# Filter & Follow: How Social Media Foster Content Curation

Avner May
Columbia University
New York, NY
avnermay@cs.columbia.edu

Augustin Chaintreau
Columbia University
New York, NY
augustin@cs.columbia.edu

Nitish Korula
Google
New York, NY
nitish@google.com

Silvio Lattanzi
Google
New York, NY
silviol@google.com

> "The Internet has turned the news industry upside down, making it more participatory, social, diverse and partisan – as it used to be before the arrival of the mass media."
> T. Standage, *The Economist*, July 7th, 2011.

## ABSTRACT

The impact of blogs and microblogging on the consumption of news is dramatic, as every day users rely more on these sources to decide what content to pay attention to. In this work, we empirically and theoretically analyze the dynamics of bloggers serving as intermediaries between the mass media and the general public.

Our first contribution is to precisely describe the receiving and posting behaviors of today's social media users. For the first time, we study *jointly* the volume and popularity of URLs received and shared by users. We show that social media platforms exhibit a natural "content curation" process. Users and bloggers in particular obey two *filtering laws*: (1) a user who receives less content typically receives more popular content, and (2) a blogger who is less active typically posts disproportionately popular items. Our observations are remarkably consistent across 11 social media data sets. We find evidence of a variety of posting strategies, which motivates our second contribution: a theoretical understanding of the consequences of strategic posting on the stability of social media, and its ability to satisfy the interests of a diverse audience. We introduce a "blog-positioning game" and show that it can lead to "efficient" equilibria, in which users generally receive the content they are interested in. Interestingly, this model predicts that if users are overly "picky" when choosing who to follow, no pure strategy equilibria exists for the bloggers, and thus the game never converges. However, a bit of leniency by the readers in choosing which bloggers to follow is enough to guarantee convergence.

## Keywords

social media; blogs; information diffusion; potential game

*SIGMETRICS'14,* June 16–20, 2014, Austin, Texas, USA.
ACM 978-1-4503-2789-3/14/06.
http://dx.doi.org/10.1145/2591971.2592010.

## 1. INTRODUCTION

One of the most transformative forces in media consumption is the recent multiplication of referrers or intermediaries. Most users still ultimately rely on content produced by professional journalists for accuracy and impartiality. But users do *not* depend any longer on traditional media to determine *what* they should see. Increasingly, people use social media to refer them to content of interest. Thus, more and more content is reaching people by traversing a network which was built from previous choices; these choices include what content users decide to post, and whose posts an individual decides to pay attention to.

The main challenge of online content curation is that the interests of users in general do not follow a simple predictable model. Users have a wide range of interests across a large set of topics; even within a topic (*e.g.,* New York sports, U.S. politics, a major geopolitical or societal event, etc.) different users will likely be interested in different subsets of the news about this topic. Social media at its core makes a remarkable promise – that through chains of referrals the torrent of information produced online can be tamed and distilled to deliver to each user the content they are interested in. In other words, the medium formed by users' connections and posting behaviors is efficient from a *communication* standpoint.

Our objective in this paper is to assess the credibility of this promise. Various claims have been made on the power of social networks to diffuse information quickly and efficiently. These typically follow from topological properties and they quote volumes of large viral cascades for content of popular appeal. Most highlight the importance of information intermediaries, such as professional journalists, bloggers, or prominent people who are well-positioned to connect different parts of the network. We wish to go beyond these structural properties and see how social media shape what content is shared: How do today's information intermediaries choose what to post? How are these choices affected by the content's popularity? Can a follower graph emerge from users' decisions that puts the right information in front of them? To address these questions, one needs to combine insight from measurement at a microscopic scale with a macroscopic analysis that predicts the efficiency of the network as a whole.

Our work empirically studies today's online crowd-curation with a new level of detail. We present an intriguing role played by information intermediaries, and analyze theoretically its consequence on the efficiency of social media. More precisely, our work makes the following contributions:

- Based on 11 social media data sets, differing in their source platforms (*e.g.,* Twitter, Facebook, Spinn3r), topics, and time periods, we demonstrate that social curation obeys a “filtering law”. This law predicts that the popularity of content received by a user, or posted by an intermediary, is inversely correlated with the amount of content she receives or posts. This relationship is robust to different ways of defining “information intermediaries”, and quantifying content popularity. Furthermore, we show that this trend is not simply a statistical consequence of the very skewed content popularity distribution. (Section 3).

- Informed by the above observations, we hypothesize that bloggers are behaving strategically, trying to post a unique, high-quality set of content, in order to attract a large following. Thus, we model the social curation system as a competition between bloggers for user attention, and call this model the “blog-positioning game.” But how does a user pick which blogger to follow? We consider two types of factors: 1) how well the content posted by an intermediary matches her interests (“endogenous factors”), and 2) “exogenous factors” such as the reputation of the intermediary and their writing style. Our model considers both types of factors, and is able to reproduce the observed filtering law. (Section 4).

- We analyze whether pure strategy Nash equilibria exist under the above model, and if so, whether these equilibria are “efficient” (in terms of the number of users satisfied with the content they receive, relative to the optimal intermediary configuration). We show that the existence of equilibria depends deeply on the way in which users pick which blogger to follow. If users precisely pick the blogger that gives them maximal utility, no equilibria exist; on the other hand, if they simply pick a blogger that is “good enough”, an equilibrium always exists. Moreover, in the resulting social network, at least half of the optimal number of users are satisfied with the content they receive. (Section 5)

Together, our empirical and theoretical findings lay the foundation for better understanding the surprising power of social networks as a communication medium.

## 2. RELATED WORK

Our work follows the classic hypothesis of a *two-step information flow* [11], where news savvy “opinion leaders” play the key role of intermediaries between the content produced by mass media and the general public. This thesis was revived using empirical evidence from Twitter [13, 20], and more recently identifying mass media and intermediaries as critical to information spread [6]. In fact, our work extends this hypothesis to model how even relatively normal users play the role of information filters for their peers. Much remains to be explored to analyze the intrinsic effectiveness of such information spread. A recent study showed that social networks allow users to be exposed to more diverse viewpoints [2]. On the other hand, study of word-of-mouth propagation showed that this spread is more likely between geographically close friends, hinting at a possible bias effect [16]. Our work is complementary as we explore a new dimension: how intermediaries help manage the *volume* of information by serving as filters and adapting to the interests of the audience. We depart from the purely empirical approach by analyzing the dynamics between bloggers as they compete for user attention.

Since a first version of our work appeared, a recent analysis of Twitter suggested that social media can be efficient at giving users information relevant to their interests [22]. It showed this efficiency is compatible with follower relationships and interests that both derive from the same underlying structure in a Kronecker graph, although this argument currently implies that the graph is undirected. Motivated by the same observation, we offer a very different starting point, by focusing on the behaviors of intermediaries in directed graphs. In our work the structure is not assumed *a priori* but naturally emerges from the interaction of various agents of the curation ecosystem.

As news moves online, how to best serve the various information needs of communities remains an important issue of public concern [18]. Previous works analyzed how to best serve this collective attention, either by analyzing temporal dynamics and treating this as a scheduling problem [19, 4, 14, 21], by adapting personalization tools [7], or by enhancing news navigation. Competition among different news offerings is generally ignored. A recent and notable exception is the debate on the effect of news aggregators. The increase of traffic they generate to news outlet was shown to come at the expense of customer loyalty, leading to a potential loss of advertising revenue for content publishers [3]. Our work is radically different as we consider competition *between* various intermediaries as well. We wish to see if the collective behavior of intermediaries can be harnessed and their incentives aligned to serve every user’s interests. Our game theoretic approach resembles recent studies on the efficiency of collective efforts [12], with two important differences: First, the context of content curation is entirely new and it involves two sides (the bloggers choosing *what* to post, and the users choosing *who* to follow). This context allows one to reuse previous bounds on efficiency [17]. Since the first version of our work was presented, a 4-page workshop paper independently derived the same observation and studied best response dynamics in curation systems [1]. Our model differs in its assumptions about how readers select their intermediaries, which has large consequences on the equilibrium properties of the game. Furthermore, in our setting, the existence of a pure strategy equilibrium requires an entirely different proof to show the game is a “potential game.”

We are only aware of a few other works analyzing news dissemination with a game theoretic background. Previous work assumed news is equally appealing to all users [9], and showed that local incentives suffice for higher quality news to reach a large audience in a random graph, by a connectivity argument. It also showed that users attempting to avoid spam may be an important limiting factor. In contrast, we are interested in how various blogs emerge to satisfy the information needs of a heterogeneous audience. In [10], a multi-topic model is analyzed to model news aggregators choosing a subset of topics in a sequential game. In contrast, our model allows for a much richer set of posting strategies, in which intermediaries can pick precisely what set of articles to post, across any topics. We assume bloggers compete for user attention through their choice of what to post, and we wish to find the conditions under which such games admit natural equilibria that are also efficient.

# 3. EVIDENCE OF FILTERING

Our first main result is to empirically validate our hypothesis that users of social media *participate* in, and potentially *benefit* from, an information filtering process. Prior work proved the existence of information intermediaries, while we explore for the first time how these intermediaries affect what information users receive. We wish to answer the following questions: Is the subset of content that people receive today through social media selected in a particular way? If so, can this filtering be understood as a consequence of the way intermediaries post information and how users select who they follow? And do the answers to these questions vary across social media platforms?

## 3.1 Data-sets

To answer the questions above, we analyze several large traces from social media. This requires (1) records of what users post in various media, (2) a way in which this information can be judged for its quality, and (3) knowledge of the social graph (to know what posts users receive). To the best of our knowledge, no data set available today allows studying all of these factors together, especially without significant subsampling. We hence gathered data from various sources (including Twitter, SPinn3r blogs, and Facebook), and enriched these data sets with complementary crawls, which we now describe.

### *Collecting information shared on social media.*

At a high level, we consider three types of social media: (1) microblogging (Twitter, denoted "TW"), (2) online social networks (Facebook, denoted "FB"), and (3) blogs (Spinn3r, denoted "Blogs"). In all these domains, we focus on posts containing URLs. We gathered a mix of *topical* traces (i.e. relating to an event) and *generic* ones (where all posts are included). We ensure that the data sets are comprehensive, containing users with a wide range of activity levels and follower amounts, and URLs of both wide and narrow appeal.

| Data sets | Users | URLs |
|---|---|---|
| | (% live today) | (% with bit.ly) |
| TW NYT | 226,512 (92%) | 7,504 (99.96%) |
| TW Bin Laden | 700,783 (75.9%) | 545,495 (19.7%) |
| TW Occupy WS | 354,117 (88.9%) | 316,408 (26.9%) |
| TW Steve Jobs | 719,025 (86.8%) | 250,644 (20.0%) |
| TW iPhone5 | 81,056 (94.6%) | 37,323 (30.7%) |
| FB iPhone5 | 330,185 (N/A) | 193,024 (14.6%) |
| Blogs All | 67,692 (N/A) | 440,933 (30.9%) |
| Blogs Obama | 13,390 (N/A) | 84,733 (40.9%) |
| Blogs Facebook | 11,643 (N/A) | 69,747 (40.5%) |
| Blogs Euro | 9,659 (N/A) | 53,001 (39.9%) |
| Blogs Mubarak | 6,546 (N/A) | 42,531 (34.6%) |

**Table 1: Number of users and URLs for our data sets. For the Twitter (TW) data sets, we report the fraction of users whose accounts are still active. We also report the fraction of URLs with a bit.ly shortener.**

- **Microblogging** (TW): Data was gathered using `DiscoverText.com` to harvest tweets from Twitter's unrestricted firehose in two ways. First, during two weeks in Dec. 2012 we collected all tweets with a URL from a particular media source; we report here results from the New York Times (NYT), though we also gathered data from CNN and FoxNews. Second, we collected all URLs associated with a particular event (*e.g.,* the "TW Bin Laden" data set contains tweets with the words "Osama" or "Bin Laden" posted during the 34 hours following his death, from 5/2/2011 at 3:30am EST to 5/3/2011 at 1:30pm EST). We similarly collected tweets on Steve Jobs' death, the Occupy Wall Street movement, and the iPhone 5 launch.

- **Online Social Networks** (FB): We gathered ≈ 1 million Facebook comments, all about the iPhone 5 release, from the GNIP Facebook API.

- **Blogs**: Finally, and in order to validate our results on data sets that are already publicly available, we use the data sets of the ICWSM 2011 data challenge [5]. It contains 386m blog posts collected from Spinn3r during the months of January-February 2011.

### *Preprocessing and additional crawls.*

**URL normalization**: To avoid duplicates due to formatting variants and URL shorteners, we normalize URLs by (1) recursively following all redirects using HTTP calls, (2) removing protocols (`http/https`) and suffixes (anything following a `#` sign, ".html", or ".htm"), and (3) removing spam.[1]

**Twitter Crawls**: We performed Twitter crawls to gather additional information about the users in our data sets.[2] First, we gathered general information about these users via the Twitter REST API, as well as a sample of up to 100 "lists" they are part of, to use for classification (see below). Additionally, to find for each user the set of URLs they received via Twitter, we crawled the active users in our NYTimes data set for the set of users they follow (called "friends" in Twitter). To keep the crawl computationally feasible, we stopped the crawl after 50k friends; following more than 50k users is rare (<0.1%) and usually implies that these edges are not relevant for information dissemination. In order to get a sense of the URLs received by users who were not active in our NYTimes dataset ("passive" users), but nonetheless received a link, we crawled users in our data set for their followers (once again capping at 50k); we then crawled a random subset of all these newly discovered passive users (who follow at least one person in our dataset) for their "friend" lists. Thus, for the users in our TW NYTimes data set, as well as for a random subset of their followers, we know who they follow, and thus what NYTimes links they received during this time period.

**URL popularity and user posting activity** were estimated using 5 different metrics to ensure the universality of our results. First, one can directly *internally* measure

[1] For these data-sets, spam URLs have generally been blacklisted (*e.g.,* spam bit.ly links land on a "warning" page). We removed users posting one or more spam URLs.

[2] Information could be gathered only from accounts that are still live today (between 75% and 95% of accounts, depending on how long before crawl data set was collected)

these in each data set by counting how many times a normalized URL was posted, and how many URLs a user posts. We complement this by measuring *externally* how popular a given URL is in two ways: (1) if this URL was ever shortened using bit.ly, we collected its number of clicks via the bit.ly public REST API.[3] We also collected the number of times each URL was shared on Facebook using the FB API. Finally, to estimate the "external activity" of a user, we calculate the average number of tweets per day by this user, from the time their account was created.

As seen in Table 1, we are successful in gathering a large number and more importantly a large range of URLs and users. The majority of URLs appear only once in our dataset, pointing to a very long tail of obscure or niche content. This set of URLs also contains a tiny minority of blockbusters which spread massively. As a consequence, the average number of posts or clicks is typically much larger than the median, and sometimes even goes above the 95% percentile. This trend is especially skewed for external popularity measures. Similarly, the overwhelming majority of users post only occasionally, while a very active minority of users end up posting most of the content.

**Defining information intermediaries** in social media is another challenge. We show it is tangential to our work as our trends apply to various definitions of intermediaries. In a "loose" sense, all active users are intermediaries, since anyone who has ever posted a URL has acted as an intermediary to their immediate surrounding. One can argue that this contains too many users reacting spontaneously to news without much of a following. Thus, we consider the stricter definition of only counting users with at least 500 followers (top 30% of active users) as intermediaries; we refer to these users as *Active VIPs* [4]. We finally use a much stricter definition where, among those VIPs, only the ones that appear in a list containing the substring "blog" are considered information intermediaries (among the $\leq$ 100 lists we crawled per user); we refer to these users as *Bloggers*. Figure 1 provides detailed statistics about these different populations for the Twitter NYTimes data set (TW NYT). In all definitions, potential information intermediaries are a very small part of the total audience (<1%) but they are responsible for most of the traffic. They post significantly more tweets per day. They also receive more URLs and follow more users.

## 3.2 The Filtering Law

We first study the relationship between the volume that one receives on social media, and the popularity of this content. Ideally, depending on how many and which intermediaries a user follows, she may be able to reduce the volume of information she receives while statistically increasing its quality. We now show that this is indeed the case.

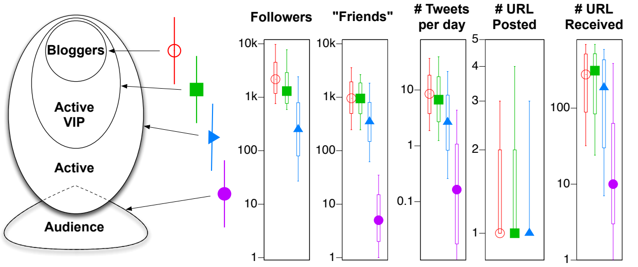


**Figure 1: Various types of users (TW NYT): Each group is associated with a symbol. We draw the median value of various parameters, together with the box plot representing the (25%,75%) percentile range, and whiskers for (10%,90%) percentiles.**

Our main observation is the following *filtering law*: URLs received by users who receive fewer URLs are disproportionately popular. This is shown in Fig. 2, where we see that the average quality of URLs received by a user decreases as the total number of URLs she receives increases.

What explains this trend? Is it a statistical illusion? Perhaps this is simply due to a replacement effect: as users receive more URLs, they may "run out" of high-quality URLs, causing a decrease in average quality. To show that this is not the case, we contrast the real trend with a random Null Hypothesis (NH) model that is constructed as follows: for each user in our data set receiving $n$ URLs, we reassign the URLs they receive by picking these $n$ URLs randomly from the same URL popularity distribution (*i.e.,* popular URLs appear more often, exactly as in the real data). This popularity distribution is determined by the number of times each URL was received. Importantly, we perform these random draws without replacement, thus not allowing a user to receive the same URL twice. The blue line in the figure denotes the Null Hypothesis, and we observe a very slight (almost flat) decreasing trend. This slight decrease comes from the following effect: Since the distribution is highly skewed, a small set of popular URLs are likely to be chosen first as the set of received URLs is constructed in this random model. For a larger set, since the set of URLs is constructed without replacement, after the first popular URLs are chosen, the random choices will be biased to less popular URLs. Note that regardless of the precise metric (*e.g.,* number of times URL received, number of times URL posted, number of Bit.ly clicks received by URL, etc.) we use to construct the URL popularity distribution for the Null Hypothesis, the resulting trend is flat. On these plots, we show the median popularity of URLs received by users in both the Null Hypothesis model, and in the real data, and we also show the 25th and 75th percentiles of these URL popularity distributions. Note that because we are plotting percentiles, and not confidence intervals, the number of users in each bucket doesn't necessarily shrink the size of the 25%-75% interval.

We observe that real data shows a trend much stronger than the replacement effect of the Null Hypothesis. For any activity level, the difference between the distributions are statistically significant: a standard Student test on the log of popularity always returns statistical difference with $p \leq 10^{-6}$. Moreover, this trend is universal: we observe it for all popularity metrics, as shown in Fig. 2, and when considering URLs received by various subsets of the population (like

[3] For precision, we consider several variants of the URL which could link to the same page, by (1) computing the "core" of the URL (removing the protocol, trailing slashes, and everything after ".htm",".html", "#"), and (2) summing the bit.ly clicks provided by the API for the following URLs: all the ones posted with the same "core" extended by varying the protocol http/https, with and without a trailing slash.

[4] We tried various thresholds here, in addition to 500 followers; however, the choice of threshold did not affect the general trends we observed. For simplicity, we only present the results for the *Active VIP* users defined according to the 500 follower threshold.

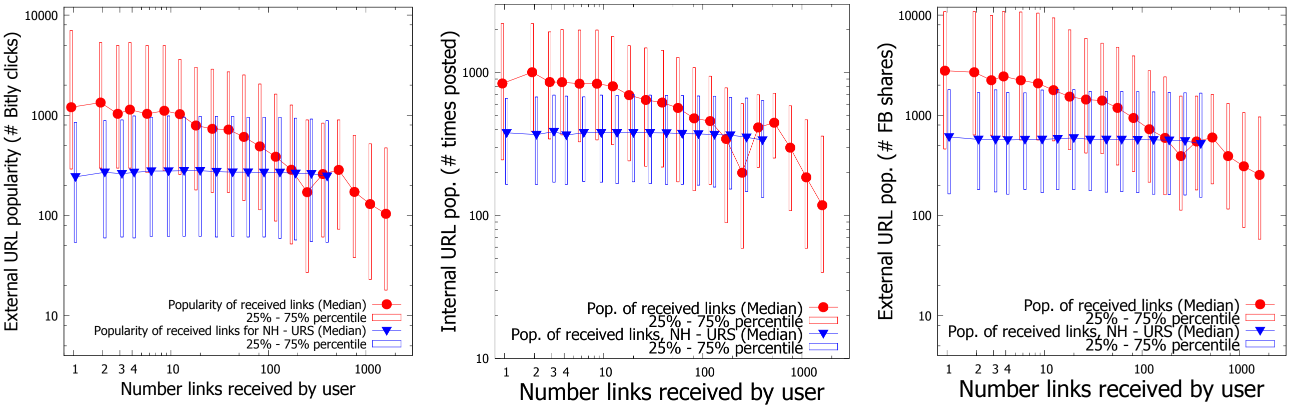


**Figure 2: The filtering law. (x-axis) Users grouped by number of URLs received, (y-axis) Popularity distribution of URLs received by these users, shown using median and 25%-75% percentiles. Red circles denote real data, blue triangles denote a random null hypothesis model. The 3 plots correspond to different metrics of URL popularity: (left) number of bit.ly clicks, (middle) number of times this URL was posted, (right) number of FB likes.**

Active or Active VIP users, and Bloggers). Our results are encouraging because they suggest that users of social media are effectively navigating the volume/quality trade-off: users who receive less are focusing on the most promising URLs.

These positive results demand further study. The only factors affecting the URLs a user receives are the intermediaries she follows, and the content they post. Are information intermediaries behaving in a way that explains this law? If so, are they naturally encouraged by social media to behave as such? Does this generally lead to an efficient propagation of information? In the rest of this section, we will examine the posting behaviors of information intermediaries, and their consequences.

## 3.3 Intermediaries' posting behavior

We begin our study of the content posted by intermediaries by observing another filtering law: URLs posted by less active intermediaries are disproportionately popular. That is, as intermediaries decrease the number of URLs they post, they are more likely to post the most popular URLs. Fig. 3 illustrates this trend; as before, the trend is much stronger than predicted by a random null hypothesis, where for each user we fix the number of URLs they post, and pick these URLs randomly according to the same overall popularity distribution. With similar tests, we observe statistical significance with $p < 10^{-3}$ (except for users posting exactly 5 URLs where $p = 0.02$). This observation suggests that intermediaries are *filtering* when choosing what content to post (more on that below).

Our trend is remarkably universal. No matter what URL popularity metric is used, or what intermediary definition is used, the trend is statistically significant. We present a few representative results in Fig. 4.

The posting filtering law shows that the network at a macroscopic level filters information to bias small volume towards popular content. However this does not detail how intermediaries select URLs at the microscopic level. It could be that they simply selectively repost a fixed fraction of the content they receive. Intermediaries who receive less, post less; those who follow them receive still less, etc. Figure 5 proves the contrary. We first observe (see top plot) that across activity levels, the popularity of URLs received by users is roughly the same. By comparison, the URLs which users choose to post are significantly more popular than those received; the difference in popularity is even more striking when only considering the subset of received URLs that users choose to repost. These gaps narrow for users posting more than 20 URLs, but this represents less than 1% of intermediaries. Moreover, Fig. 5 (bottom) shows that the average number of URLs received by intermediaries is more or less constant across activity levels. However, users

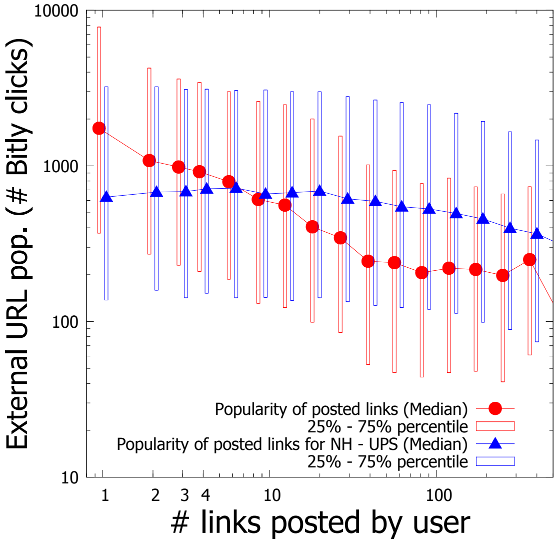


**Figure 3: The posting filtering law. (x-axis) Users grouped by number of URLs posted, (y-axis) popularity distribution (by number of bit.ly clicks) of URLs posted by these users; median and 25%-75% percentiles shown. Red circles denote real data, blue triangles denotes a random null hypothesis model.**

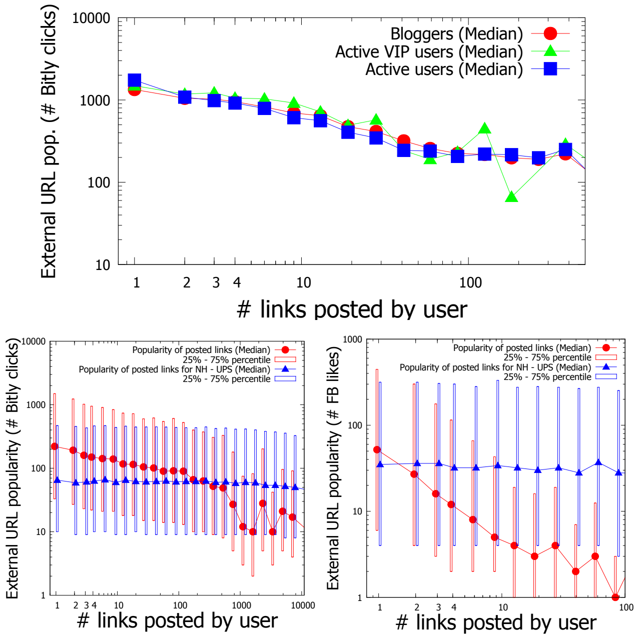


**Figure 4: The posting filtering law is universal (same axes as Figure 3): (top) posting filtering law is unchanged when the definition of intermediaries changes, (bottom left) Blogs All data set, (bottom right) FB Iphone5 data set (y-axis denotes FB likes).**

with different activity levels differ very much in their reposting frequencies. In other words, intermediaries more or less receive similar URLs (both in number and popularity), but they choose to repost different subsets of it, usually composed of URLs that are more popular. Note that in addition to "reposting" URLs they receive, users can also post URLs which they discover from outside Twitter; however, even these are quite popular on average (as can be seen by the gap between the green-triangle line and the blue-square line in Fig. 5).

Finally, we studied the factors affecting the relative "success" of an information intermediary, defined by her number of followers. We noticed that in general, intermediaries who post more URLs have more followers, though this trend shows diminishing returns. Figure 6 plots the raw measure of success (followers in Twitter, number of likes received by their comments in Facebook) for users of various activity levels. We observe that success correlates with user's activity (roughly in proportion of the square root of activity, also plotted for reference. Note that we do not perform precise trend fitting here). This trend is observed across our data sets. Note that we do not claim causality: it is also possible that popular people tend to become more active. One can also interpret these diminishing returns naturally in light of the filtering law. Posting more often correlates with more occasions to get noticed and gather followers and likes. At the same time, it corresponds to posting less popular URLs, which perhaps contribute to overall success to a smaller degree. For future work, we intend to study what other attributes of intermediaries (*e.g.,* average popularity of URLs they post, earliness) correlate with success.

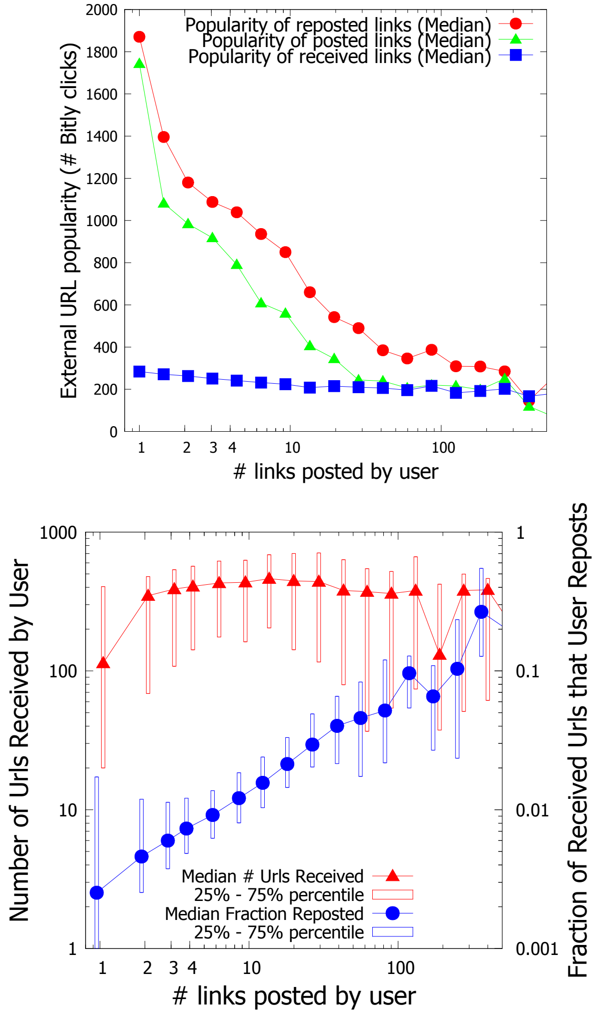
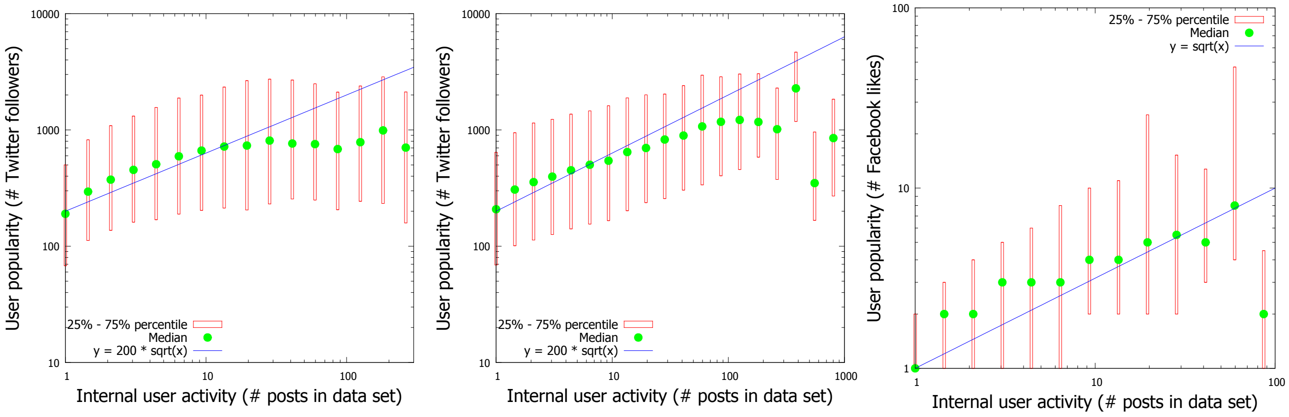


**Figure 5: Selection of popular content by intermediaries: (x-axis) User posting activity in number of URLs, (top y-axis) popularity distribution of URLs received by these users, and the popularity distributions of URLs posted or reposted by the users. (bottom left y-axis) distribution of number of URLs received plotted as red triangle, (bottom right y-axis) distribution of reposting fraction plotted as blue circle. (Boxplots show 25%-75% percentiles).**

In summary, we have observed two important phenomena about the posting behavior of intermediaries: First, they are selective in what they post, such that the average popularity of URLs posted by an intermediary is inversely proportional to the amount they post. Secondly, this difference in posting behavior is explained not by differences in the sets of URLs *received* by these intermediaries, but rather by their distinct decisions of what to post. In effect, we see that intermediaries differ widely in their posting behaviors. In the next section, we attempt to better understand why such a wide range of posting behaviors might arise.



(a) TW Bin Laden

(b) TW Occupy WS

(c) FB iPhone5

**Figure 6: The diminishing return of posting activity: (x-axis) user internal activity in number of URLs, (y-axis) user "success" (followers, likes).**

# 4. INFORMATION FILTERING MODEL

The previous section showed that information intermediaries pick different subsets of content to post, the size of which also relates to content popularity. Understanding this behavior in the context of social media efficiency as a whole raises two questions: Can the intermediaries' behavior be understood as the optimization of some natural objective? If so, would this derived model imply that social curation is efficient?

### *Motivation for Model.*

It seems natural to assume that intermediaries wish to increase the attention they receive. The factors that determine which intermediaries a user chooses to pay attention to could be classified into two groups: First, *endogenous* factors related to the content posted by the intermediary, and *exogenous* factors which capture all other variables, such as the identity of the blogger, her writing quality and style, etc. Ideally, a model would allow for both types of factors, reproduce our observations, and remain tractable.

We first note that a model based purely on exogenous factors cannot satisfactorily answer the two questions above. If users choose who to follow independently of the content posted, this could allow for outcomes in which users never receive content they are interested in. Furthermore, in this type of model, bloggers can't optimize their posting behavior to gather a larger audience. This type of model is both unrealistic, and makes it impossible to study the questions we are interested in.

In contrast, a model with endogenous factors immediately provides a natural explanation for the filtering law. Consider the following simple example: All content items belongs to a single topic, and vary only based on their quality. Every user has an "interest threshold" indicating that they are interested in seeing all content above that quality level. In response these user interests, each blogger picks a "posting threshold" and posts all items above that quality level. Each user follows the blogger whose posting threshold is closest to their interest threshold, breaking ties randomly. In this scenario, how would a strategic bloggers behave? Assuming that there was a wide range of user interest thresholds, it would be in the best interest of bloggers to "spread out" their posting thresholds – for some to only post the highest quality content (to satisfy users only interested in the highlights), while others post a larger set (to attract more interested users). If all bloggers posted the same content, any of the bloggers could change their posting strategy and attract a larger set of users who prefer them to the rest of the bloggers.

This simple unidimensional scenario is sufficient to reproduce our main findings: (i) Users who receive less content tend to receive the most popular content. (ii) Intermediaries have very different posting thresholds / frequencies. (iii) Less active intermediaries / bloggers generally post more popular content.

### *Model Definition: "Blog-Positioning Game".*

We now formally describe our model, which we call the "blog-positioning game". This model accounts for arbitrary content spaces, both endogenous and exogenous attributes of bloggers, and different ways in which users may decide which blogger to follow. We define and motivate each component of the game below:

**Sets of Agents**: We use $B$ to denote the set of all bloggers (these are the strategic agents in our game), and $U$ to denote the set of users in the "audience".

**Content Space**: We assume that every item of content lives in an arbitrary space $\Omega$. Every day $k$, a set of news items $X_k \subset \Omega$ are generated from $\Omega$ following an arbitrary, possibly probabilistic, but time-invariant process.

**Blogger strategies**: We assume that each blogger $b$ picks a filtering function $F_b\colon \Omega \to \{0,1\}$, that indicates for each possible news item, whether or not the blogger will post it. We will denote the strategy space of all possible filtering functions $F_b\colon \Omega \to \{0,1\}$ by $\sigma$. So, the strategy vectors $s$ in this game live in the space $\sigma^{|B|} \equiv \Sigma$, where each element of $s$ corresponds to a specific blogger's choice of filtering function. Note that our model can extend to the case where the blogger filtering functions are stochastic. In this setting, $F_b \colon \Omega \to [0,1]$ indicates, for each possible point in the space, the probability that the blogger will post an item at this point. All our results carry over to the case of stochastic filtering functions by replacing the utility of a user with their expected utility, but for ease of notation, we restrict our analysis to the deterministic case.

The set of items a blogger posts on a given day is not always the same since the content generated on a given day is created randomly. However, it seems natural to assume that when the same exact content is produced, then the blogger will select the same exact subset to post. Thus, we are interested in *pure strategies* for bloggers, modeling blogs that apply an invariant filter to content items that are generated randomly every day. This contrasts with a *mixed strategy*, where the blogger initially fixes a set of strategies, and then randomly chooses one of them each day based on a coin toss. Mixed strategies are more general, but seem unnatural: why would a blogger decide to randomly change her behavior from day to day, in a potentially drastic fashion, when her target audience has deterministic tastes? [5]

**Blogger objectives**: We assume that bloggers are attempting to maximize their number of followers (or expected number of followers). More specifically, they are attempting to maximize the amount of attention they receive from their followers. Each follower is worth 1 "attention unit"; in the case where a user chooses to follow several bloggers, this attention unit is split across these bloggers.

**User Utility**: How much utility does a user $u$ get from following blogger $b$ with filtering function $F_b$? We assume a user u gets utility $V_u(F_b, b) \in \mathbb{R}$ from following blogger $b$ with filtering function $F_b$. Note that we include $b$ as a separate parameter of $V_u$ so that a user's utility from following a blogger can depend on the identity of the blogger (this can capture all exogenous factors), and not simply on the content they post (the endogenous factors). Also, note that this utility could be negative, in which case the user would not choose to follow this blogger. In general, we always allow the option for a user to not follow any blogger and get utility 0 (we assume for simplicity's sake that if a user has the choice between following a blogger who gives them utility 0, and not following any blogger, they will choose to not follow any blogger).

For the sake of illustration, we can specify the following functional form for $V_u$, even though the specifics of how $V_u$ is defined do not affect our theoretical results. We can assume each user $u$ has an interest function $h_u \colon \Omega \to \mathbb{R}$ that assigns a real number to every item, according to its position in $\Omega$. This value represents how much the user would enjoy seeing this item. If a user receives a set $X$ of content, we assume they receive utility $\sum_{x \in X} h_u(x)$ (or it could be a function of this sum, with diminishing returns). So, if on day $k$ a user $u$ chooses to follow a blogger $b$, they will receive utility $\sum_{x \in F_b(X_k)} h_u(x)$, where $F_b(X_k) = \{x \in X_k : F_b(x) = 1\}$ is the set of items posted by blogger $b$ on day $k$. The value of this sum may vary by day as content gets randomly produced. Thus, we can define the utility $V_u(F_b, b)$ that this user gets from following $b$ as this sum's expected value. More explicitly, $V_u(F_b, b) = \mathrm{E}\left[\sum_{x \in F_b(X_k)} h_u(x)\right]$, where this expectation is taken over the probability distribution from which the daily set of news $X_k$ is drawn from $\Omega$ (as well as over the stochastic filtering function, when appropriate).

**How do users pick what blogger to follow?**: We specify 3 separate ways in which a bloggers could pick who to follow. It turns out that the precise way in which a user picks her blogger has an enormous effect on the equilibria of the game. The three models are:

1. **Greedy Model**: In this model, each user $u$ chooses to follow blogger $b = \arg\max_B V_u(F_b, b)$. Note that, in order for a user to behave this way, she must have perfect information about every blogger's posting behavior, a strong assumption to make.

2. **Satisficing Model**: In this model, we assume a user "satisfices", meaning that they randomly pick a blogger to follow amongst the ones that make them "happy enough". More formally, we assume each user $u$ has a utility threshold $v_u$, and that they pick which blogger to follow randomly from the set of bloggers $b$ such that $V_u(F_b, b) \geq v_u$ [6]. An alternative and equivalent interpretation is that a user splits their attention evenly across these bloggers. If we let $S_u(s) = \{b \mid V_u(F_b, b) \geq v_u\}$ denote the set of bloggers who provide user $u$ with at least the required utility under strategy vector $s \in \Sigma$, then every blogger $b \in S_u$ receives a fraction $\frac{1}{|S_u|}$ of $u$'s attention (in expectation). This model can be seen as relaxing the assumption that users have perfect information about the blogger behaviors, and could perhaps be seen as more realistic as a result.

3. **Satisficing Model with Blogger Abilities**: In this model, we assume different bloggers have different instrinsic "abilities", which allow us to represent various external factors that can contribute differently to each blog's success (*e.g.,* writing quality, fame, the time since the blog was created). We model bloggers' abilities as follows: We suppose that each blogger has an associated ability $a_b \in [0,1]$, and *if* he provides a user with at least her required utility $v_u$, the user decides whether to follow him or not based on his ability. In particular, a user flips a biased coin that gives heads with probability $a_b$; if the coin lands heads, the user follows the blogger, otherwise she does not. As in the previous section, we can assume a user shares her attention equally among all the bloggers for whom the coin comes up heads (or that the user picks uniformly at random from these bloggers). Note that this model is a strictly more general than the previous model, because if all bloggers have ability $a_b = 1$, then these two satisficing models become equivalent. The primary reason we consider this more general model is that it allows us to better explain differences in follower amounts between bloggers that post very similar content.

[5] Note that mixed strategies are fundamentally different from stochastic filtering functions: The latter model minor day-to-day "random" variations in an essentially consistent filtering process, but the former imply that bloggers explicitly make a (perhaps widely) different random choice each day based on the outcomes of a coin toss.

[6] Of course, users need not calculate utility explicitly; they may have a general notion of whether a blog is "good enough" to keep them well-informed.

**Formal game definition**: Our "blog-positioning game" is a simultaneous game, in which all bloggers need to pick their posting strategy $F_b \in \sigma$, in order to maximize their (expected) number of followers. We assume that all bloggers have perfect knowledge of all users' utility functions $V_u$, and of which strategy the users are employing to pick their blogger. Given a strategy vector $s \in \Sigma$ containing every blogger's choice of filtering function $F_b$, each blogger can precisely calculate $V_b(s)$, the number of followers they expect to receive under this strategy vector.

In the section that follows, we analyze the pure strategy Nash equilibria of this game, and in addition study how "efficient" these equilibria are by considering the "price of anarchy" of the game. We perform this analysis for all three variants of the game, corresponding to the 3 different ways users can pick who to follow.

# 5. OUTCOMES OF BLOG COMPETITION

We now formally analyze the outcomes of the model presented in Section 4. As mentioned, the precise way in which users pick their blogger has unexpectedly important consequences on the nature of the game's equilibria. We are interested in 2 properties of these games: (i) Whether the game has pure strategy Nash equilibria, and (ii) Whether the equilibria of the game are "efficient" relative to the social optimum (we define social welfare to be the number of users who follow at least one blogger. More details below). At a high-level, we will show 3 primary results:

1. Under the greedy user model, there are cases in which no pure strategy Nash equilibrium exists.
2. Under the satisficing user models (with and without blogger abilities), the game is an "exact potential game" and always admits a pure strategy Nash equilibrium.
3. The price of anarchy (ratio of social optimum to the social welfare under the worst possible Nash equilibria) of the game under either satisficing user model is at most 2, and this bound is tight.

## 5.1 Greedy user model

The first question that we explore is: does the system admit a pure strategy equilibrium in the greedy user model? The answer is negative. In fact, we can prove that there even exist simple cases where only two bloggers are present, in which no such equilibrium exists.

Proposition 1. *Under the greedy user model, there exist cases where no pure strategy Nash equilibrium exists.*

Proof. We will prove this by showing a specific, relatively simple example of the blog-positioning game, in which no pure strategy Nash equilibrium exists. The example is as follows: Suppose that the generative process of news creation creates one item daily in each of 3 points of $\Omega$: $p_1, p_2$ and $p_3$, where $p_1$ is the most important item, and $p_3$ the least. Let us assume that we have a population of 66 users, distributed in three groups who value items differently:

| Number of users | Value for | | |
|---|---|---|---|
| | $p_1$ | $p_2$ | $p_3$ |
| 30 | $+1$ | $+1$ | $+1$ |
| 20 | $+1$ | $+1$ | $-1$ |
| 16 | $+1$ | $-1$ | $-1$ |

Then with the *greedy user model*, there is no pure strategy equilibrium in this scenario even with two bloggers.

First note that we can assume without loss of generality that each blogger picks from one of the following 3 strategies: (i) $s_1$, posting only $p_1$ (ii) $s_2$, posting $p_1$ and $p_2$, and (iii) $s_3$, posting all 3 items. In fact, other strategies such as only posting $p_3$, or posting $p_1$ and $p_3$, are dominated. Now, we can construct the $3 \times 3$ payoff matrix for this game (where the first number in each pair corresponds to the number of followers the row blogger gets when they play a given strategy, and the second number corresponds to the column blogger's payoff):

| | $s_1$ | $s_2$ | $s_3$ |
|---|---|---|---|
| $s_1$ | $(33, 33)$ | $(16, \mathbf{50})$ | $(\mathbf{26}, 40)$ |
| $s_2$ | $(\mathbf{50}, 16)$ | $(25, 25)$ | $(20, \mathbf{30})$ |
| $s_3$ | $(40, \mathbf{26})$ | $(\mathbf{30}, 20)$ | $(25, 25)$ |

This payoff matrix is computed by considering how many followers each blogger would get, depending on their strategy and their opponent's strategy, under the greedy user model. For example, if both bloggers play strategy $s_1$, they give all 66 users utility 1, and thus they each in expectation receive 33 users. All other elements of this matrix are computed similarly. For each column, we put in bold the optimal payoff for the row player (indicating their best response), and similarly for each row we put in bold the best response for the column player. The fact that there is no pure strategy Nash equilibrium is immediately evident from the fact that no element of this payoff matrix contains both elements in bold font. This indicates that there is no scenario in which both bloggers are playing optimally, given the behavior of the other. So there is always incentive for at least one blogger to deviate from their strategy, and this proves that no pure strategy Nash equilibrium exists in this game. □

We note that though this proposition applies to a specific news generation process and scoring function, it is in fact possible to come up with many such examples, particularly in continuous content spaces $\Omega$, with continuous user interest functions $i_u$. The greedy strategy followed by users generally forbids that bloggers that only use a deterministic strategy converge to a stable Nash equilibrium. This seems at odds with the state of the blogosphere. Though chaotic, it seems to be a reasonably stable environment, with blogs following fairly consistent strategies.

Note that the greedy model seems a natural simple first step, but it generally ignores any exogenous factors differentiating bloggers, and the search costs, frictions or lack of perfect information that in practice make it difficult for a user to pick the blogger who is "optimal" for her interest. How is the expected behavior of the game altered if one changes this greedy model to assume that users *satisfice*: that is, they pick a blogger who is "good enough". In such cases, bloggers still compete for attention but, surprisingly, this relaxation of the greedy model is sufficient for the game to always reach a stable state with deterministic filtering.

## 5.2 Satisficing user models

We now discuss the theoretical results for the satisficing user models. We will prove all our results for the more general model with blogger abilities, and thus all the results hold for the model without blogger abilities as well (recall that the model without blogger abilities is a special case of the model with blogger abilities, where all bloggers have ability 1).

The way we will prove that pure strategy Nash equilibrium exist will be by using results from potential games; in particular, our approach is inspired by the proof of existence of an equilibrium in [12], following a result from [15]. We review the relevant notation, as well as this corollary, below; the notation is adapted from [15]:

We consider games $\Gamma$ with $n$ players $B = \{1, \ldots, n\}$. Each player $i$ picks a strategy from their set of strategies $\sigma_i$. The set of strategy profiles is $\Sigma = \sigma_1 \times \sigma_2 \times \ldots \times \sigma_n$. Player $i$ has payoff function $V_i \colon \Sigma \to \mathbb{R}$. For $A \subseteq B$, we denote the complement of $A$ by $-A$, and we denote the Cartesian product $\times_{i \in A} \sigma_i$ by $\Sigma_A$. For sets $A$ with one or 2 elements $\{i\}$ or $\{i, j\}$, we denote $\Sigma_{-A}$ by $\Sigma_{-i}$ or $\Sigma_{-i,j}$. Lastly, for a function $G \colon \Sigma \to \mathbb{R}$, we will sometimes use $G$ as if it were a function $G \colon \sigma_i \times \Sigma_{-i} \to \mathbb{R}$, or as a function $G \colon \sigma_i \times \sigma_j \times \Sigma_{-i,j} \to \mathbb{R}$.

Definition 1. *A game $\Gamma$ is an exact potential game if there exists a function $\Phi \colon \Sigma \to \mathbb{R}$ such that for any player $i$, any $x_i, z_i \in \sigma_i$, and any $s_{-i} \in \Sigma_{-i}$,*

$$V_i(x_i, s_{-i}) - V_i(z_i, s_{-i}) = \Phi(x_i, s_{-i}) - \Phi(z_i, s_{-i}).$$

Definition 2. *A game $\Gamma$ is a finite game if it has a finite number of players, each with a finite number of strategies (*i.e., *both $B$ and $\Sigma$ are finite).*

Now, we review Corollary 2.9 from [15], which we will use to prove that our game is an exact potential game:

Theorem 2. *A game $\Gamma$ is an exact potential game if and only if $\forall i, j \in B$, $\forall a \in \Sigma_{-i,j}$, $\forall x_i, x'_i \in \sigma_i$ and $\forall x_j, x'_j \in \sigma_j$, the following holds:*

$$V_i(x'_i, x_j, a) - V_i(x_i, x_j, a) + V_i(x'_i, x'_j, a) - V_i(x'_i, x_j, a) +$$
$$V_i(x_i, x'_j, a) - V_i(x'_i, x'_j, a) + V_i(x_i, x_j, a) - V_i(x_i, x'_j, a) = 0.$$

Now, we review a (slightly modified) version of Corollary 2.2 from [15], which we will need as well in our proof:

Theorem 3. *Every finite exact potential game possesses a pure strategy Nash equilibrium.*

Now we can prove our equilibrium result:

Theorem 4. *The blog-positioning game in the satisficing user model with blogger abilities is an exact potential game, and has a pure strategy Nash equilibrium.*

Proof. To prove the statement we use Theorem 2, adapting its notation to our setting. Let $\Sigma_{-i,j}$ denote the strategy space of all bloggers except $i$ and $j$. According to Theorem 2, we must prove that for any pair of bloggers $i, j \in B$, for any $s_{-i,j} \in \Sigma_{-i,j}$, and for any posting strategies $F_i, F'_i, F_j, F'_j \in \sigma$,

$$\begin{aligned} &V_i(F'_i, F_j, s_{-i,j}) - V_i(F_i, F_j, s_{-i,j}) + V_j(F'_i, F'_j, s_{-i,j}) + \\ -\, &V_j(F'_i, F_j, s_{-i,j}) + V_i(F_i, F'_j, s_{-i,j}) - V_i(F'_i, F'_j, s_{-i,j}) + \\ +\, &V_j(F_i, F_j, s_{-i,j}) - V_j(F_i, F'_j, s_{-i,j}) = 0 \end{aligned}$$

where $V_i(s)$ is blogger $i$'s payoff under strategy profile $s \in \Sigma$ (as above).

First, let us compute the utility of a blogger $i$ when she selects a posting strategy $F_i$. Recall that $S_u(s)$ denotes the set of bloggers which "satisfy" user $u$. We abuse notation and write $F_i \in S_u$ if blogger $i$ is in $S_u$ when she plays $F_i$. We use $U_i = \{u \in U \colon F_i \in S_u\}$ to denote the users to whom a blogger $i$ would provide their required utility if she picked a posting set equal to $F_i$. Similarly, we let $U'_i = \{u \in U \colon F'_i \in S_u\}$. $U_j$, and $U'_j$ are defined similarly for blogger $j$. Let $G(u)_{-i}$ be the random variable for the number of bloggers followed by $u$ not including $i$, when these bloggers play $s_{-i} = (F_j, s_{-i,j})$. Similarly, let $G(u)_{-i,j}$ be the random variable for the number of bloggers followed by $u$ not including $i, j$, when those bloggers play $s_{-i,j}$:

$$\begin{aligned} & V_i(F_i, F_j, s_{-i,j}) \\ = & \sum_{u \in U_i} a_i \,\mathrm{E}\left[\frac{1}{G(u)_{-i} + 1}\right] \\ = & \sum_{u \in U_i - U_j} a_i \,\mathrm{E}\left[\frac{1}{G(u)_{-i} + 1} \,\middle|\, u \notin U_j\right] \\ + & \sum_{u \in U_i \cap U_j} a_i \,\mathrm{E}\left[\frac{1}{G(u)_{-i} + 1} \,\middle|\, u \in U_j\right] \\ = & \sum_{u \in U_i - U_j} a_i \,\mathrm{E}\left[\frac{1}{G(u)_{-i,j} + 1}\right] \\ + & \sum_{u \in U_i \cap U_j} a_i \overline{a_j} \,\mathrm{E}\left[\frac{1}{G(u)_{-i,j} + 1}\right] \\ + & \sum_{u \in U_i \cap U_j} a_i a_j \,\mathrm{E}\left[\frac{1}{G(u)_{-i,j} + 2}\right], \end{aligned}$$

where $\overline{a_j} = 1 - a_j$. Subsequently, we use the following notation:

$e(u) = \mathrm{E}\left[\frac{1}{G(u)_{-i,j}+1}\right]$,

$e'(u) = \mathrm{E}\left[\frac{1}{G(u)_{-i,j}+2}\right]$.

Now, we can rewrite the utility of blogger $i$ playing $F_i$ as:

$$\begin{aligned} & a_i \sum_{u \in U_i - U_j} e(u) + a_i \overline{a_j} \sum_{u \in U_i \cap U_j} e(u) + a_i a_j \sum_{u \in U_i \cap U_j} e'(u) \\ = & a_i \sum_{u \in U_i} e(u) - a_i \sum_{u \in U_i \cap U_j} e(u) + a_i(1 - a_j) \sum_{u \in U_i \cap U_j} e(u) \\ & + a_i a_j \sum_{u \in U_i \cap U_j} e'(u) \\ = & a_i \sum_{u \in U_i} e(u) - a_i a_j \sum_{u \in U_i \cap U_j} e(u) + a_i a_j \sum_{u \in U_i \cap U_j} e'(u) \end{aligned}$$

Now we can write the formula of Theorem 2:

$$\begin{aligned} V_i(F'_i, F_j, s_{-i,j}) - V_i(F_i, F_j, s_{-i,j}) + V_j(F'_i, F'_j, s_{-i,j}) \\ -V_j(F'_i, F_j, s_{-i,j}) + V_i(F_i, F'_j, s_{-i,j}) - V_i(F'_i, F'_j, s_{-i,j}) \\ +V_j(F_i, F_j, s_{-i,j}) - V_j(F_i, F'_j, s_{-i,j}) = \end{aligned}$$

$$\begin{aligned}
a_i\sum_{u\in U_i'}e(u) &- a_ia_j\sum_{u\in U_i'\cap U_j}e(u)+a_ia_j\sum_{u\in U_i'\cap U_j}e'(u)\\
-(a_i\sum_{u\in U_i}e(u) &- a_ia_j\sum_{u\in U_i\cap U_j}e(u)+a_ia_j\sum_{u\in U_i\cap U_j}e'(u))\\
+a_j\sum_{u\in U_j'}e(u) &- a_ia_j\sum_{u\in U_i'\cap U_j'}e(u)+a_ia_j\sum_{u\in U_i'\cap U_j'}e'(u)\\
-(a_j\sum_{u\in U_j}e(u) &- a_ia_j\sum_{u\in U_i'\cap U_j}e(u)+a_ia_j\sum_{u\in U_i'\cap U_j}e'(u))\\
+a_i\sum_{u\in U_i}e(u) &- a_ia_j\sum_{u\in U_i\cap U_j'}e(u)+a_ia_j\sum_{u\in U_i\cap U_j'}e'(u)\\
-(a_i\sum_{u\in U_i'}e(u) &- a_ia_j\sum_{u\in U_i'\cap U_j'}e(u)+a_ia_j\sum_{u\in U_i'\cap U_j'}e'(u))\\
+a_j\sum_{u\in U_j}e(u) &- a_ia_j\sum_{u\in U_i\cap U_j}e(u)+a_ia_j\sum_{u\in U_i\cap U_j}e'(u)\\
-(a_j\sum_{u\in U_j'}e(u) &- a_ia_j\sum_{u\in U_i\cap U_j'}e(u)+a_ia_j\sum_{u\in U_i\cap U_j'}e'(u))
\end{aligned}$$

with one line for each of the 8 terms of the formula. Now, if you look carefully down each of the three columns of the above formula, you will notice that all the terms in each column cancel out. For example, in the first column of terms, the expressions in rows 1 and 6 cancel out. Similarly, in the second and third columns of terms, the expressions in rows 2 and 7 cancel out. Thus, adding these all together, we obtain a sum of 0, which completes the proof that the blog-positioning game is an exact potential game. Now, by simply noticing that this game is a finite game when $\Omega$ is finite, and the number of bloggers is finite, we are able to conclude that a pure-strategy Nash equilibrium always exists[7]. □

We now define the *efficiency* of a certain strategy profile $s \in \Sigma$ as the fraction or number of users who end up following at least one blogger under this configuration of the bloggers; note that this is equivalent to the sum of all the bloggers' utilities. We will now analyze the efficiency of the pure strategy Nash equilibrium of this game, by finding a bound on its price of anarchy. We recall the concept of price of anarchy:

Definition 3. *Let $W : \Sigma \to R^+$ be the social welfare function of a game, where $\Sigma$ is the space of strategy profiles. Let $E \subseteq \Sigma$ be the set of strategy profiles which are equilibria. The price of anarchy is defined as:*

$$PoA = \frac{\max_{s\in\Sigma} W(s)}{\min_{s\in E} W(s)}$$

The closer to 1 the price of anarchy of a game is, the more robust it is to its players behaving selfishly. The idea here is that if players behave selfishly, we expect their behavior to converge to an equilibrium; and if the price of anarchy is close to 1, that implies that the social welfare at this equilibrium must be close to the social optimum.

We will now prove that the blog-positioning game is efficient, in the sense that it has a low bound on its price of anarchy.

Theorem 5. *The blog-positioning game under the model of satisficing users with blogger abilities has price of anarchy at most 2.*

Proof. From the definition of the Price of Anarchy, let $s^*$ be the strategy profile in $\Sigma$ that maximizes $W(s^*)$ and let $s \in E$ be the strategy profile in $E$ that minimizes $W(s)$. For each blogger $i$, let $s_i^*$ denote the strategy of $i$ in $s^*$, and $s_i$ denote the strategy of $i$ in the equilibrium strategy profile $s$; similarly, $s_{-i}$ denotes the strategy of all bloggers except $i$ under strategy profile $s$.

Let $Sat_u(s)$ and $Sat_u(s^*)$ denote the set of bloggers which satisfy user $u$ under strategy profiles $s$ and $s^*$ respectively. Let $B_u = Sat_u(s^*) - Sat_u(s)$, the set of bloggers who satisfy $u$ under $s^*$ which didn't satisfy $u$ under $s$. Let $P_u(s)$ denote the probability that user $u$ follows at least 1 blogger under strategy profile $s$. Note that $P_u(s) = 1 - \prod_{i\in Sat_u(s)}(1-a_i)$. Similarly, $P_u(s^*) = 1 - \prod_{i\in Sat_u(s^*)}(1-a_i)$. Since the total utility of the system is defined as the number of users who follow at least one blogger, the difference between $P_u(s^*)$ and $P_u(s)$, when summed over all users $u$, must account for the difference in total welfare between $W(s^*)$ and $W(s)$. Below, we will bound the difference between $P_u(s^*)$ and $P_u(s)$.

We first define the function $P_u'(\cdot)$, such that for any subset $A$ of the bloggers $B$, $P_u'(A) = 1 - \prod_{i\in A}(1-a_i)$ denotes the probability that $u$ follows at least 1 blogger in $A$, under the assumption that $u$ is satisfied by all bloggers in $A$. Thus $P_u(s) = P_u'(Sat_u(s))$. Now, for any blogger $i$, user $u$, and strategy vector $s$, let $b_{i,u}(s)$ denote the expected attention blogger $i$ receives from user $u$ under strategy vector $s$. Note that for $i \in Sat_u(s)$, $b_{i,u}(s) = a_i \operatorname{E}\left[\frac{1}{1+G(u,s)_{-i}}\right]$, where $G(u,s)_{-i}$ denotes the random variable for the number of bloggers followed by $u$, not including blogger $i$, under strategy vector $s$. We also use $G(u,s)$ to denote the random variable for the total number of bloggers followed by $u$ under strategy vector $s$.

To show the claim, we first prove two technical lemmas:

Lemma 1. *Let $\{x_1,\ldots,x_n\}$ be a set of numbers $x_i$ where $0 \le x_i \le 1$. Then $\sum_{i=1}^n x_i \ge 1 - \prod_{i=1}^n(1-x_i)$*

Proof of Lemma 1: We use induction to prove the equivalent inequality: $\prod_{i=1}^n(1-x_i) \ge 1 - \sum_{i=1}^n x_i$. As the base case ($n=1$), note that $\prod_{i=1}^n(1-x_i) = 1-x_1 = 1-\sum_{i=1}^n x_i$, so the inequality holds. For the inductive case, note the following: $\prod_{i=1}^{n+1}(1-x_i) = (1-x_{n+1})\prod_{i=1}^n(1-x_i) \ge (1-x_{n+1})(1-\sum_{i=1}^n x_i) = (1-\sum_{i=1}^n x_i) - x_{n+1} + x_{n+1}\sum_{i=1}^n x_i \ge 1 - \sum_{i=1}^{n+1} x_i$. □

Lemma 2. *If $\{A_1, A_2\}$ is a partition of a set $A \subseteq B$, then $P_u'(A) = P_u'(A_1) + (1 - P_u'(A_1))P_u'(A_2)$.*

Proof of Lemma 2: Recall that $P_u'(A) = 1 - \prod_{i\in A}(1-a_i)$. From this definition we can directly prove the claim:

[7] These results extend to the case of an infinite and/or continuous content space $\Omega$. No matter the nature of the content space, the space $\sigma$ of posting strategies can always be divided into $2^{|U|}$ discrete areas, where each area corresponds to the set of posting strategies that would result in a subset $Z \subseteq U$ of the users being "satisfied". A blogger's payoff simply depends on which of these areas she chooses her posting strategy from. Thus, regardless of the nature of the content space, or of the user's utility functions, this can be seen as a finite game.

$$\begin{aligned}
& P'_u(A_1) + (1 - P'_u(A_1))P'_u(A_2) \\
= \; & 1 - \prod_{i \in A_1}(1 - a_i) + (\prod_{i \in A_1}(1 - a_i))(1 - \prod_{i \in A_2}(1 - a_i)) \\
= \; & 1 - \prod_{i \in A_1}(1 - a_i) + \prod_{i \in A_1}(1 - a_i) - \prod_{i \in A_1 \cup A_2}(1 - a_i)) \\
= \; & 1 - \prod_{i \in A_1 \cup A_2}(1 - a_i)) = P'_u(A) \qquad \square
\end{aligned}$$

Now we will prove that for any user $u$:

$$P_u(s) + \textstyle\sum_{i \in B_u} b_{i,u}(s_i^*, s_{-i}) \geq P_u(s^*).$$

In fact,

$$\begin{aligned}
& P_u(s) + \sum_{i \in B_u} b_{i,u}(s_i^*, s_{-i}) \\
= \; & P_u(s) + \sum_{i \in B_u} a_i \, \mathrm{E}\left[\frac{1}{1 + G(u,s)_{-i}}\right] \\
\geq \; & P_u(s) + \sum_{i \in B_u} a_i \Pr[G(u,s)_{-i} = 0] \\
= \; & P_u(s) + \sum_{i \in B_u} a_i \Pr[G(u,s) = 0] \\
= \; & P_u(s) + \sum_{i \in B_u} a_i (1 - P_u(s)) \\
= \; & P_u(s) + (1 - P_u(s)) \sum_{i \in B_u} a_i \\
\geq \; & P_u(s) + (1 - P_u(s))(1 - \prod_{i \in B_u}(1 - a_i)) \\
= \; & P'_u(Sat_u(s)) \\
+ \; & (1 - P'_u(Sat_u(s))) \cdot P'_u(B_u) \\
\geq \; & P'_u(Sat_u(s^*) \cap Sat_u(s)) \\
+ \; & \Big( (1 - P'_u(Sat_u(s^*) \cap Sat_u(s))) \\
* \; & P'_u(Sat_u(s^*) - Sat_u(s)) \Big) \\
= \; & P'_u(Sat_u(s^*)) \\
= \; & P_u(s^*),
\end{aligned}$$

where the third step follows from the fact that user $u$ definitely does not follow blogger $i$ under strategy vector $s$ (because $i \in B_u$, the set of bloggers which satisfy $u$ in $s^*$ but not $s$), so $G(u,s) = G(u,s)_{-i}$, the sixth follows from Lemma 1, the seventh by the definitions of $P_u$ and $P'_u$, the eighth from the fact that $P'_u(Sat_u(s)) \geq P'_u(Sat_u(s^*) \cap Sat_u(s))$, and the ninth from Lemma 2. All the other steps follow from the definitions or direct algebraic manipulations.

Now we are ready to prove the main claim that $2W(s) \geq W(s^*)$. To do this, we take the final inequality, and sum over all users $u$:

$$\begin{aligned}
P_u(s^*) - P_u(s) \quad & \leq \quad \sum_{i \in B_u} b_{i,u}(s_i^*, s_{-i}) \\
\Rightarrow \sum_u (P_u(s^*) - P_u(s)) \quad & \leq \quad \sum_u \sum_{i \in B_u} b_{i,u}(s_i^*, s_{-i}).
\end{aligned}$$

Notice, that $\sum_u P_u(s^*) - \sum_u P_u(s) = W(s^*) - W(s)$. So:

$$\begin{aligned}
W(s^*) - W(s) \quad & \leq \quad \sum_u \sum_{i \in B_u} b_{i,u}(s_i^*, s_{-i}) \\
& = \quad \sum_{i \in B} \sum_{u : i \in B_u} b_{i,u}(s_i^*, s_{-i}) \\
& \leq \quad \sum_{i \in B} \sum_{u \in U} b_{i,u}(s_i^*, s_{-i}) \\
& = \quad \sum_{i \in B} V_i(s_i^*, s_{-i}) \\
& \leq \quad \sum_{i \in B} V_i(s_i, s_{-i}) \\
& = \quad W(s).
\end{aligned}$$

Note that the last inequality follows from the fact that $V_i(s_i^*, s_{-i}) \leq V_i(s_i, s_{-i})$, because $s$ is a Nash equilibrium. This concludes the proof that $2W(s) \geq W(s^*)$. □

We note that one can also use the framework of [17] to prove Theorem 5, but this requires a somewhat laborious demonstration that the blog game is a "valid utility system"; in our setting, it is easier to give a direct bound. Furthermore, this framework cannot be used to prove Theorem 4 and similar results on the existence of pure strategy equilibria.[8]

Next we prove that the theorem above is essentially tight.

Proposition 6. *The price of anarchy of the system in the satisficing user model with blogger abilities is at least $2 - \epsilon$ for any $\epsilon > 0$.*

Proof. We present a simple equilibrium example with the claimed gap. Suppose that in the system there are only $n$ bloggers, each with ability 1, and $2n-1$ users. Furthermore, suppose that $n$ users are interested in the item in position $p_1$ and give positive score only to it, and each user $n+i$ is interested only in the item in position $i$ and gives positive score only to it.

In this scenario there is a strategy $s^*$ that covers all users; one blogger covers the first $n$ users and then each other blogger can cover a distinct user. So strategy $s^*$ has welfare $2n-1$. But now consider the equilibrium in which all the bloggers are covering the first $n$ users; this is an equilibrium because no blogger can increase her share of user attention by moving, and the welfare of this equilibrium is $n$. The PoA is then equal to $\frac{2n-1}{n}$. Thus, $\epsilon > \frac{1}{n}$ yields the claim. □

# 6. DISCUSSION

Our results demonstrated a new prevalent effect in social media: Users of various activity levels exert different content selection within a topic, varying from broad-interest posts to more niche content. As our empirical results suggest, this phenomenon is present at various scales, for various social media and topics. This observation has important consequences. Most importantly even a population of intermediaries with various abilities can collectively serve efficiently the interests of an arbitrary population of users. A simple reward by audience size, that is naturally implemented today by advertising, is sufficient. However, this result is a non-trivial one as it depends on reader's aggressiveness to react to various news offerings.

Our work relies on a few assumptions that also point to interesting future research. First of all, we deliberately avoided domain specific metrics of quality and success in

[8] The blog positioning game does not satisfy the requirements of a "basic utility system", which would be required to prove the existence of a pure strategy Nash equilibrium. Therefore, Theorem 4 does not follow from previous work of [17] and others.

our empirical study, to study how blogs behave in function of popularity or quality within a topic. This was shown to be fruitful as it applies well to blogs focusing on a single topic of expertise. However, when more domain specific metrics are available, relating filtering to multiple dimensions such as political leaning and diversity of topics can bring additional insight on the ingredient of successful information intermediaries. Second, in our model a two-hop information flow is sufficient to explain how a few selected news reach a very large audience. Indeed, recent results highlighted that viral dissemination generally goes only through a few hops [8]. Nevertheless, a generalization of information filtering through multi-hop dissemination could offer even more insights. As an interesting example, we note that all our results extend to the case where users cannot pick any blog but need to choose one in a given *subset*. Many related topics – such as predicting the success of a post or a user, and understanding information cascades – can be revisited using a game theoretical model of information intermediaries. Our work is a stepping stone to understand these issues as it provides such a tool and shows how it already relates to users' behavior in today's social media.

## 7. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation, under grant ICES-1139915 and CNS-1254035. We are grateful to `Texifter` for making data-sets collected with `DiscoverText.com` available to us, and especially to Stuart Shulman and Josh Sowalsky for their time and guidance. We also thank participants of the 3rd Workshop on Social Computing and User Generated Content, The New York Computer Science and Economics day, and the first Interdisciplinary Workshop on Information and Decision in Social Networks, for early feedback and discussion.

## 8. REFERENCES

[1] Z. Abbassi, N. Hegde, and L. Massoulié. Distributed Content Curation on the Web. *Proceedings of ACM WPIN-NETECON Workshop*, 2013.

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

[3] S. Athey and M. Mobius. The impact of news aggregators on internet news consumption: The case of localization. *working paper*, pages 1–36, 2012.

[4] L. Backstrom, J. M. Kleinberg, and R. Kumar. Optimizing web traffic via the media scheduling problem. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009.

[5] K. Burton, N. Kasch, and I. Soboroff. The ICWSM 2011 spinn3r dataset. *Proceedings of the International Conference Weblogs and Social Media (ICWSM)*, 2011.

[6] M. Cha, F. Benevenuto, H. Haddadi, and K. 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.

[7] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, 2007.

[8] S. Goel, D. J. Watts, and D. G. Goldstein. The structure of online diffusion networks. In *EC '12: Proceedings of the 13th ACM Conference on Electronic Commerce*, 2012.

[9] M. Gupte, M. Hajiaghayi, L. Han, L. Iftode, P. Shankar, and R. Ursu. News posting by strategic users in a social network. *WINE '09: Proceedings of the 5th International Workshop on Internet and Network Economics*, pages 632–639, 2009.

[10] P. R. Jordan, U. Nadav, K. Punera, A. Skrzypacz, and G. Varghese. Lattice games and the economics of aggregators. *WWW '12: Proceedings of the 21st international conference on World Wide Web*, 2012.

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

[12] J. M. Kleinberg and S. Oren. Mechanisms for (mis)allocating scientific credit. In *STOC '11: Proceedings of the 43rd annual ACM symposium on Theory of computing*, 2011.

[13] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, 2010.

[14] J. Leskovec, L. Backstrom, and J. M. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009.

[15] D. Monderer and L. S. Shapley. Potential games. *Games and Economic Behavior*, 14(1):124–143, 1996.

[16] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida. On word-of-mouth based discovery of the web. In *IMC '11: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, 2011.

[17] A. Vetta. Nash equilibria in competitive societies, with applications to facility location, traffic routing and auctions. In *FOCS: Proceedings of the The 43rd Annual IEEE Symposium on Foundations of Computer Science*, pages 416–425, 2002.

[18] S. Waldman. The Information Need of Communities. *Report to the Federal Communications Commission (FCC)*, pages 1–478, 2011.

[19] F. Wu and B. A. Huberman. Popularity, novelty and attention. *EC '08: Proceedings of the 9th ACM conference on Electronic commerce*, 2008.

[20] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. *WWW '11: Proceedings of the 20th international conference on World wide web*, 2011.

[21] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM '11: Proceedings of the fourth ACM international conference on Web search and data mining*, 2011.

[22] R. B. Zadeh, A. Goel, K. Munagala, and A. Sharma. On the precision of social and information networks. In *COSN '13: Proceedings of the first ACM conference on Online social networks*, 2013.