\Big)^{\frac{1}{(1+\frac{\alpha}{2})}}
\end{aligned}$$

If a graph is network normal, then $\lambda_i < \big(\frac{\alpha}{2C \cdot (-\mu_{min} - 1)}\big)^{\frac{1}{(1+\frac{\alpha}{2})}}$. If μ_{min} increases with the size of the graph and $\lim_{|G_t| \to \infty} |\mu_{min}| = \infty$ (which is true for many classes of graph), then $\forall i, \lambda_i \to 0$ (and thus, $\zeta_i \to \infty$). Thus, when a graph is network normal, there is also wisdom of the private crowd. □

4.4 Conclusion

Social networks and social media enabled information sharing at unprecedented scale. We expanded on the notion of wisdom of the crowd to also encompass some cost associated with gathering or sharing information. We introduced a network game between nodes, allowing them to trade-off the cost of sharing information with the reward of better information. We studied the evolution of the equilibria in the growth of a graph derived from Twitter and showed that while many nodes are better off as the graph grows, a significant fraction is worse off. We showed that this results from the hierarchical evolution of the graph. Using some simple hierarchical models, we showed a connection between the imbalance in the graph and the failure to achieve ideal information sharing. Further, we showed a general result, based on spectral features of the graph, which gives a sufficient condition for the wisdom of the private crowd property to be attained.

Conclusion

In this thesis, we aimed to better understand information exchange in online social networks. In these studies, we primarily focused on two major questions: (1) To what extent can we exploit the relationships between different type of data to infer properties of diffusion and (2) How is the structure of a network related to the diffusion of information? To answer these questions, we relied on several analysis techniques including cross-domain inference techniques, predictive modeling and game theory.

In these studies, we found that one needs surprisingly little information in order to learn features of the data. The key to being able to do so is in combining different types of information that reveal different facets of the data. We use this cross-domain techniques in several settings. The first setting is one in which we identify even passive individuals online by integrating private browsing behavior with the network information and the corresponding visible online actions. The second is one in which we use early sharing information, combined with network based features, to predict performance of posts. With these crucial pieces of information, we developed a two stage prediction model to accurately estimate the popularity of a link.

We also saw how network structure and relationships is a key piece of information across many contexts. While many researchers have previously used these features, we have shown new techniques and mechanisms through which we can exploit this information. It proves useful in identifying users through their uniqueness in their graph position and friends. It also provides extra knowledge when predicting diffusion by accounting for the effect of different types of individuals.

Further, we saw how network structure provides crucial knowledge in modeling user behavior. We showed how public good models can be used to model content curation and information flow on social networks. We also prove that the network structure has strong effects on the resultant equilibria. We showed that the observed specialization arises from the relationship of spectral features of the network and the lifetime of the content.

4.4.1 Open Questions

Our results lead to many interesting open questions across different aspects of information diffusion in a network. From an inference perspective, it is interesting to understand the limits of prediction and inference in a network (both of user data and of network-related data). For instance, we showed that an hour of data was adequate to predict popularity – is a minute then enough? ten minutes? Can this limit be broken by integrating additional pieces of data (from say, additional networks)? When predicting user identity, can we use similar techniques to identify other properties of the user? We relied on some degree of uniqueness (k-anonymity) of users. Even if users are not uniquely identifiable, we could potentially exploit their degree of anonymity to infer some interests and properties. Further, are there measures that a user can take to resist such privacy-breaking efforts?

There are also many active research questions in the area of content production. Current models are limited in the complexity of either the network they study or the user behavior. One way to model content production mechanisms more realistically is to assume heterogeneity in individual utility functions. Moreover, in growing and evolving networks, what are features of real world networks that result in better and worse outcomes? Are there features of individuals that make them better suited to effectively learn in different types of networks? As a network designer, are there more optimal ways to suggest edges to add to the network?

While some of these questions might be answered through social networks analysis techniques, newer techniques from other fields might be useful to answer others.

Bibliography

Acemoglu, Daron, Asuman Ozdaglar, and A ParandehGheibi (2010). “Spread of (mis) information in social networks.” In: *Games and Economic Behavior* 70.2, pp. 194–227.

Acemoglu, Daron et al. (2011). “Bayesian learning in social networks.” In: *The Review of Economic Studies* 78.4, pp. 1201–1236.

Adamic, Lada, Thomas Lento, and Andrew Fiore (2012). *How You Met Me*. URL: `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4681`.

Allouch, Nizar (2013). “The Cost of Segregation in Social Networks.” In: *SSRN Electronic Journal*.

— (2015). “On the Private Provision of Public Goods on Networks.” In: *Journal of Economic Theory* forthcoming, pp. 1–34.

Altman, Eitan (2015). “Trend detection in social networks using Hawkes processes.” In: pp. 1–15.

An, J et al. (2011). “Media landscape in Twitter: A world of new conventions and political diversity.” In: *Proceedings of the International Conference Weblogs and Social Media (ICWSM)*, pp. 18–25.

An, J et al. (2014). “Sharing political news: the balancing act of intimacy and socialization in selective exposure.” In: *EPJ Data Science*.

Ayenson, Mika et al. (2011). “Flash cookies and privacy II: Now with HTML5 and ETag respawning.” In: *World Wide Web Internet And Web Information Systems*.

Bakshy, Eytan et al. (2011). “Everyone’s an influencer: quantifying influence on twitter.” In: *WSDM ’11: Proceedings of the fourth ACM international conference on Web search and data mining*. ACM Request Permissions.

Bakshy, Eytan et al. (2012). “The Role of Social Networks in Information Diffusion.” In: *WWW ’12: Proceedings of the 21st international conference on World Wide Web*.

Bala, Venkatesh and Sanjeev Goyal (1998). “Learning from Neighbours.” In: *Review of Economic Studies* 65.3, pp. 595–621.

Ballester, Coralio, Antoni Calvó-Armengol, and Yves Zenou (2006). “Who’s Who in Networks. Wanted: The Key Player.” In: *Econometrica* 74.5, pp. 1403–1417.

Bergstrom, Theodore, Lawrence Blume, and Hal Varian (1986). “On the private provision of public goods.” In: *Journal of Public Economics* 29.1, pp. 25–49.

Bindel, D, Jon M Kleinberg, and S Oren (2011). “How Bad is Forming Your Own Opinion?” In: *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pp. 57–66.

Bosagh Zadeh, Reza et al. (2013). “On the Precision of Social and Information Networks.” In: *Proceedings of the First ACM Conference on Online Social Networks*. COSN ’13. Boston, Massachusetts, USA: ACM, pp. 63–74. ISBN: 978-1-4503-2084-9. DOI: 10.1145/2512938.2512955. URL: http://doi.acm.org/10.1145/2512938.2512955.

Boyd, Danah, Scott Golder, and Gilad Lotan (2010). “Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter.” In: *System Sciences (HICSS), 2010 43rd Hawaii International Conference*. Vol. 0. Honolulu, HI: IEEE, pp. 1–10. ISBN: 978-1-4244-5509-6. DOI: 10.1109/HICSS.2010.412. URL: http://dx.doi.org/10.1109/HICSS.2010.412.

Bramoullé, Yann and Rachel Kranton (2006). “Public goods in networks.” In: *Journal of Economic Theory* 135.1, pp. 478–494.

Bramoullé, Yann et al. (2014). “Strategic interaction and networks.” In: *American Economic Review* 104.3, pp. 898–930.

Brouwer, Andries E and Willem H Haemers (2012). *Spectra of graphs*. Springer.

Cha, Meeyoung et al. (2009). “Analyzing the video popularity characteristics of large-scale user generated content systems.” In: *IEEE/ACM Transactions on Networking (TON* 17.5, pp. 1357–1370.

Cha, Meeyoung et al. (2010). “Measuring User Influence in Twitter: The Million Follower Fallacy.” In: *Proceedings of the International Conference Weblogs and Social Media (ICWSM)*.

Cha, Meeyoung et al. (2012). “The World of Connections and Information Flow in Twitter.” In: *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 42.4, pp. 991–998.

Chebolu, P. and P Melsted (2008). “PageRank and the random surfer model.” In: *Proc. of ACM-SIAM SODA’08*. San Francisco, USA.

Cheng, Justin et al. (2014). “Can cascades be predicted?” In: *WWW ’14: Proceedings of the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee.

Chessa, Michela, Jens Grossklags, and Patrick Loiseau (2015). “A Game-Theoretic Study on Non-Monetary Incentives in Data Analytics Projects with Privacy Implications.” In: *CoRR* abs/1505.02414. URL: `http://arxiv.org/abs/1505.02414`.

DeMarzo, P M, D Vayanos, and J Zwiebel (2003). “Persuasion Bias, Social Influence, and Unidimensional Opinions.” In: *The Quarterly Journal of Economics* 118.3, pp. 909–968.

Degroot, Morris H (1974). “Reaching a Consensus.” In: *Journal of the American Statistical Association* 69.345, pp. 118–121.

Dow, P Alex, Lada A Adamic, and Adrien Friggeri (2013). “The Anatomy of Large Facebook Cascades.” In: *Proceedings of the International Conference Weblogs and Social Media (ICWSM)*.

Easley, David and Arpita Ghosh (2013). “Incentives, gamification, and game theory: an economic approach to badge design.” In: *EC ’13: Proceedings of the fourteenth ACM conference on Electronic commerce*. ACM Request Permissions.

Eckersley, Peter (2010). “How unique is your web browser?” In: *Privacy Enhancing Technologies*, pp. 1–18.

Farahat, Ayman and Michael C Bailey (2012). “How effective is targeted advertising?” In: *WWW ’12: Proceedings of the 21st international conference on World Wide Web*. ACM Request Permissions.

Feick, Lawrence F and Linda L Price (1987). “The Market Maven: A Diffuser of Marketplace Information.” In: *Journal of Marketing* 51.1, pp. 83–97.

Füredi, Z. and J. Komlós (1981). “The eigenvalues of random symmetric matrices.” In: *Combinatorica* 1, pp. 233–241. ISSN: 02099683. DOI: `10.1007/BF02579329`. URL: `http://dx.doi.org/10.1007/BF02579329`.

Gabielkov, Maksym et al. (2016). “Social Clicks: What and Who Gets Read on Twitter?” In: *ACM SIGMETRICS / IFIP Performance 2016*. Antibes Juan-les-Pins, France. URL: `https://hal.inria.fr/hal-01281190`.

Galeotti, A and Sanjeev Goyal (2010). “The law of the few.” In: *American Economic Review* 100.4, pp. 1468–1492.

Galton, Francis (1907). “Vox Populi.” In: *Nature* 75.1949, pp. 450–451.

Ghaderi, Javad and R Srikant (2013). “Opinion Dynamics in Social Networks: A Local Interaction Game with Stubborn Agents.” In: *ACC ’13: Proceedings of the American Control Conference*. arXiv: 1208.5076v1 [cs.GT].

Ghosh, Arpita, Satyen Kale, and Preston McAfee (2011). “Who moderates the moderators?: crowdsourcing abuse detection in user-generated content.” In: *the 12th ACM conference*, pp. 167–176.

Ghosh, Arpita and Preston McAfee (2011). “Incentivizing high-quality user-generated content.” In: *WWW ’11: Proceedings of the 20th international conference on World wide web*. ACM Request Permissions.

— (2012). “Crowdsourcing with Endogenous Entry.” In: *WWW ’12: Proceedings of the 21st international conference on World Wide Web*.

Gill, Phillipa et al. (2013). “Follow the money: understanding economics of online aggregation and advertising.” In: *IMC ’13: Proceedings of the 13th ACM SIGCOMM conference on Internet measurement*.

Goel, Sharad, Duncan J Watts, and Daniel G Goldstein (2012). “The structure of online diffusion networks.” In: *EC ’12: Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM Request Permissions.

Goel, Sharad et al. (2010). “Anatomy of the long tail: ordinary people with extraordinary tastes.” In: *WSDM ’16: Proceedings of the ninth ACM international conference on Web search and data mining*.

Goel, Sharad et al. (2016). “The Structural Virality of Online Diffusion.” In: *Management Science* 62.1, pp. 180–196. DOI: 10.1287/mnsc.2015.2158. URL: http://dx.doi.org/10.1287/mnsc.2015.2158.

Goldstein, D G, R P McAfee, and Siddarth Suri (2014). “The wisdom of smaller, smarter crowds.” In: *EC ’14: Proceedings of the fifteenth ACM conference on Economics and computation*.

Goldstein, D.G., S. Goel, and D.J. Watts (2015). *System for tracking diffusion*. US Patent 8,990,341. URL: https://www.google.com/patents/US8990341.

Golub, Benjamin and Matthew O Jackson (2010). “Naive learning in social networks and the wisdom of crowds.” In: *American Economic Journal: Microeconomics* 2.1, pp. 112–149.

Gomez-Rodriguez, Manuel, Jure Leskovec, and Andreas Krause (2012). “Inferring Networks of Diffusion and Influence.” In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 5.4.

Hubert, A. B., T. Hubert, and C. Mugizi (2006). “A Random-Surfer Web-Graph Model.” In: *Proc. of ANALCO'06*. New York, USA.

Ioannidis, Stratis and P Loiseau (2013). “Linear regression as a non-cooperative game.” In: *Web and Internet Economics*.

Isaacman, Sibren et al. (2011). “Distributed rating prediction in user generated content streams.” In: *RecSys '11: Proceedings of the fifth ACM conference on Recommender systems*. ACM Request Permissions.

Jones, Eric, Travis Oliphant, Pearu Peterson, et al. (2001–). *SciPy: Open source scientific tools for Python*. [Online; accessed ¡today¿]. URL: `http://www.scipy.org/`.

Kamath, Krishna Y et al. (2013). “Spatio-temporal dynamics of online memes: a study of geo-tagged tweets.” In: *WWW '13: Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee.

Katz, Elihu (1957). “The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis.” In: *Public Opinion Quarterly* 21.1, p. 61.

Kempe, David, Jon M Kleinberg, and Éva Tardos (2003). “Maximizing the spread of influence through a social network.” In: *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*.

Kittur, A et al. (2007). “Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie.” In: *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*.

Kumar, Ravi, Jasmine Novak, and Andrew Tomkins (2010). “Structure and evolution of online social networks.” In: *Link Mining: Models, Algorithms, and Applications*, pp. 337–357.

Kwak, Haewoon et al. (2010). “What is Twitter, a social network or a news media?” In: *WWW '10: Proceedings of the 19th international conference on World wide web*. ACM.

Kwon, Sejeong and Meeyoung Cha (2014). “Modeling Bursty Temporal Pattern of Rumors.” In: *Proceedings of the International Conference Weblogs and Social Media (ICWSM)*.

Leskovec, Jure, Lada A Adamic, and Bernardo A Huberman (2007). “The dynamics of viral marketing.” In: *ACM Transactions on the Web* 1.1, 5–es.

Leskovec, Jure, Ajit Singh, and Jon M Kleinberg (2006). “Patterns of Influence in a Recommendation Network.” In: *Advances in Knowledge Discovery and . . .* Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 380–389.

Leskovec, Jure et al. (2007). “Cascading Behavior in Large Blog Graphs Patterns and a Model.” In: *Proc. SIAM International Conference on Data Mining*.

Leskovec, Jure et al. (2008). “Microscopic evolution of social networks.” In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 462–470.

Li, Yanhua et al. (2013). “Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships.” In: *WSDM ’13: Proceedings of the sixth ACM international conference on Web search and data mining*. ACM Request Permissions.

Liu, D. C. and J. Nocedal (1989). “On the Limited Memory BFGS Method for Large Scale Optimization.” In: *Math. Program.* 45.3, pp. 503–528. ISSN: 0025-5610. DOI: 10.1007/BF01589116. URL: http://dx.doi.org/10.1007/BF01589116.

Lorenz, M O (1905). “Methods of measuring the concentration of wealth.” In: *Publications of the American Statistical Association* 9.70, pp. 209–219.

May, Avner et al. (2014). “Filter & Follow: How Social Media Foster Content Curation.” In: *SIGMETRICS ’14: Proceedings of the ACM International conference on Measurement and modeling of computer systems*. New York, New York, USA: ACM Press, pp. 43–55.

McMahan, H B et al. (2013). “Ad Click Prediction: a View from the Trenches.” In: *KDD ’16: Proceedings of the 22th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Meeder, Brendan et al. (2011). “We know who you followed last summer: inferring social link creation times in twitter.” In: *WWW ’11: Proceedings of the 20th international conference on World wide web*. ACM Request Permissions.

Mislove, Alan et al. (2010). "You are who you know: inferring user profiles in online social networks." In: *WSDM '16: Proceedings of the ninth ACM international conference on Web search and data mining*, pp. 251–260.

Montjoye, Yves-Alexandre de et al. (2013). "Unique in the Crowd: The privacy bounds of human mobility." In: *Scientific Reports* 3.

Myers, SA, C Zhu, and Jure Leskovec (2012). "Information Diffusion and External Influence in Networks." In: *KDD '12: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. scholar: 5B22CD56-A65B-43CA-95AB-4E21D8FE35A6.

Narayanan, Arvind, Elaine Shi, and Benjamin I. P. Rubinstein (2011). "Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge." In: *CoRR* abs/1102.4374. URL: http://arxiv.org/abs/1102.4374.

Narayanan, Arvind and V Shmatikov (2008). "Robust De-anonymization of Large Sparse Datasets." In: *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pp. 111–125.

Narayanan, Arvind and Vitaly Shmatikov (2009). "De-anonymizing Social Networks." In: *Security and Privacy, 2009 30th IEEE Symposium on*, pp. 173–187.

Olejnik, Lukasz, Tran Minh-Dung, and Claude Castelluccia (2014). "Selling off privacy at auction." In: *In Proceedings of the Network and Distributed System Security Symposium (NDSS)*.

Pedarsani, Pedram and Matthias Grossglauser (2011). "On the privacy of anonymized networks." In: *KDD '11: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Request Permissions, pp. 1235–1243.

Pinto, Henrique, Jussara M Almeida, and Marcos A Gonçalves (2013). *Using early view patterns to predict the popularity of youtube videos*. New York, New York, USA: ACM.

Ramachandran, Arthi and Augustin Chaintreau (2015a). "The Network Effect of Privacy Choices." In: *SIGMETRICS Perform. Eval. Rev.* 43.3, pp. 59–62. ISSN: 0163-5999. DOI: 10.1145/2847220.2847240. URL: http://doi.acm.org/10.1145/2847220.2847240.

— (2015b). "Who Contributes to the Knowledge Sharing Economy?" In: *Proceedings of the 2015 ACM on Conference on Online Social Networks*. COSN '15. Palo Alto, California, USA: ACM, pp. 37–48. ISBN: 978-1-4503-3951-3. DOI: 10.1145/2817946.2817963. URL: http://doi.acm.org/10.1145/2817946.2817963.

Richardson, Matthew, Ewa Dominowska, and Robert Ragno (2007). “Predicting clicks: estimating the click-through rate for new ads.” In: *WWW ’07: Proceedings of the 16th international conference on World Wide Web*. ACM.

Rizoiu, Marian-Andrei et al. (2017). “Expecting to Be HIP: Hawkes Intensity Processes for Social Media Popularity.” In: *Proceedings of the 26th International Conference on World Wide Web*. WWW ’17. Perth, Australia: International World Wide Web Conferences Steering Committee, pp. 735–744. ISBN: 978-1-4503-4913-0. DOI: 10.1145/3038912.3052650. URL: https://doi.org/10.1145/3038912.3052650.

Rodrigues, Tiago et al. (2011). “On word-of-mouth based discovery of the web.” In: *IMC ’11: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM Request Permissions.

Rodriguez, Manuel Gomez, David Balduzzi, and Bernhard Schölkopf (2011). “Uncovering the temporal dynamics of diffusion networks.” In: *Proceedings of ICML*.

Rodriguez, Manuel Gomez, Jure Leskovec, and Bernhard Schölkopf (2013). “Structure and dynamics of information pathways in online media.” In: *WSDM ’13: Proceedings of the sixth ACM international conference on Web search and data mining*. ACM Request Permissions.

Roesner, Franziska, Tadayoshi Kohno, and David Wetherall (2012). “Detecting and defending against third-party tracking on the web.” In: *NSDI’12: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association.

Rusinowska, Agnieszka et al. (2011). “Social Networks: Prestige, Centrality, and Influence.” In: *Relational and Algebraic Methods in Computer Science: 12th International Conference, RAMICS 2011, Rotterdam, The Netherlands, May 30 – June 3, 2011. Proceedings*. Ed. by Harrie de Swart. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 22–39. ISBN: 978-3-642-21070-9. DOI: 10.1007/978-3-642-21070-9_2. URL: http://dx.doi.org/10.1007/978-3-642-21070-9_2.

Sharad, Kumar and George Danezis (2014). “An Automated Social Graph De-anonymization Technique.” In: *WPES ’14: Proceedings of the 13th Workshop on Privacy in the Electronic Society*. ACM Request Permissions.

Sharma, Naveen Kumar et al. (2012). “Inferring who-is-who in the Twitter social network.” In: *WOSN ’12: Proceedings of the 2012 ACM workshop on Workshop on online social networks*. ACM Request Permissions.

Shen, Huawei et al. (2014). *Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes*. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8370.

Szabo, Gabor and Bernardo A Huberman (2010). “Predicting the popularity of online content.” In: *Communications of the ACM* 53.8.

Takács, G et al. (2009). “Scalable collaborative filtering approaches for large recommender systems.” In: *The Journal of Machine Learning Research* 10, pp. 623–656.

Wang, G et al. (2013). “Wisdom in the social crowd: an analysis of quora.” In: *WWW ’13: Proceedings of the 22nd international conference on World Wide Web*.

Watts, Duncan J (2006). “Empirical analysis of an evolving social network.” In: *Science* 311.5757, p. 88.

Wu, Shaomei et al. (2011). “Who says what to whom on twitter.” In: *WWW ’11: Proceedings of the 20th international conference on World wide web*. ACM Request Permissions.

Yang, Jaewon and Jure Leskovec (2011). “Patterns of temporal variation in online media.” In: *WSDM ’11: Proceedings of the fourth ACM international conference on Web search and data mining*. ACM Request Permissions.

Zaman, Tauhid, Emily B Fox, and Eric T Bradlow (2014). “A Bayesian Approach for Predicting the Popularity of Tweets.” In: *Annals of Applied Statistics* 8.3, pp. 1583–1611.

Zhang, Yuchen et al. (2011). “User-click Modeling for Understanding and Predicting Search-behavior.” In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’11. San Diego, California, USA: ACM, pp. 1388–1396. ISBN: 978-1-4503-0813-7. DOI: 10.1145/2020408.2020613. URL: http://doi.acm.org/10.1145/2020408.2020613.

Zhao, Qingyuan et al. (2015). “SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity.” In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’15. Sydney, NSW, Australia: ACM, pp. 1513–1522. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2783401. URL: http://doi.acm.org/10.1145/2783258.2783401.