The Cost of Sharing Information in a Social World

Thesis proposal

Arthi Ramachandran
Department of Computer Science
Columbia University
arthir@cs.columbia.edu

October 31, 2016

Abstract

Big data is becoming a ubiquitous term used to indicate challenges in modeling data because of scale and complexity of the data. 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, a crucial question that big data allows us to examine is to model and analyze the flow of information within the network.

The first part of this thesis proposal discusses methods to 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 proposal characterizes the structural bias in a network by modeling the underlying information flow as a commodity of trade.

Contents

1 Introduction

"Big Data" is an increasingly common term used to refer to massive datasets which often have complex relationships, both between participants and between different types of data. While some techniques used to to analyze these datasets are mature methods, many current analyses involve trying to model and understand these complex relationships. Indeed, these relationships act as connectors between different types of information, allowing us to leverage data from different sources.

One key question that current research struggles with is understanding the flow of information in such a system, as well as the biases that result. Our work focuses on these questions in the context of a social network, where information flow is between participants in the social network. Social networks are now one of the major sources of web referrals [1]. Through the connected sharing that occurs, they function as content disseminators as well as content filters. This flow of information through a social network results in a dynamic that both provides more information and adds a layer of complexity in the analysis.

In this proposal, we tackle two major themes (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?

1.1 Cross Domain Analysis

The first theme we study is '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. 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 indentification.

The use of pseudonyms and semi-anonymous user names in social networks give a sense of anonymity to users of the network. However, we show that, in fact, this is not the case. Combining information from Twitter with clicking behavior, we show that with relatively little information, we can identify the twitter user (Section 3; published in *Conference for Online Social Networks 2014*).

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. On a macro-scale, we show that these can inferred from publicly available data (Section 4; published in *Sigmetrics 2016*, published in *Workshop on News and Public Opinion 2016* and submitted to *World Wide Web 2017*).

We address the following broad questions in these papers, with the goal of leverage publicly available information to learn private information:

- How unique is a user's content? (Section 3.2)
- How can we use a user's uniqueness to deanonymize a user from their click behavior? (Section 3.3)
- 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 4.2)

[1] http://j.mp/1qHkuzi

- Can we leverage our observations to predict private information, such as the clicks a link obtains? (Section 4.3)

1.2 Structural Biases in Social Networks

Our second theme revolves around how information flows in a social network and the biases that result. At large scales of data, previously hard to detect hidden dynamics begin to emerge as observable phenomenon. 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. There is not a strict assignment of effort and reward but rather, this is distributed across the network. We show that this redistribution has non-negligible effects. There is bias based on the structure of the network and this allocation of effort can be related to your position in the network.

Further while the network is composed of millions of individual, each such individual acts on their own to maximize their individual goals. Hence, the behavior of the complex network is the result of the actions (algorithms) that each of these individuals follow. We build a model of information sharing where reward of reading and searching for information is socialized (Section 5; published in *Conference for Online Social Networks 2015*). In this setting, we show that structural features of the network result in polarized equilibria where some individuals can free-ride on others' efforts.

We build on this idea by extending the model to socialize the cost of privatizing information (Section 6; published in *Workshop on the Economics of Networks, Systems and Computation 2015*; submitted to *World Wide Web 2017*). 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.

We address the following broad questions in these papers, 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 5.2)
- Can we develop a model to reproduce our observations? What are the implications of such a model? (Section 5.3)
- When users seek to preserve privacy in the information they share, what happens with an evolving graph? Do users gain from greater access to information through their network of friends? (Section 6.2)
- Can we replicate our observations with models? (Section 6.3)

2 Data

In our studies to better understand the dynamics of information flow, we rely on several datasets:

- **KAIST** (see [22] for more details): This dataset contains the entire Twitter graph from August 2009 and consists of 8m users and 700m links. 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. Further, we filtered the tweets by news domains (*e.g.,* `nytimes.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`).

- **DIGG**: Within **KAIST**, we focused on the domain `digg.com`. This dataset consists of 216k unique URLs tweeted by 44k unique users. These users represent the population that displays some interest in the domain. The associated network is derived from the 52m links connecting these individuals.

- **NYT** (see [62] for more details): This dataset contains all the Twitter posts containing a URL from the nytimes.com domain during a full week of December 2011. In parallel, we crawled the follower-followee relationship at the same time in order to construct the URLs that each user received. The final dataset totals 346k unique users receiving a total 22m tweets with URL (including multiplicity). Of these, there are 70k unique links.

- **CLICKS** (see [33] for more details): This dataset concentrates on five domains of news media chosen among the most popular on Twitter: 3 news media channel BBC, CNN, and Fox News, one newspaper, The New York Times, and one strictly-online source, The Huffington Post. The dataset is comprised of two sources: Twitter, for information about sharing and impressions, and `bit.ly`, for click information. We focus only on the URLs shortened by `bit.ly`. We monitor the Twitter Spritzer stream, a 1% sample of Twitter's public statuses, for these media references. Then, for multiple time points, we query both Twitter and `bit.ly` to gather data about the impressions, shares and clicks at each time point. In all, over the course of a month, we have 39.3k urls.

- **BUZZFEED** (see [84] for more details): This dataset includes all tweets published by any of BuzzFeeds Twitter accounts over a 7-day period from February 23 to March 2, 2016 (4K tweets) and August 3-17, 2016. We focused on original tweets, excluding retweets of non-BuzzFeed user tweets, in order to preserve uniformity of type of data being shared. While Twitter Analytics provides information about the readership, that information is only available to the account owner. We used Twitters REST API to scrape all tweets published by BuzzFeed accounts and all public retweets of those tweets over the span of the same 7 days. To further aggregate publicly available readership data, we crawled `bit.ly` on the BuzzFeed URLs with a bit.ly shortened URL (2.4K tweets). We collected hourly data for the time periods of hours 1, 2, 3, 4, 5 8, 9 12, 13 24.

3 Social Media + Clicks = Identity

3.1 Background and Overview

The use of 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 [68, 69, 31] to identify or infer information about users of online services, and social media in particular [79, 64, 65].

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 [22, 53, 6, 75, 21, 88, 42]. Indeed, measuring and predicting the success of a cascade is still a matter of controversy [22, 10]. Validating those studies with individual data about *which* users clicked *which* links sent by *whom* requires data unavailable outside researchers at social media provider [90, 11]. This paper 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. (§ 3.2)
- 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

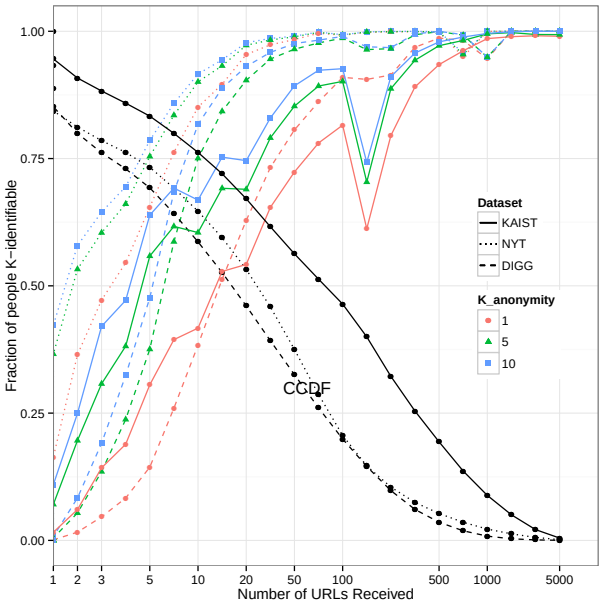


Figure 1: Fraction of social media users receiving more than n URLs, and those receiving a unique sets of URLs among them.

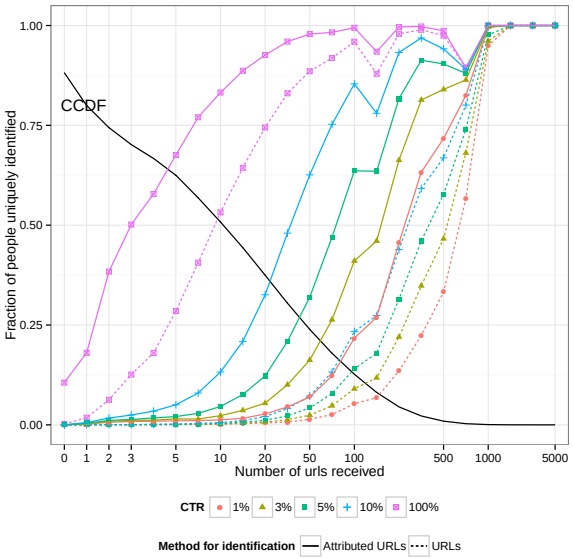


Figure 2: Fraction of users with at least n URLs received, and the proportion that are identified for various click generation rates and two methods.

method, using recent work on influence inference. When inference is accurate, the method promises identification in less than 4 or 5 clicks. (§ 3.3)

3.2 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[31]. 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 [68]. 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 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 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*[2]. 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 distribution [41]. 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.

3.3 Deanonymization Algorithm and Select Findings

Our results highlight a new promising domain of application of sparsity methods to identify social media users, based on the content that they receive and the individuals participating in it.

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.

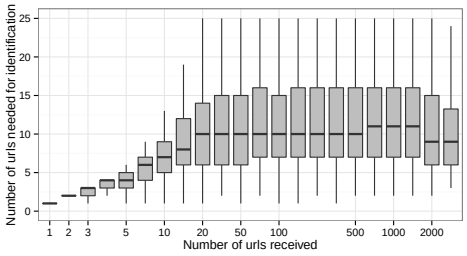


Figure 3: Distribution of the number of visits needed to successfully re-identify a node as a function of the size of its receiving URLs set.

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 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

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

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 `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 [74, 93]. 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 2 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.

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 [45, 77]. 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. 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 our paper, we conduct simulation assuming that the attribution steps succeeds with some probability in finding the source, or otherwise introduces an attribution error.

We applied our method to *NYT* and recovered a significant fraction of the individuals. Figure 2 compares the performance of the two methods – the baseline method using just URLs and the modified RAI method wth 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.

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$(v) \leftarrow (\cup_{f \text{ post } u \text{ \& } v \text{ visits } u \text{ via } f}$Followers$(f)) \cap$ 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$(v) \leftarrow (\cup_{f \text{ post } u}$Followers$(f)) \cap$ I(v);
        $u \leftarrow$ next visited url;
    end
end

4 Combining multiple social media sources to predict clicks

4.1 Background and Overview

When available data is lacking, inference methods applied to social networks can prove to be especially useful. One realm in which this is the case is in 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 propagations to quantify influence online [22, 10], how news sharing is affected by social networks [6, 7], and various mechanisms and drivers behind retweeting links [53, 18]. 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 [32, 63].

However, we show in this paper 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. (§4.2)

- Equipped with this insight, we develop a model which predicts temporal impressions, and then temporal clicks from those impression predictions. (§4.3)

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 [67] or using more complex models for the underlying diffusion [80, 94]. 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 [91], reformulating the problem into several stages [25], and classifying cascades based on their temporal evolution [89]. 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 [81]. 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?

4.2 Understanding Click Dynamics

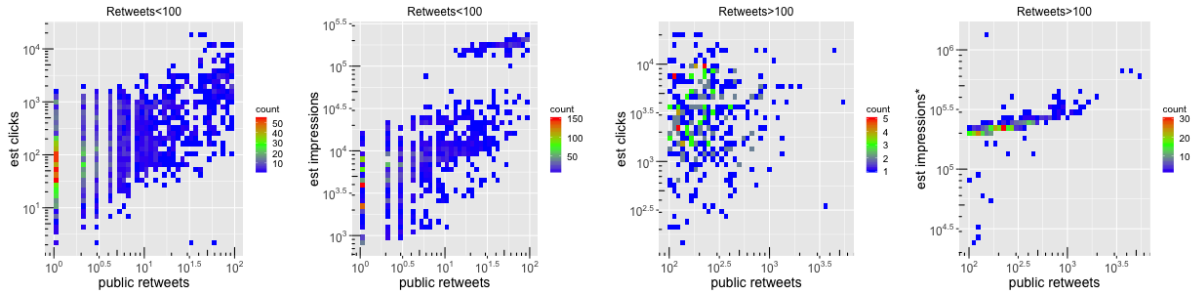
For online content, the click through rate (CTR) is defined as the probability that a reader clicks on the link when that it is shown to her. It has been studied and used in multiple applications including models of web surfing [46], ranking of search results [24], and optimization of online ads [74, 32]. 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 how we compute the audience size. 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, we consider 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 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 [**?**].

Figure 4 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.



(a) **Retweets** $<$ **100**. Strong positive correlation in both clicks and impressions

(b) **Retweets** $>$ **100**. Positive correlation in impressions, but no correlation in clicks.

Figure 5: **Retweets vs Estimated Clicks, Estimated Impressions** The effect of sharing on clicks and impressions, at different sharing magnitudes.

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 [**?**]. 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.

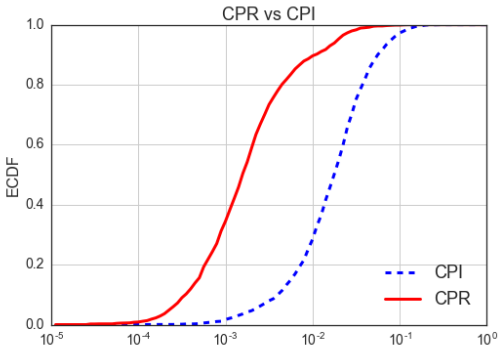


Figure 4: Clicks per Reception (Red) and Clicks per Impression (Blue). CPI is computed from the publisher dataset. CPR is computed entirely from public data.

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 $r = -0.089$, p value$= 1.70e^{-5}$). 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 setting, and define reach by re-shares rather than clicks [67].

We see a threshold effect at ~ 100 retweets, above which the clicking and sharing relationship changes. When number of retweets is < 100 (Figure 5a), clicks and shares are positively correlated (Pearson's $r = 0.530$). However, past this threshold, increasing retweets does not translate to increasing clicks (Figure 5b)(Pearson's $r = 0.060$, p value$= 0.242$). A diminishing impression

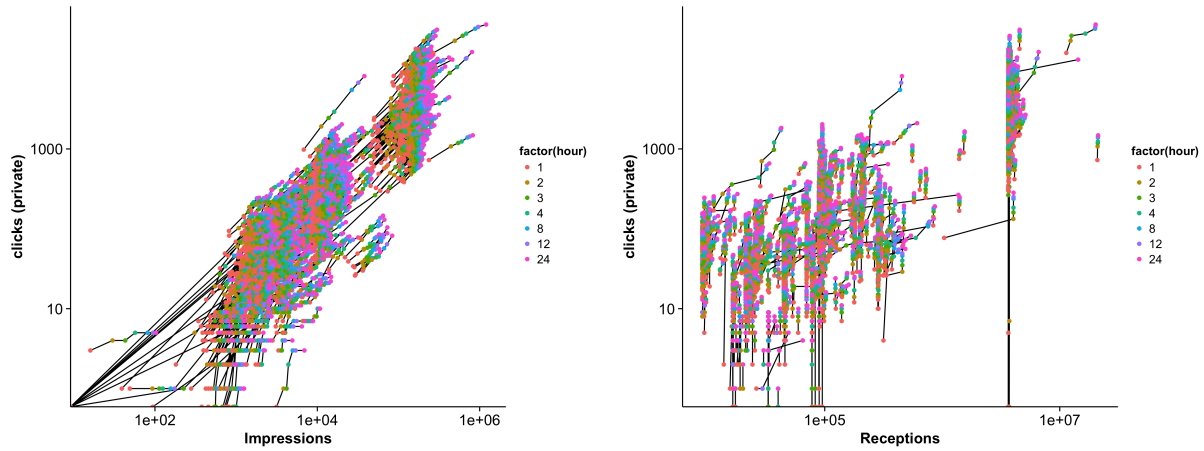


Figure 6: Trajectory plot for each URL with clicks on the y-axis and impressions (left) or receptions (right) on the x-axis. Each line represents a single URL with the colored points indicating the time of the measurement.

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 $r = 0.603$ (p value = $4.71e^{-39}$) (Figure 5b). 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.

To better understand the relationship between posting behavior and clicks or impressions, we study the evolution of the characteristics of a link evolve. Trajectory plots map the the evolution of each url (Figure 6). In these, each URL is represented as a line with each time point indicated with a different colour. Both the clicks and impressions are cumulative. In the impressions trajectory plot (left), we observe that most of the trajectories run roughly parallel to each other, indicating a close linear relationship between the two quantities. The difference when using private impression information vs public receptions is staggering. 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. The different clusters roughly correspond to the different buzzfeed accounts and we can see that overall, they exhibit similar behavior.

We also examine how the distribution of CPR and CPI changes with time (Figure 7). 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 7 (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.

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, 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.

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 §4.3) 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 §4.3) 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.

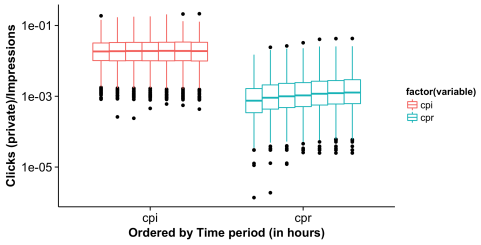


Figure 7: Clicks per Reception (Blue) and Clicks per Impression (Red) ordered by the time period. CPI is relatively stable over time, while CPR shows some increase.

4.3 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.

Predicting Impressions from Receptions 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 8), 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 ([**?**]).

We propose a simple memoryless model in which a fraction q of the receptions are 'activated' in each time period. 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_i be the number of receptions in hour i and y_i be the cumulative estimated number of impressions in hour i. Then we

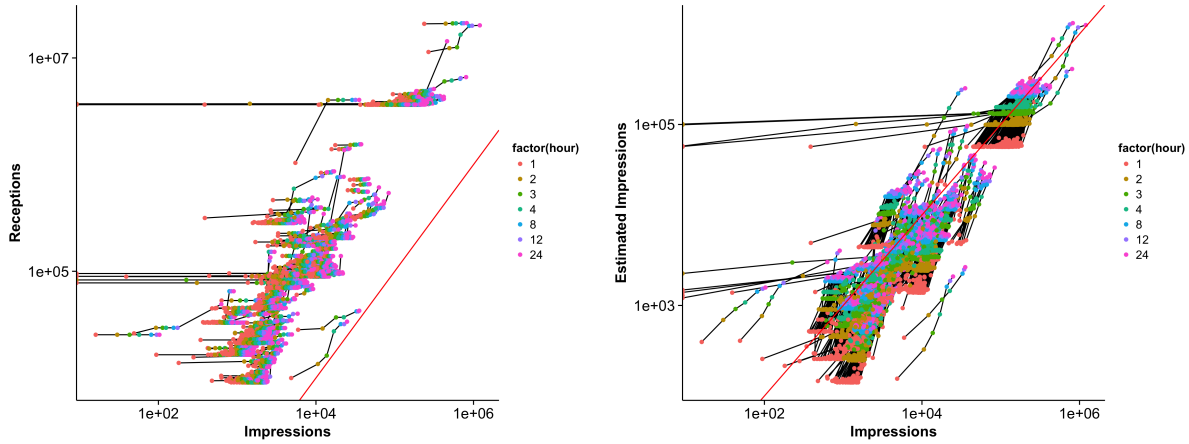


Figure 8: Relationship of Impressions to Estimated Impressions. Each line represents a single URL and each point represents the time period. The red line is the identity line for comparison.

have

$$y_i = \sum_{j=1}^{i} s \cdot (1 - q^{i-j+1}) x_j$$

In practice, we find that $q = 0.75$ and $s = 0.15$ work well.

Predicting Clicks from Impressions We derive an estimate of clicks for every hour period that we have an estimate of impressions. We chose $q = 0.75$ for our impressions estimate. Since clicks per impression remains rather stable over time (Figure 7), we estimate clicks by multiplying cumulative estimated impressions at hour i (y_i) with CPI at hour 1. Let z_i be the cumulative estimated number of clicks in hour i. Then

$$z_i = CPI_1 \cdot y_i$$

CPI_1 is the number of `bit.ly` clicks at hour i divided by y_1.

For completeness, we also explore as done in past work the possibility that the geometric transform for estimating impressions doesn't entirely capture the growth of impressions with respect to receptions. Therefore, we tested the addition of an exponential decay factor, similar to that used in [**?**] to model the temporal relaxation of retweeting dynamics. (*i.e.,* $z_i = CPI_1 \cdot y_i \cdot \exp(-i/\tau)$).

We evaluated the mean squared error (MSE) and median absolute percentage error (MAPE) with respect to actual (private) click data, for our different click estimate versions. As seen in Figure 9, the estimate of clicks with constant hour 1 CPI, with no decay factor added, actually has lower MSE and MAPE than untouched bit.ly clicks, across all time periods. It additionally has higher R^2 for all hours, than bit.ly clicks does. The exponential decays doesn't improve quality – when τ is low ($\tau = 48$) vastly underestimates clicks at later time periods, but converges to constant hour 1 click estimates when it is high ($\tau = [96, 384]$). These results are also consistent with the observed residuals, where our constant estimator which leverages estimated impressions has smaller residuals than untouched bit.ly clicks do, across all time periods (Figure 10).

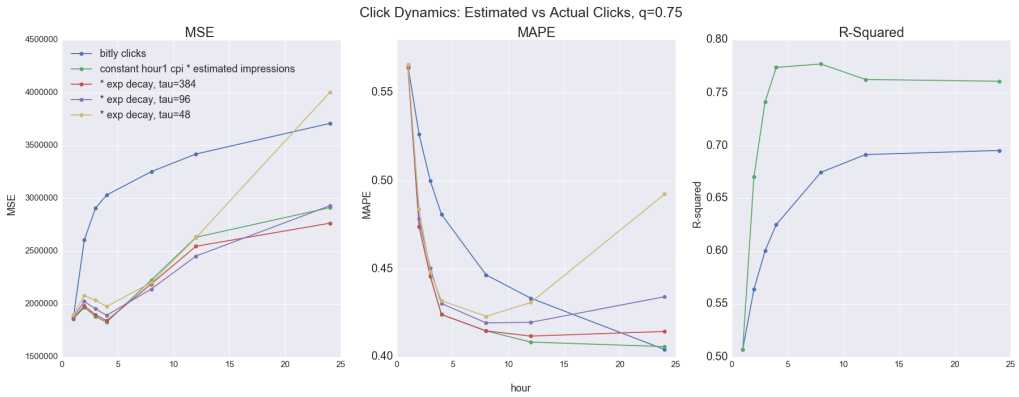


Figure 9: Evaluation of Click Estimates vs Actual Click Values. These click estimations are derived from impressions estimated with q=0.75. The constant-hour1-CPI-based estimation of clicks, along with its exponentially decayed counterparts, is compared to the raw bitly clicks over time. The metrics for goodness-of-fit evaluation are Mean Squared Error (Left), Median Absolute Percent Error (Center), and R^2 (Right).

Predicting Future Clicks In the previous sections, 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. We further develop a model using these estimated clicks and impressions to predict future clicks.

We created a linear model for predicting the total number of (private data) clicks on a URL, from just using our hour 1 (public data) estimates of clicks (1C) and impressions (1I). Using 20-fold cross validation and forward step-wise feature subset selection, we arrived at an average predictive R^2 of 0.80 and MAPE of 0.34%. This can be seen in Figure 11. The model is defined as:

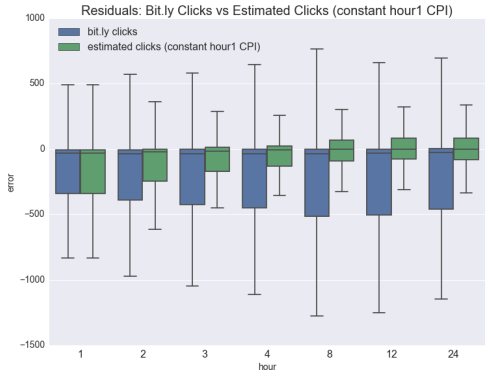
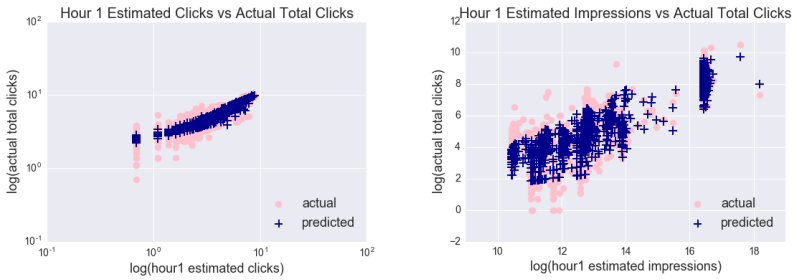


Figure 10: Click Estimation Residuals. These are the residuals between estimated clicks and actual clicks, by time period.

$$\log(TotalClicks) = \beta_0 + \beta_1 \log(1C) + \beta_2 \log(1I).$$

In fact, using the first hour of clicks *alone* proves predictive of future success ($R^2 = 0.76$, $MAPE = 40\%$). 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. The use of a log (base 10) transformation on all of the variables improved the R^2 and MAPE values, even given that we transform our log predictions back to linear space.



(a) Predicted and Actual Total Clicks vs Estimated Hour 1 Clicks

(b) Predicted and Actual Total Clicks vs Estimated Hour 1 Impressions

Model	**Hour1 Estimated Clicks (β_1)**	**Hour1 Estimated Impressions (β_2)**	**Intercept (β_0)**	**MAPE**	**R^2**
Hour 1 Est. Clicks Only	0.9386	0	1.6918	40%	0.76
Hour 1 Est. Clicks + Hour 1 Est. Imp.	0.7626	0.2296	-0.6689	34%	0.80

(c) Regressor Coefficients, Median Absolute Percent Error and R^2. The response and predictor variables included are all transformed by natural log. The R^2 and MAPE values reflect the goodness of fit of the log linear model's predictions transformed back into linear space.

Figure 11: **OLS Linear Regression for Predicting Total Clicks from Estimated Hour 1 Clicks and Impressions.**

4.4 Proposed Work

We have evaluated and verified our click prediction method primarily on BuzzFeed data. However, it remains to be seen how well this translates to prediction of other media sources. There are a few limitations to overcome in this application that we need to address. Firstly, most news media sources have CPR an order of magnitude lower than buzzfeed and hence might behave somewhat differently. Secondly, while the parameters for prediction worked well for the BuzzFeed data, we need to develop a method to learn them for other sources (without access to the intermediate private impressions). Lastly, from a data perspective, we only have public bit.ly clicks for other news sources. While public bit.ly clicks strongly correlate to the clicks observed by a publisher (from their own analytics feeds), this correlation isn't perfect. We need to account for this in our prediction – how does it change when we predict public information rather than private information?

In addition to further developing our own method, we need to develop comparisons to currently existing methods. While no methods currently exist to predict clicks (either dynamically or the final count) for media sources, there is a significant body of research dealing with predicting cascade sizes and shares. For those that can be extended to predicting clicks, it would be useful to see how our methods compare.

5 Specialization in Static Networks....

5.1 Background and Overview

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?

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* [15], 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 [4]. The general network case was studied more recently [13, 20, 19].Even for that simple case, predictions vastly differ: On the one hand, a study of small effects [19] 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 [20]. 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. 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 [21, 88]. However, the dynamics of participation and influence remains elusive. For instance, relying on number of followers to judge an influencer can be misleading [22, 10] and predicting who is successful at an individual level was shown to be generally unreliable [10].

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

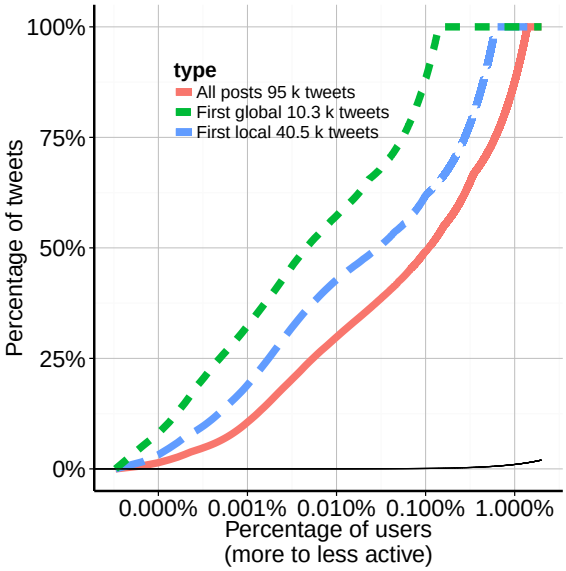


Figure 12: Lorenz curve (*i.e.,* cumulative share of the top x% nodes in the audience seen as a function of x) comparing production of tweets and original content for `cnn.com` from KAIST

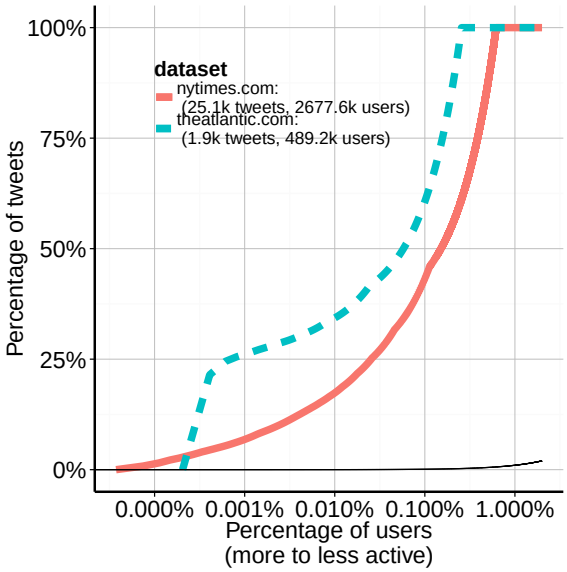


Figure 13: Lorenz curve for "first local tweets in two different domains.

rather than just the just well known people. In fact, counter-intuitively, original content production is skewed towards less active and connected people. (Section 5.2).

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. 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 5.3).

5.2 The existence of specialization

Imbalanced content creation Unsurprisingly, in social media like Twitter, a small fraction of users are responsible for a large part of the activity. We quantify this concentration using the NYTIMES dataset and use the Lorenz curve [60], or the cumulative share of the top x % of users as a function of x, in Figure 12. 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. Equiva-

lently, 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 repost 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.

Effect of Time Finally, we study the factors quantitatively affecting specialization. To take an example, first, we show in Figure 13 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.

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 14 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.

5.3 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

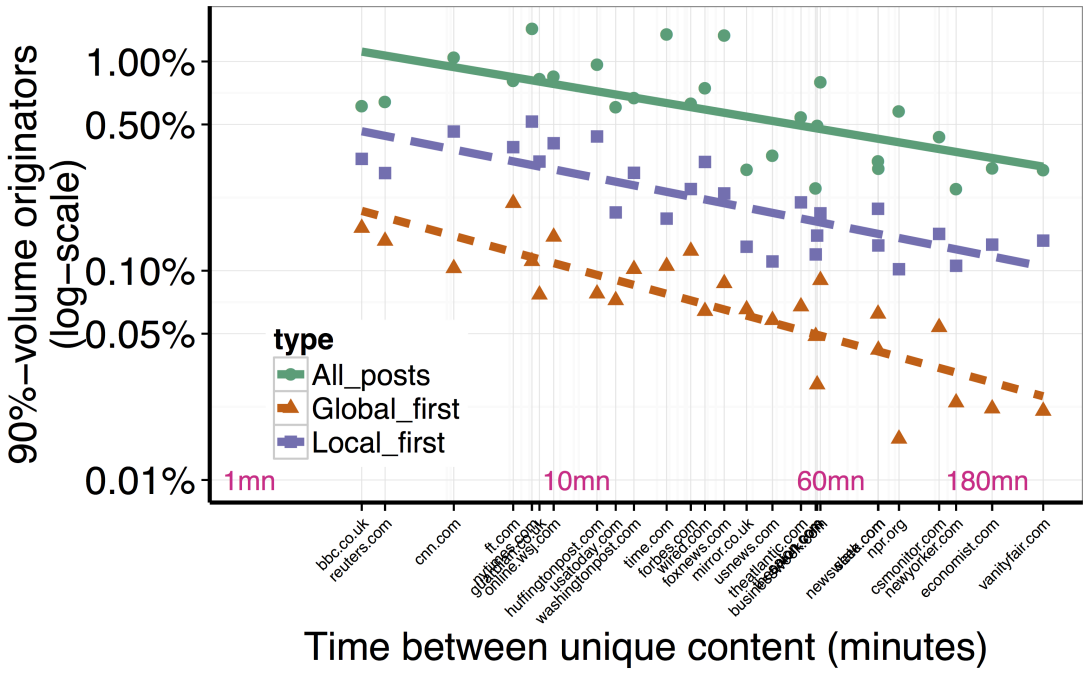


Figure 14: Concentration of sharing compared to the shelf-life for each media source. Each point is the fraction of the audience responsible for 90% of the tweet volume of the media source.

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 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. 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.

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). 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.

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 $\theta = 1$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 $x \mapsto x \exp(x)$.*

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*

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

The case of symmetric graphs is interesting because, as we show in Section 5.3, this symmetric equilibrium need not always be a unique or stable equilibrium.

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 $\gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ 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 satifies the condition: $1 + \frac{1}{\lambda_{\min}} < \gamma_i'(w_i + y_{-i}) < 1$ [4]. 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.

Theorem 3 (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*

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

The quantity $\hat{\tau}$ of G specifies the condition under which a unique Nash equilibrium exists. Table 1 details the value of $\hat{\tau}$ for various regular graphs ([73]).

Table 1: 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	$-2\sqrt{np}$	$\left(\frac{1}{2\sqrt{np}-1}\right)^{\frac{1}{2}} e^{\frac{1}{2(2\sqrt{np}-1)}}$
Star	$-\sqrt{n-1}$	$\left(\frac{1}{\sqrt{n-1}-1}\right)^{\frac{1}{2}} e^{\frac{1}{2(\sqrt{n-1}-1)}}$
Complete Bipartite	$-\frac{n}{2}$	$\left(\frac{2}{n-2}\right)^{\frac{1}{2}} e^{\frac{1}{n-2}}$

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 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.

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

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 [73] . 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 15), 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 asymetric graphs (figure 16). 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.

We see that specialization can occur as a result of the degree distribution (as in asymetric 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.

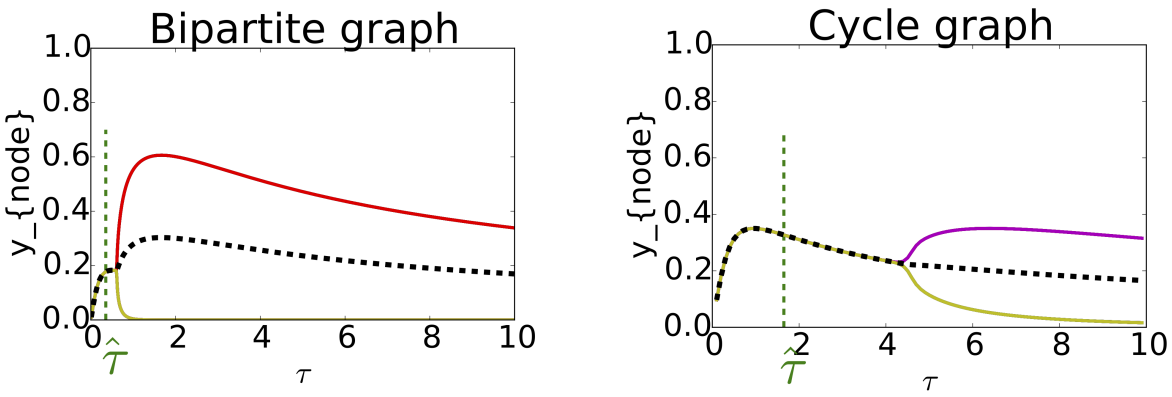


Figure 15: Differing effort levels in the Nash Equilibrium (y-axis) with different τ (x-axis) in symmetric graphs. Each node (of $n = 20$ nodes) is represented by a line in the figure. The unique equilibrium ($\tau < \hat{\tau}$) is always symmetric. (left) Complete bipartite graph (right) Cycle graph.

6 ...Leads to Biases in Dynamic Networks

6.1 Background and Overview

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 [35], to recent online applications like collaborative encyclopedia [51], question answering [83] or prediction games [43]. 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.

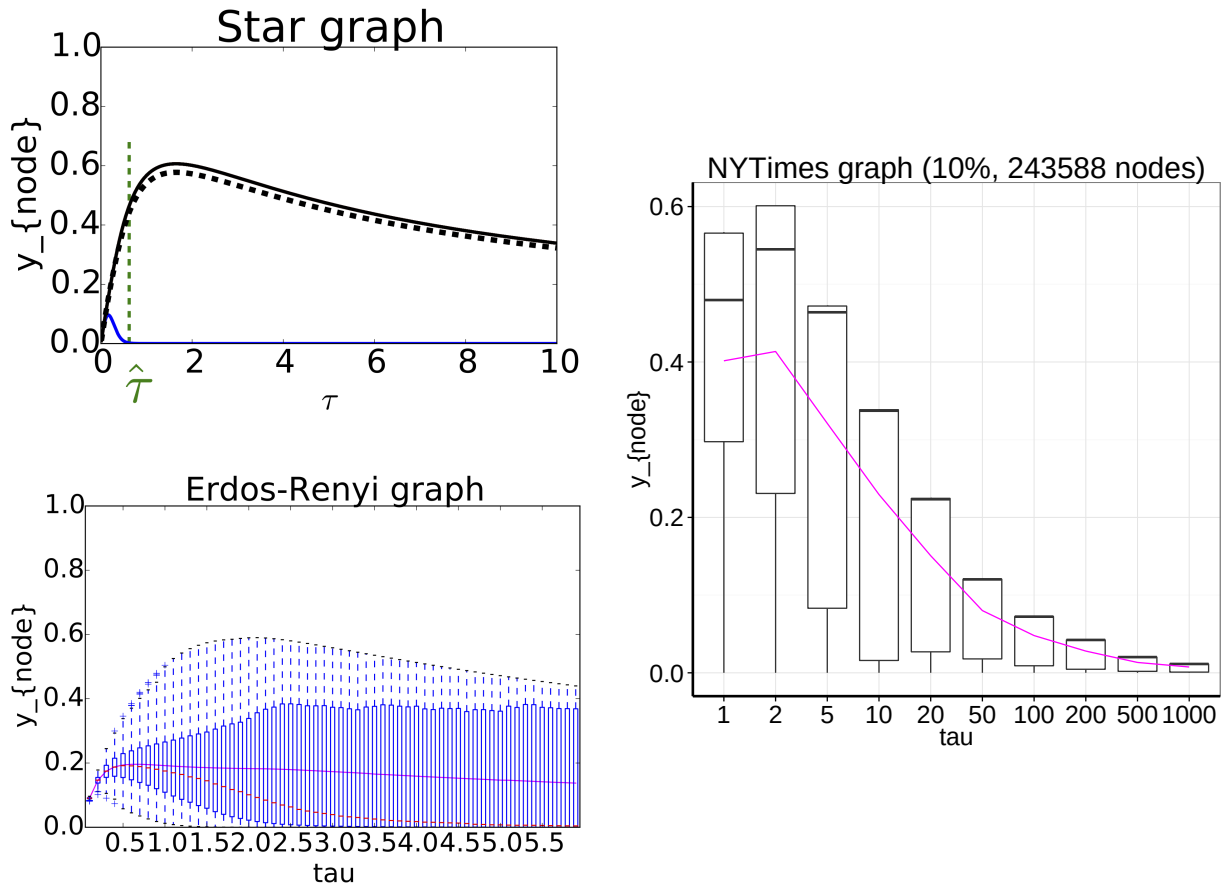


Figure 16: Differing effort levels with different τ in asymetric graphs. Each bar represents the distribution of the amount of effort by all the nodes. The pink line is the average effort of all the nodes. (top-left) Star graph. (bottom-left) Erdös-Renyi graph ($n = 1000, p = 0.01$). (right) Randomly sampled NYTimes graph with 243k nodes.

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 6.2).

- 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 6.3).

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 [38] and personalization [82, 48]. 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 crowdsourcing models [40, 39] 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 [47, 26, 71]. 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.

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, $\phi(\lambda) : \min_{\lambda_i} J_i(\lambda)$ 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 5. *In the estimation problem, the best response of an agent i, to its neighbors' effort is given by $\zeta_i^2 \lambda_i^\alpha = \frac{1}{C^2}$.*

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 paper, 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 [26]. 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 (which was shown in an earlier workshop version of this paper [71]).

Claim 6. *For a d-regular graph, a symmetric Nash Equilibrium always exists and is given by $\lambda^* = \left(\frac{1}{C^2 \cdot (d+1)^2}\right)^{\frac{1}{\alpha+2}}$. Moreover, if the d is increasing in the size of the graph, wisdom of the private crowd always exists.*

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 6.3).

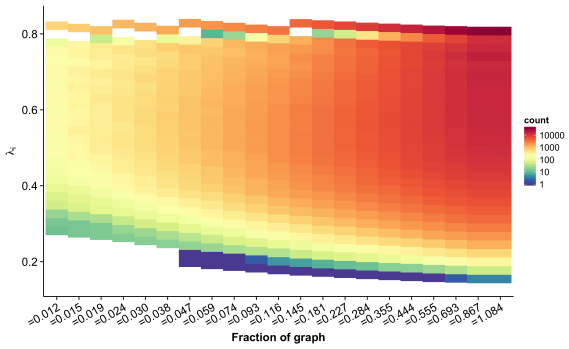


Figure 17: Heatmap of distribution of λ_i for increasing large fractions of the graph.

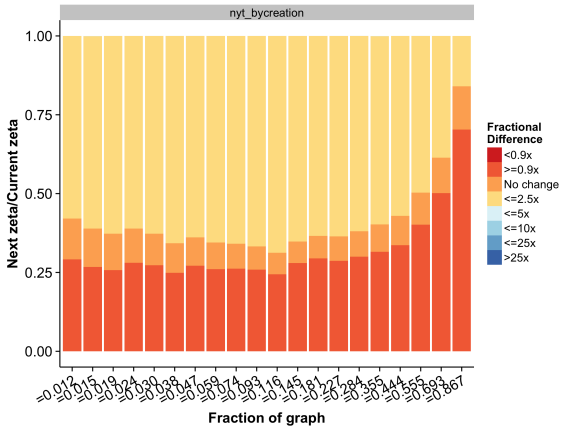


Figure 18: Fractional change in ζ_i from one stage to the next. We see that with larger graphs, there are, in fact, more nodes who lose than gain.

6.2 Evolving Social Graphs

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 [72]. 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).

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 17). 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 18). 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 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 19). 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 20). 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.

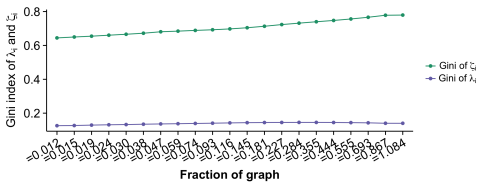


Figure 19: Gini index measured of the distributions of λ_i and ζ_i. The gini index is a measure of inequality – lower values indicate less inequality.

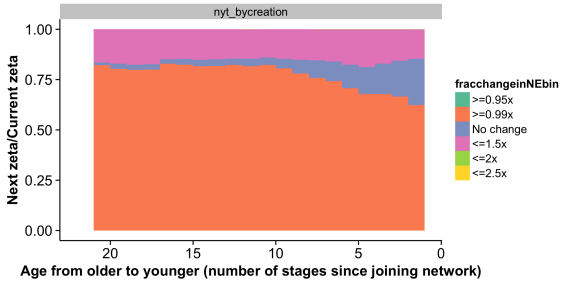


Figure 20: Fractional change in ζ_i from the penultimate stage to the last stage, split by the relative twitter age on the x-axis.

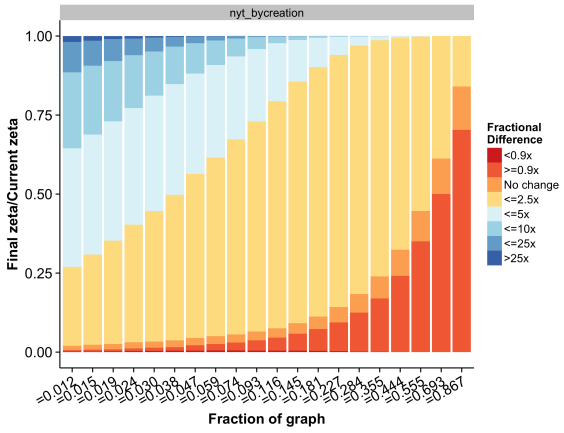


Figure 21: Fractional change from current stage to the final network.

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 21 shows the change in a node's total precision from each stage to the final graph. 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.

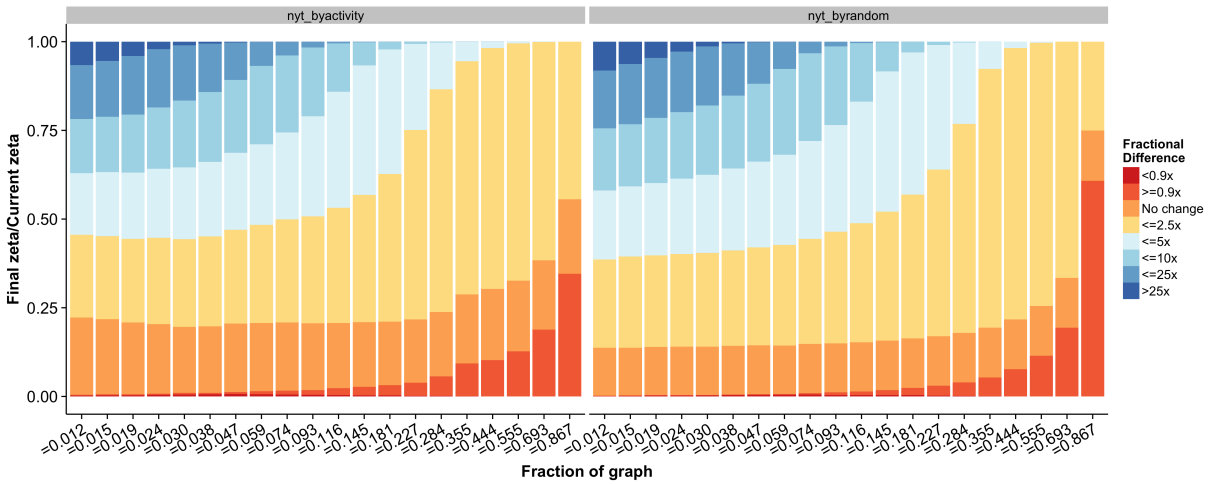


Figure 22: Fractional change from current stage till the final stage for the graph evolving by activity (right) and randomly (left).

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 18) and when it becomes active in the network (Figure 22). 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.

6.3 Network Structure and Inequality

Bipartite Graph In networked public goods, the bipartite graph proves to be especially interesting. As we will see in §6.3, 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 7. *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.*

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.

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 [3, 4, 20, 19].

Theorem 8. *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 and have $|\mu_{min}|$ as an increasing function of $|G_t|$, as $|G_t| \to \infty$, the crowd is privately wise.*

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 μ_{min}. For many graph families, $|\mu_{min}|$ increases with the size of the graph (*e.g.,* $\mu_{min} = -d$ for a d-regular graph). Thus we see that the condition of network normality is sufficient for the crowd to be privately wise.

6.4 Proposed Work

We see that segregation in networks clearly leads to suboptimal outcomes for certain sections of the network. Indeed, we observe such outcomes in real world networks when a certain fraction of the population loses from contact with more individuals. One interesting question to examine is whether we can identify a segregation of real world graphs that lead to such an outcome. For example, are the different genders of Twitter users segregated enough to explain which are the users who lose in the expanding network?

Another question to further explore is how robust these results are for other networks. We have observed in the NYTIMES dataset that the structure of evolution has a bearing on the characteristics of nodes who gain or lose with the graph expansion. We expect that the evolution other networks occurs in a similar manner as in NYTIMES, but we would like to solidify these results in networks such as citation networks, mobility networks or even other social networks such as instagram or pinterest.

7 Research plan

Table 2 shows my plan for completion of research. Thus, I plan to defend my thesis in Apr 2016.

Timeline	Work
2011	Complete coursework
2012	Complete candidacy
2013	Complete teaching requirement
	Begin work on deanonymization
2014	Complete work on deanonymization
2015	Complete work on curation and specialization of information
2016	Complete work networked privacy
	Complete work on click prediction
	Complete Thesis Proposal
2017	Extend work on networked privacy and click prediction
	Write and defend thesis (Apr 2017)

Table 2: Plan for completion of my research

References

[1] Daron Acemoglu, M A Dahleh, I Lobel, and Asuman Ozdaglar. Bayesian learning in social networks. *The Review of Economic Studies*, 78(4):1201–1236, 2011.

[2] Daron Acemoglu, Asuman Ozdaglar, and A ParandehGheibi. Spread of (mis) information in social networks. *Games and Economic Behavior*, 70(2):194–227, 2010.

[3] Nizar Allouch. The Cost of Segregation in Social Networks. *SSRN Electronic Journal*, 2013.

[4] Nizar Allouch. On the Private Provision of Public Goods on Networks. *Journal of Economic Theory*, forthcoming:1–34, 2015.

[5] N Alon, M Feldman, O Lev, and M Tennenholtz. How Robust is the Wisdom of the Crowds? *IJCAI*, 2015.

[6] J An, Meeyoung Cha, Krishna Gummadi, and Jon Crowcroft. Media landscape in Twitter: A world of new conventions and political diversity. In *Proceedings of the International Conference Weblogs and Social Media (ICWSM)*, pages 18–25, 2011.

[7] J An, D Quercia, Meeyoung Cha, Krishna Gummadi, and Jon Crowcroft. Sharing political news: the balancing act of intimacy and socialization in selective exposure. *EPJ Data Science*, 2014.

[8] Jisun An, Daniele Quercia, and Jon Crowcroft. Recommending investors for crowdfunding projects. In *WWW '14: Proceedings of the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee, April 2014.

[9] S. Asur and B A Huberman. Predicting the Future with Social Media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, pages 492–499, 2010.

[10] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 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, February 2011.

[11] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada A Adamic. The Role of Social Networks in Information Diffusion. In *WWW '12: Proceedings of the 21st international conference on World Wide Web*, January 2012.

[12] Venkatesh Bala and Sanjeev Goyal. Learning from Neighbours. *Review of Economic Studies*, 65(3):595–621, July 1998.

[13] Coralio Ballester, Antoni Calvó-Armengol, and Yves Zenou. Who's Who in Networks. Wanted: The Key Player. *Econometrica*, 74(5):1403–1417, September 2006.

[14] Abhijit Banerjee. A Simple Model of Herd Behavior. *The Quarterly Journal of Economics*, 107(3):797–817, August 1992.

[15] Theodore Bergstrom, Lawrence Blume, and Hal Varian. On the private provision of public goods. *Journal of Public Economics*, 29(1):25–49, February 1986.

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

[17] J Bollen, H Mao, and X Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2011.

[18] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference*, volume 0, pages 1–10, Honolulu, HI, January 2010. IEEE.

[19] Yann Bramoullé, Y Bramoullé, R Kranton, Rachel Kranton, M D'Amours, and Martin D'amours. Strategic interaction and networks. *American Economic Review*, 104(3):898–930, 2014.

[20] Yann Bramoullé and Rachel Kranton. Public goods in networks. *Journal of Economic Theory*, 135(1):478–494, July 2006.

[21] Meeyoung Cha, F Benevenuto, H Haddadi, and Krishna Gummadi. The World of Connections and Information Flow in Twitter. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 42(4):991–998, 2012.

[22] Meeyoung Cha, H Haddadi, F Benevenuto, and Krishna Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the International Conference Weblogs and Social Media (ICWSM)*, 2010.

[23] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Networking (TON*, 17(5):1357–1370, 2009.

[24] P. Chebolu and P Melsted. Pagerank and the random surfer model. In *Proc. of ACM-SIAM SODA'08*, San Francisco, USA, 2008.

[25] Justin Cheng, Lada A Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *WWW '14: Proceedings of the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee, April 2014.

[26] Michela Chessa, Jens Grossklags, and Patrick Loiseau. A game-theoretic study on non-monetary incentives in data analytics projects with privacy implications. *CoRR*, abs/1505.02414, 2015.

[27] Blerim Cici, Athina Markopoulou, Enrique Frias-Martinez, and Nikolaos Laoutaris. Assessing the potential of ride-sharing using mobile and social data: a tale of four cities. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 201–211. ACM, 2014.

[28] R L Cross and A Parker. *The hidden power of social networks*. Harvard Business School Press, 2004.

[29] Pranav Dandekar, Ashish Goel, and David T Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796, April 2013.

[30] Abhimanyu Das, Sreenivas Gollapudi, Rina Panigrahy, and Mahyar Salek. Debiasing social wisdom. *the 19th ACM SIGKDD international conference*, pages 500–508, August 2013.

[31] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, 2013.

[32] Ayman Farahat and Michael C Bailey. How effective is targeted advertising? In *WWW '12: Proceedings of the 21st international conference on World Wide Web*. ACM Request Permissions, April 2012.

[33] Maksym Gabielkov, Arthi Ramachandran, Arnaud Legout, and Augustin Chaintreau. Social Clicks: What and Who Gets Read on Twitter? In *ACM SIGMETRICS / IFIP Performance 2016*, Antibes Juan-les-Pins, France, June 2016.

[34] A Galeotti and Sanjeev Goyal. The law of the few. *American Economic Review*, 100(4):1468–1492, 2010.

[35] Francis Galton. Vox Populi. *Nature*, 75(1949):450–451, March 1907.

[36] Gary L Geissler and Steve W Edison. Market Mavens' Attitudes Towards General Technology: Implications for Marketing Communications. *Journal of Marketing Communications*, 11(2):73–94, June 2005.

[37] Javad Ghaderi and R Srikant. Opinion Dynamics in Social Networks: A Local Interaction Game with Stubborn Agents. *ACC '13: Proceedings of the American Control Conference*, 2013.

[38] Arpita Ghosh, Satyen Kale, and Preston McAfee. Who moderates the moderators?: crowdsourcing abuse detection in user-generated content. *the 12th ACM conference*, pages 167–176, June 2011.

[39] Arpita Ghosh and Preston McAfee. Incentivizing high-quality user-generated content. In *WWW '11: Proceedings of the 20th international conference on World wide web*. ACM Request Permissions, March 2011.

[40] Arpita Ghosh and Preston McAfee. Crowdsourcing with Endogenous Entry. In *WWW '12: Proceedings of the 21st international conference on World Wide Web*, February 2012.

[41] Sharad Goel, Andrei Broder, Evgeniy Gabrilovich, and Bo Pang. Anatomy of the long tail: ordinary people with extraordinary tastes. *WSDM '16: Proceedings of the ninth ACM international conference on Web search and data mining*, February 2010.

[42] Sharad Goel, Duncan J Watts, and Daniel G Goldstein. The structure of online diffusion networks. In *EC '12: Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM Request Permissions, June 2012.

[43] D G Goldstein, R P McAfee, and Siddarth Suri. The wisdom of smaller, smarter crowds. *EC '14: Proceedings of the fifteenth ACM conference on Economics and computation*, 2014.

[44] Benjamin Golub and Matthew O Jackson. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010.

[45] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring Networks of Diffusion and Influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4), February 2012.

[46] A. B. Hubert, T. Hubert, and C. Mugizi. A random-surfer web-graph model. In *Proc. of ANALCO'06*, New York, USA, 2006.

[47] Stratis Ioannidis and P Loiseau. Linear regression as a non-cooperative game. *Web and Internet Economics*, 2013.

[48] Sibren Isaacman, Stratis Ioannidis, Augustin Chaintreau, and Margaret Martonosi. Distributed rating prediction in user generated content streams. In *RecSys '11: Proceedings of the fifth ACM conference on Recommender systems*. ACM Request Permissions, October 2011.

[49] Krishna Y Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. 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, May 2013.

[50] Elihu Katz. The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis. *Public Opinion Quarterly*, 21(1):61, 1957.

[51] A Kittur, E Chi, B A Pendleton, B Suh, and T Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007.

[52] I Kremer, Ilan Kremer, Yishay Mansour, Y Mansour, Motty Perry, and M Perry. Implementing the "Wisdom of the Crowd". *Journal of political Economy*, 2014.

[53] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*. ACM, April 2010.

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

[55] PF Lazarsfeld, Bernard Berelson, and Hazel Gaudet. *The peoples choice: how the voter makes up his mind in a presidential campaign.* Columbia University Press, 1948.

[56] Yabing Liu, C Kliman-Silver, R Bell, Balachander Krishnamurthy, and Alan Mislove. Measurement and analysis of osn ad auctions. *COSN '14: Proceedings of the 2nd ACM conference on Online social networks*, pages 139–150, 2014.

[57] Ilan Lobel, I Lobel, Evan Sadler, and E Sadler. Social learning and aggregate network uncertainty. In *EC '13: Proceedings of the fourteenth ACM conference on Electronic commerce*, page 677, New York, New York, USA, June 2013. ACM.

[58] Charles G Lord, Lee Ross, and Mark R Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098, November 1979.

[59] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025, May 2011.

[60] M O Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219, 1905.

[61] Travis Martin, Jake M Hofman, Amit Sharma, Ashton Anderson, and Duncan J Watts. Exploring limits to prediction in complex social systems. In *WWW '16: Proceedings of the 25th International Conference on World Wide Web*, February 2016.

[62] Avner May, Augustin Chaintreau, Nitish Korula, and Silvio Lattanzi. Filter & Follow: How Social Media Foster Content Curation. In *SIGMETRICS '14: Proceedings of the ACM International conference on Measurement and modeling of computer systems*, pages 43–55, New York, New York, USA, 2014. ACM Press.

[63] H B McMahan, G Holt, D Sculley, M Young, and D Ebner. Ad Click Prediction: a View from the Trenches. *KDD '16: Proceedings of the 22th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013.

[64] Brendan Meeder, Brian Karrer, Amin Sayedi, R Ravi, Christian Borgs, and Jennifer Chayes. 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, March 2011.

[65] Alan Mislove, B Viswanath, Krishna Gummadi, and P Druschel. You are who you know: inferring user profiles in online social networks. *WSDM '16: Proceedings of the ninth ACM international conference on Web search and data mining*, pages 251–260, 2010.

[66] Elchanan Mossel, Allan Sly, and Omer Tamuz. Asymptotic learning on Bayesian social networks. *Probability theory and related fields*, pages 1–31, 2012.

[67] SA Myers, C Zhu, and Jure Leskovec. Information Diffusion and External Influence in Networks. *KDD '12: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012.

[68] Arvind Narayanan and V Shmatikov. Robust De-anonymization of Large Sparse Datasets. *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125, 2008.

[69] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing Social Networks. *Security and Privacy, 2009 30th IEEE Symposium on*, pages 173–187, 2009.

[70] K Olmstead, A Mitchell, and Tom Rosenstiel. *Navigating News Online: Where people go, how they get there, and what lures them away*. Pew Research Center's Project for Excellence in Journalism, 2011.

[71] Arthi Ramachandran and Augustin Chaintreau. The network effect of privacy choices. *SIGMETRICS Perform. Eval. Rev.*, 43(3):59–62, November 2015.

[72] Arthi Ramachandran and Augustin Chaintreau. The Network Effect of Privacy Choices. *Proceedings of ACM NETECON Workshop*, 43(3):59–62, November 2015.

[73] Arthi Ramachandran and Augustin Chaintreau. Who contributes to the knowledge sharing economy? In *Proceedings of the 2015 ACM on Conference on Online Social Networks*, COSN '15, pages 37–48, New York, NY, USA, 2015. ACM.

[74] Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*. ACM, May 2007.

[75] Tiago Rodrigues, Fabrício Benevenuto, Meeyoung Cha, Krishna Gummadi, and Virgílio Almeida. 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, November 2011.

[76] Manuel Gomez Rodriguez, David Balduzzi, and Bernhard Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *Proceedings of ICML*, 2011.

[77] Manuel Gomez Rodriguez, Jure Leskovec, and Bernhard Schölkopf. 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, February 2013.

[78] Adam Sadilek and Henry Kautz. Modeling the impact of lifestyle on health at scale. In *WSDM '13: Proceedings of the sixth ACM international conference on Web search and data mining*. ACM Request Permissions, February 2013.

[79] Naveen Kumar Sharma, Saptarshi Ghosh, Fabrício Benevenuto, Niloy Ganguly, and Krishna Gummadi. 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, August 2012.

[80] Huawei Shen, Dashun Wang, Chaoming Song, and Albert-Lszl Barabsi. Modeling and predicting popularity dynamics via reinforced poisson processes, 2014.

[81] Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8), August 2010.

[82] G Takács, I Pilászy, B Németh, and D. Tikk. Scalable collaborative filtering approaches for large recommender systems. *The Journal of Machine Learning Research*, 10:623–656, 2009.

[83] G Wang, K Gill, M Mohanlal, and H Zheng. Wisdom in the social crowd: an analysis of quora. In *WWW '13: Proceedings of the 22nd international conference on World Wide Web*, 2013.

[84] Lucy Wang, Arthi Ramachandran, and Augustin Chaintreau. Measuring click and share dynamics on social media: A reproducible and validated approach, 2016.

[85] Ting Wang, Dashun Wang, and Fei Wang. Quantifying herding effects in crowd wisdom. In *the 20th ACM SIGKDD international conference*, pages 1087–1096, New York, New York, USA, 2014. ACM Press.

[86] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. Automatic crime prediction using events extracted from twitter posts. In *SBP'12: Proceedings of the 5th international conference on Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer-Verlag, April 2012.

[87] FMF Wong, S Sen, and M Chiang. Why watching movie tweets won't tell the whole story? In *Proceedings of the 2012 ACM workshop . . .* , 2012.

[88] Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. Who says what to whom on twitter. In *WWW '11: Proceedings of the 20th international conference on World wide web*. ACM Request Permissions, March 2011.

[89] Jaewon Yang and Jure Leskovec. 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, February 2011.

[90] Reza Bosagh Zadeh, Ashish Goel, Kamesh Munagala, and Aneesh Sharma. On the precision of social and information networks. In *COSN '15: Proceedings of the third ACM conference on Online social networks*. ACM Request Permissions, October 2013.

[91] Tauhid Zaman, Emily B Fox, and Eric T Bradlow. A Bayesian Approach for Predicting the Popularity of Tweets. *Annals of Applied Statistics*, 8(3):1583–1611, September 2014.

[92] Georgios Zervas, Davide Proserpio, and John Byers. The Impact of the Sharing Economy on the Hotel Industry: Evidence from Airbnb's Entry in Texas. *EC '15: Proceedings of the sixteenth ACM conference on Economics and computation*, pages 1–36, January 2014.

[93] Yuchen Zhang, Weizhu Chen, Dong Wang, and Qiang Yang. 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, pages 1388–1396, New York, NY, USA, 2011. ACM.

[94] Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. 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, pages 1513–1522, New York, NY, USA, 2015. ACM.