Game in the Newsroom: Greedy Bloggers for Picky Audience

AVNER MAY, Columbia University
AUGUSTIN CHAINTREAU, Columbia University
NITISH KORULA, Google Research, New York
SILVIO LATTANZI, Google Research, New York

The impact of blogs and microblogging on the consumption of news is dramatic. This has been recently quantified, as data from microblogging websites measured the importance of intermediaries in the dissemination of news content to a large audience. In this work, we analyze experimentally and theoretically the dynamics of users serving as intermediaries between the mass media and a general audience. Our first contribution is to describe at a macroscopic level the posting behaviors of today's social media users, as well as which posting strategies correlate with a user's popularity. We show that users obey a "filtering law", meaning that less active users post more popular content on average than more active users; we do this by studying jointly for the first time the volume and popularity of URLs shared by users. We also show that users who selectively post the most important content generally have more followers, across activity levels. We perform this analysis on 10 social media data sets, and show that our observations are consistent across them. As our second contribution, we attempt to understand these posting behaviors by introducing a blog positioning game and showing that it can lead to equilibria in which the needs of a picky audience are optimized, while bloggers are naturally rewarded with followers for the overall system's improvement. Interestingly, this model predicts that greedy readers often destabilize the system, while a bit of leniency by the readers in choosing bloggers to follow is enough to stabilize it.

> "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*, 2011.

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* rely anymore on traditional media to set *what* they should see. Increasingly, people rely on others (whether friends or other individuals active on 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: what content do users decide to post? and based on this, whose posts does an individual decide to pay attention to?

Most studies to date focus on understanding how some news can become overly popular through such sharing. Assuming the graph is fixed and known, how to identify, predict and sometimes encourage popular exposure all seem natural research questions. Our hypothesis turns this problem upside down: we aim at exploring how a graph can emerge from users decisions that put the right information in front of them. Furthermore we want to understand the efficiency of such a structure.

First, we show a "filtering law" which predicts that even users of low activity can play an intermediary role by selecting only the most important news. This result also highlights that the relationship between the volume and the importance/quality of an intermediary's posts can be exploited by users advantageously to choose the amount of news they wish to receive. Second, we show that users who selectively post the most popular content, across activity levels, generally have more followers. A potential explanation of this result, which we investigate at length, is that the curation choices made by an intermediary can be a result of their potential audience, as well as of competition with others to retain it. For this reason, we choose to analyze a simplified model in which intermediaries and users are both perfectly strategic and may modify the graph at any time. While, even for very simple such games, instability is produced by users behaving greedily, we prove that a graph emerges that always satisfies at least half of the maximum number of users when users instead follow a simple satisficing strategy. Finally our results quantify for the first time the efficiency of social media as a curation system, and more generally demonstrate its promise.

2. RELATED WORK

Our work follows the classic hypothesis of a *two-step information flow* [Katz 1957], where opinion leaders that are also news savvy play the key role of intermediaries between media producing content and a large audience. This thesis was revived using empirical evidence of a similar effect occurring on Twitter [Kwak et al. 2010; Wu et al. 2011], and more recently identifying mass media and intermediaries as critical to information spread [Cha et al. 2012]. In fact, our work extends this hypothesis to model how even relatively normal users with moderate activity 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 or equilibrate viewpoints [An et al. 2011]. On the other hand, study of word-of-mouth showed propagation happens more likely between geographically closed friends, hinting at a possible bias effect [Benevenuto et al. 2011]. Our work is complementary as we explore a new dimension: how intermediaries help manage the *volume* of information by adapting to the interests of the audience. In contrast with previous work, we wish to go a step beyond a pure empirical approach: first, by proving that intermediaries emerge naturally from a collaborative curation ecosystem, and second, by informing the design of efficient incentives for such information sharing systems.

As news moves online, how to best serve the various information needs of communities remains an important issue of public concern [Waldman 2011]. Previous works analyzed how to best serve this collective attention, either by analyzing temporal dynamics and treating this as a scheduling problem [Wu and Huberman 2008; Backstrom et al. 2009; Leskovec et al. 2009; Yang and Leskovec 2011], by adapting personalization tools [Das et al. 2007] or by enhancing news navigation [Ahmed et al. 2011; Shahaf and Guestrin 2012]. They typically ignore competitions among different news offerings. A recent and notable exception is the recent debate about the potential damaging effect of news aggregators on the revenue of traditional content publishers [Krakovsky 2013]. Recent results showed that the increase of traffic they generate to news outlet comes at the expense of customer's loyalty [Athey et al. 2012], leading to potential loss of advertising revenue [Athey and Mobius 2012]. Our work is radically different as we consider competition *between* various intermediaries as well. We wish to harness the collective behavior of intermediaries and align their incentive to serve each and every information need. Our game theoretic approach resembles recent studies on the efficiency of crowdsourcing [Ghosh and McAfee 2011], but it is applied for the first time in the context of content curation.

We are only aware of a few works analyzing news dissemination with a game theoretic background. Previous work assumed news are equally appealing to all users [Gupte et al. 2009], and showed that a local incentive suffices for higher quality news to reach a large audience in random graph, by a connectivity argument. It also shows users attempting to avoid spam may be an important limiting factor. In contrast, we are interested in how various blogs emerge to address the need of a heterogeneous audience. In [Jordan et al. 2012], 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 how many, and what set of articles to post *within* a given topic; this allows players to leverage another differentiation. Our analysis needs to account for a repeated game as blogs may decide to increase or decrease the volume of information with time to appeal to a different population. We wish to find the conditions under which such games admit a natural equilibrium that is also efficient.

3. DO BLOGS FILTER? SHOULD THEY?

Our data sets. In order to begin to understand the different ways in which intermediaries are behaving as quality filters in social media, and what posting strategies are associated with more followers, we need to gather information about (1) what users post in various media, (2) a way in which this information can be judged for its quality, (3) some measure of the success a user has as an intermediary of information. We rely on simple metrics of quality and success that we explain below. Note that for a given content item or a given user they may not be absolutely accurate, however we believe that such measures are sufficient to make relative statistical comparison and observe collective trends. Since, to the best of our knowledge, no data set available today allows studying all of these factors together, we gathered data from various sources and conducted some complementary crawls, that we now describe.

3.1. Collecting and Preprocessing data

We first gather data on users' *posts* in social media. In order to avoid the complexity of understanding the way in which bloggers choose what content to post across various topics, we choose to focus on data sets in which all the posts pertain to the same topic. We study posts during major events, which create blockbuster URLs, as well as less popular niche content.

Table I. Users' Posting activity (left) and URL Popularity (on the right) for each of our datasets, all measured inside the data sets (by post) and through external crawl: The Median for all metrics is low, it indicates that more than half of the users (resp. URLs) are not that active (resp. not that popular). On the contrary, 95% percentile and average can be large, indicating some very active users (resp. very popular URLs). Bloggers have comparatively a high activity (more engaged), in contrast with FB users (less engaged).

Data sets	Users (% live today on Twitter)	Activity Distribution mean [median - 95%perc.]		URLs (% with click count from bit.ly)	Popularity Distribution mean [median - 95%perc.]	
		Posts	Tweets/day		Posts	Bit.ly clicks
TW Bin Laden	700,783 (75.9%)	2.4 [1 - 6]	9.6 [3.3 - 38.9]	545,495 (19.7%)	3.1 [1 - 8]	917 [7 - 713]
TW Occupy WS	354,117 (88.9%)	3.1 [1 - 10]	8.7 [2.7 - 35.2]	316,408 (26.9%)	3.5 [1 - 9]	550 [7 - 664]
TW Steve Jobs	719,025 (86.8%)	1.5 [1 - 4]	8.3 [2.8 - 33.9]	250,644 (20.0%)	4.5 [1 - 8]	1,189 [8 - 1,275]
TW iPhone5	81,056 (94.6%)	1.3 [1 - 2]	19.6 [5.8 - 74.6]	37,323 (30.7%)	2.7 [1 - 7]	2,632 [34 - 3,940]
FB iPhone5	330,185 (N/A)	1.41 [1 - 3]	N/A	193,024 (14.6%)	2.4 [1 - 4]	1,857 [36 - 3,811]
				"likes": 45,792 URLs, popularity 8.5 [2 - 19]		
Blogs All	67,692 (N/A)	16.5 [2 - 42]	N/A	440,933 (30.9%)	2.5 [1 - 7]	398 [16 - 914]
Blogs Obama	13,390 (N/A)	14.7 [2 - 44]	N/A	84,733 (40.9%)	2.3 [1 - 7]	769 [23 - 1,324]
Blogs Facebook	11,643 (N/A)	10.1 [2 - 23]	N/A	69,747 (40.5%)	1.7 [1 - 4]	977 [27 - 1,778]
Blogs Euro	9,659 (N/A)	9.5 [2 - 23]	N/A	53,001 (39.9%)	1.7 [1 - 5]	944 [20 - 1,563]
Blogs Mubarak	6,546 (N/A)	13.7 [2 - 37]	N/A	42,531 (34.6%)	2.1 [1 - 7]	1,214 [19 - 1,304]

— **Microblogging.** Data was gathered using `DiscoverText.com` to harvest tweets from the Twitter ("TW") public API, for various events[1]: the deaths of Steve Job's and Osama Bin Laden, the Occupy Wall Street movement, and the launch of iPhone 5.

— **Online Social Networks.** We also use a set of 1m Facebook comments collected for one event (iPhone 5 release) from the GNIP Facebook API.

[1]As an example, the Bin Laden data set contains 1,977,716 tweets with the words "Osama" or "Bin Laden" posted on the night and day following his death (from 5/2/2011 at 3:30am EST to 5/3/2011 at 1:30pm EST).

— **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 [Burton et al. 2011]. It contains 386m blog posts collected from spinn3r during the months of January-February 2011.

We do acknowledge that, in the first two data sets, some users are not bloggers per se, even if we already restrict to those users posting an article containing a topic-relevant URL. We already saw that considering only users with sufficient importance (*e.g.,* a minimum number of followers) identifies bloggers quite well and that our results generalize. This will be further analyzed in a future version of this work.

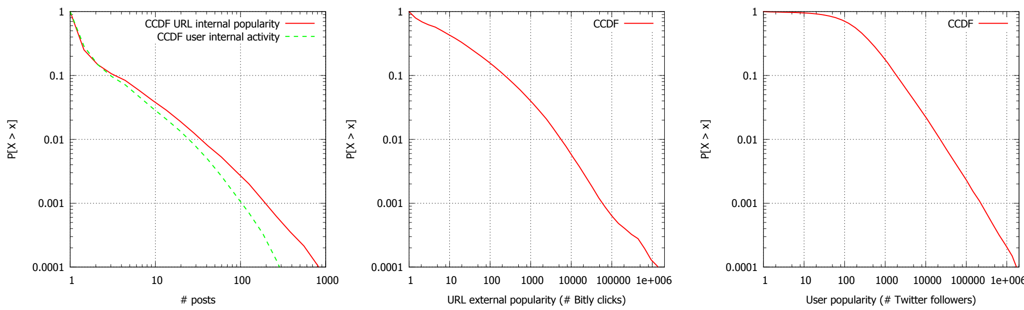
We crawled each of these data to gather all posts, tweets and comments that contain URLs. Many of these are duplicate with formatting variants and URL shorteners, so we first run a normalization technique by (1) following all redirects using HTTP call, (2) removing headers (`http/https`) and suffix (anything following a # sign, ".html", or ".htm") (3) removing spam[2]. Our data sets vary considerably in size, ranging from between forty thousand and a million unique normalized URLs shared.

A user's activity (resp. a URL popularity), can directly be measured internally in the data-set by counting the number of posts (resp. the number of times a URL is posted). Active users and popular URLs are more likely to be high in this *internal* metric. However, and especially when the data set is small, it is important to validate our results also using *external* measures of activity and popularity. First, this allows us to remove the effect of the sampling bias: Since there is no way to ensure that no post is missing, a user may be more active or a URL more popular than the data set reveals. Second, it offers an overview of activity and popularity that is not impacted by the particular search term that were applied to gather these data.

We obtained an external measure of a user's activity by crawling through the Twitter public API for their account creation's time and total number of tweets. That only applies to accounts that are still live today after the event, which are fortunately always larger than 75% more than year after the event, and 94% when the data had been collected for a couple of months. Similarly, we obtained an external measure of a normalized URL popularity by crawling for its bit.ly shortener and collecting the number of clicks it receives.[3] For Facebook, since each comment can receive one or several "likes" (24% of them do so), we can compute the popularity of a URL as a function of the total number of likes received by comments containing it in our dataset.

We plot internal and external metrics of URL's popularity and user's activity and success. As seen in Table I, user activity and URL popularity have very low median and very high variance: a majority of users (resp. URLs) post only once or twice (resp. are only posted once), but a handful of users or URLs are disproportionately active or popular. We observed that all metrics on all data sets are power-law distributed on their [90%-percentile, 99.9%-percentile] range. One consequence is that the average is

[2]Since most of these data-sets are some months old, spam URL have generally been blacklisted by now, and in particular all bit.ly links redirect to a standard "warning" page. We treat any such redirects as spam and remove all users that have posted at least one malicious URLs.
[3]To compute the "Bit.ly click count" of a URL, we consider several variants of the URL which could link to the same page. We compute the list of variants by first computing the "core" of the URL (we remove the protocol, any trailing slashes, and everything after ".htm",".html", "#". We then query the Bit.ly API for the number of clicks received by all of the following URLs: 1) All the original URLs with the same "core", 2) The four variants of the URL, built from the URL core by different combinations of the URL's protocol (http or https) and whether or not it has a trailing slash. We then consider the popularity of a URL with this core to be the sum of the number of Bit.ly clicks received by each of these versions of the URL. We observe that between 20-40% of our URLs have a bit.ly shorteners, especially since Bit.ly is the official URL shortener for several media (*e.g.,* nyti.ms, wapo.st, ti.me).



(a) Internal Metrics (# posts) (b) Ext. Popularity (# bit.ly clicks) (c) Ext. "Success" (# followers)

Fig. 1. Marginal distributions of various activity, popularity, and success metrics (TW Bin Laden)

generally well above the median and close to the 90%-percentile of the distribution. It is especially pronounced for external popularity measures.

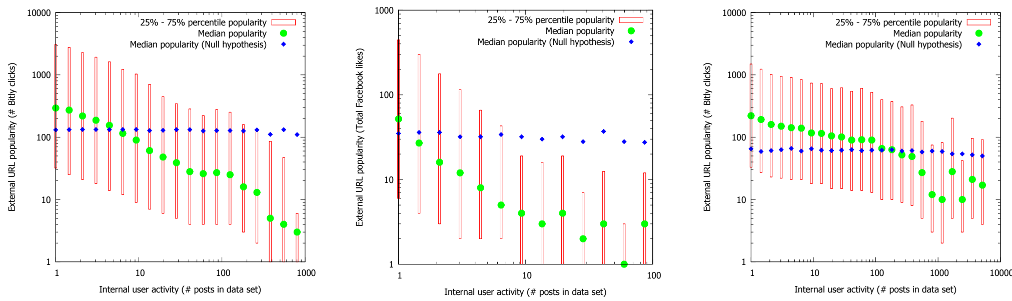
Finally, to judge the success of a user as an information intermediary, we simply crawl Twitter for her number of followers. This simplistic measure is generally noisy especially for very large values (most people above 100,000 followers [the top 0.08% of users] are either organizations or famous celebrities who do not necessarily relay other information), but for the medium range from 200 (median amount of followers) to 20,000 (top 1% of users in terms of amount of followers), we found this metric reliable to estimate the relative importance that a node has reached. For a user of our Facebook data, we compute the relative measure of her success as the total number of likes received by her comments.

3.2. Activity-Popularity Tradeoff

We study for the first time the dependence between the activity of a user and the popularity of the URLs she posts. One intuition is that users who are more active are more engaged, hence they better select the URLs they post. In addition, they may have in general more followers (see below) and hence the URLs they choose may as an effect become more popular.

To analyze this trend, we place users in various bins according to the number of URLs they post (*i.e.,* their internal activity), and plot in Figure 2 the popularity distributions (in bit.ly clicks) of the URLs in each bin. Surprisingly we observe that users who are less active pick their content disproportionately among the most popular URLs: users who post only once in TW Bin Laden pick half of the time a URL with at least 300 bit.ly clicks (*i.e.,* a URL in the top 8% most popular); they pick one with at least 3000 bit.ly (*i.e.,* in the top 2%) a quarter of the time. In comparison, those who posted 40 times already see a median popularity of 30 clicks, and the most active who posted 800 URLs see 75% of these with less than 7 clicks (*i.e.,* bottom 50% URLs). In fact, all percentiles appear to decrease linearly in a log-log plot, across all levels of activity.

Validation with Null Hypothesis. One might argue that the observations can simply be explained by the number of URLs shared by a user. A user posting only a few URLs can pick them to be extremely popular, but perhaps a user posting many more "runs out" of these blockbuster URLs and is forced to post less popular content. We prove that this effect is not enough to explain the observed trend; we simulate a null hy-



(a) TW Bin Laden (b) FB iPhone5 (c) Blog All

Fig. 2. The filtering law *"Users who post less tends to post popular URL disproportionately"*. (x-axis) user internal activity in number of URLs, (y-axis) popularity (bit.ly, likes) of URLs posted by these users.

pothesis model in which all users post URLs, without repetition, according to the same popularity distribution, by reshuffling all edges representing posts in the user-URL bipartite graph while preserving the degree of users (*i.e.,* their internal activity) and URLs (*i.e.,* their internal popularity). We compare the results of these null hypotheses to the trends in the real data, and find that they differ entirely: in the null hypothesis, URL popularity is essentially uncorrelated with the activity level of the poster. This proves that the filtering law is significantly stronger than what a random model predicts (note in addition that y-axis here is in logarithmic scale).

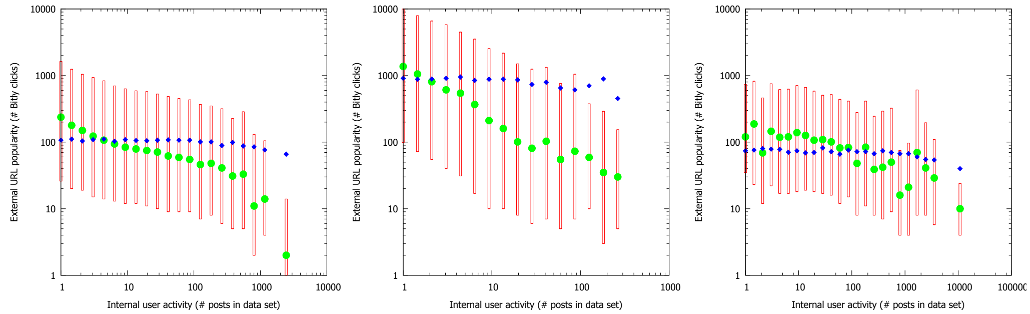
We also obtained preliminary evidence indicating that the filtering law is not an effect of less active users receiving more popular URLs, another potential confounding factor. This will be reported in a future version of this paper.

Validation with other data sets and metric. We observe that this phenomena generalizes to posts on social networking sites (Fig. 2 (b)) and Blogs (Fig. 2 (c)), and to popularity metrics like Facebook "likes". Interestingly, the popularity drop remains always linear in log-log plot; furthermore, it is always steeper as users in the network on average post less: Facebook's filtering effect is more pronounced, blogs see a milder effect and they typically have more engaged users. It holds across topics for all data sets, as shown in Figure 3. It also holds for other metrics: Figure 4 plots the same results when users are grouped according to their activity as observed externally on Twitter. Except for very passive users who post less than once every ten days, we see the exact same trend as with internal activity, with remarkably similar numbers (in Bin Laden data set, median of 300 clicks for people posting daily, dropping to 30 clicks as people post on average 40 times a day). Again comparison with the null hypothesis proves that this effect is significant. The only exception is a couple of points at the far right: extremely famous Twitter accounts like CNN breaking news (@*cnnbrk*), which post a lot of links that become instantly popular due to their direct followers.

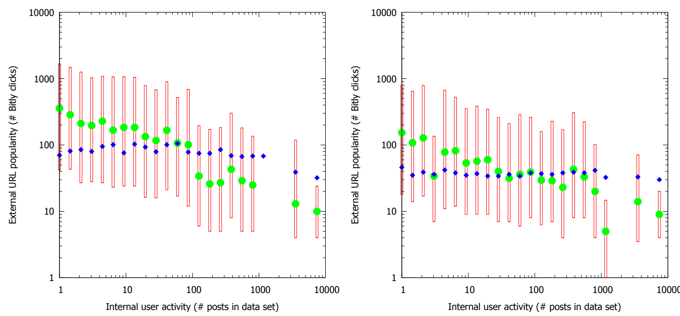
3.3. What make successful intermediaries?

We have seen that filtering is latent in today's social media, as users with various activity levels post content differently. We now wish to study if filtering is a competitive advantage for a user playing the role of an information intermediary. This motivates us to study more generally the main factors that correlate with metrics of success.

Activity correlates with success, but shows a diminishing return. Figure 5 plots the raw measure of success (followers in Twitter, number of likes received by their comments in Facebook) for users of various activity levels. While we observe that activity

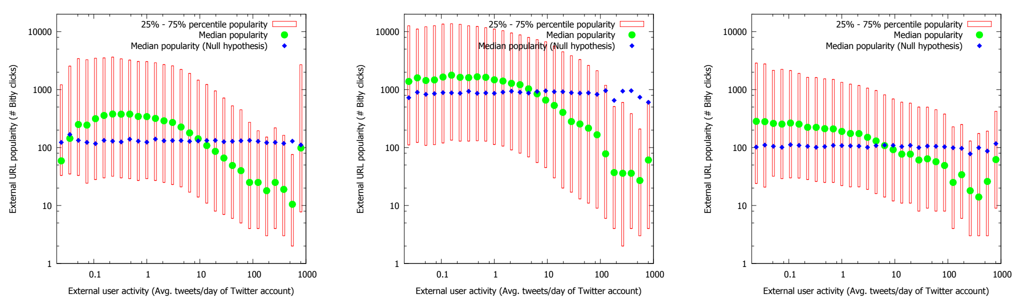


(TW Occupy) (TW Steve Jobs) (Blog Obama)



(Blog Mubarak) (Blog Euro)

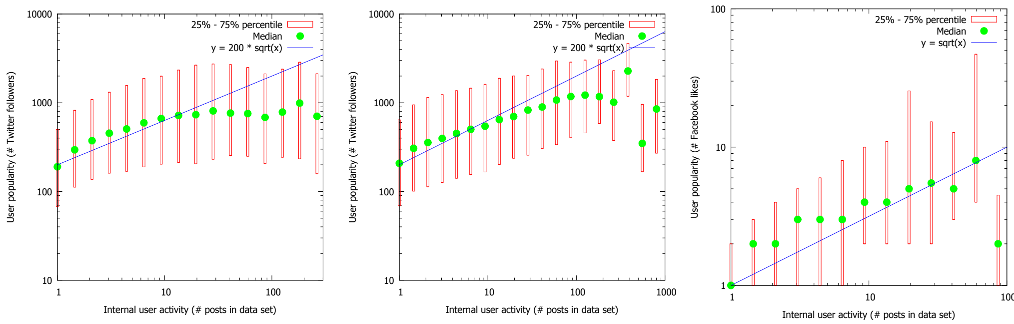
Fig. 3. The filtering law validated on more data sets: (x axis) internal activity in # of posts, (y axis) external popularity in # of bit.ly clicks.



(a) TW Bin Laden (b) TW Steve Jobs (c) TW Occupy WS

Fig. 4. The filtering law observed using external activity: (x-axis) activity in Tweets/day, (y axis) external popularity in bit.ly clicks. URLs clicks gradually drops with User's activity, except for passive users (far left), and a few famous accounts (far right).

is in general positively correlated with success, we see that they are not related in a linear manner. This trend is observed in all data sets we have, on almost the full range of activity (with the exception of very high value in which popularity drops). We also observed the same trend to hold for external activity for the range of users tweeting between 0.1 and 1000 tweets a day. It may indicate that activity is an ingredient to get noticed and adopted as information intermediaries. As a reference point, 50-75% of accounts posting twenty tweets a day have above 300 followers; this is the case only



(a) TW Bin Laden (b) TW Occupy WS (c) FB iPhone5

Fig. 5. The diminishing return of posting activity: (x-axis) user internal activity in number of URLs, (y-axis) user "success" (followers, likes). We include the plot of the square-root function as a common benchmark in these plots.

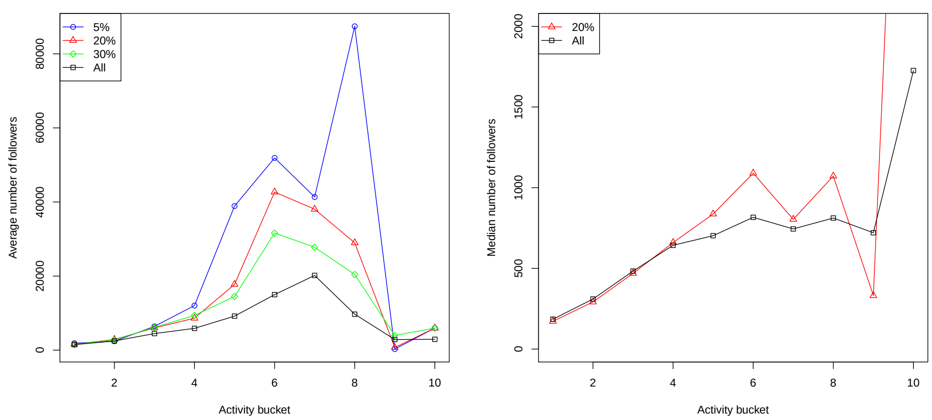
for 25% of the account who post once a day. Note that it is also possible that popular people also tends to become more active. In the plots of Figure 5 we include the

One can also interpret this result naturally in light of our previous experiments. Posting more often correlates with more occasions to get noticed and gather followers and likes. At the same time, it correlates with posting URLs that have less popular appeal, which implies that their effect on the overall success as an information intermediaries has a diminishing return.

The role of filtering. To understand further how filtering plays a role in the success of information intermediaries, we analyzed in our data sets whether users who post more 'selectively' in general have more followers. To do this, we assigned to each user a selectivity score which is the median popularity of the URLs they posted, in terms of the number of bit.ly clicks. In order to control for the effect of activity on blogger success, we bucketed all users by their activity level. Then, within each activity level, we assigned each user a 'selectivity percentile' score based on how selective they were relative to the other users in their bucket (a score of 100 means a user is the most selective user at their activity level, 0 is least selective). We then considered 4 groups of users; the 5% most selective users, 20% most selective, 30% most selective, and the set of all users. In Figure 6 (a), we show the mean number of followers of all these groups of users, bucketing the users in each group by their internal activity level. In Figure 6 (b), we show similar results for the median number of followers; as the trend is less pronounced, the lines are harder to distinguish and we only show one selective group. For example, the blue line in figure 6 (a) represents the top 5% most selective users; we bucketize these users by their internal activity level on the x-axis (most active users on the right), and consider the mean number of followers of users in these buckets. As an example, the 5th least active bucket here has an average number of followers around 4000. As you can see, in general the more selective a group of users is, the higher their average follower count. This trend is consistent, with the mean number of followers decreasing as the selectivity of the user group decreases. In the 6th least active bucket of users, for example, the top 5% most selective users have an average of around 5000 followers, compared with around 4000, 3000, and 1500 followers for the 20%, 30%, and complete groups respectively. The trend is less striking in the case of the median number of followers for these various groups of users (figure $6b$), but even here we see a higher median number of followers amongst the more selective users, especially at the higher activity levels. The fact that the trend is much stronger for the mean can be explained by there being a larger fraction of users who are very successful amongst the most selective users, thus raising the average number of followers considerably, but not the median as much.

Posting early correlates with success. We have seen that people who post more, and post more selectively, generally have more followers. One more question we wanted to answer was: do users who post links quickly (eg, they are generally one of the first to post a given article) generally have more followers as well? To do that, we analyzed a data set containing all 450k “nyti.ms” links (NYTimes.com articles shortened with URL shortener) shared on Twitter over 2 weeks (between 12/15/12 and 12/27/12). Given that within this data set, each article was generally shared many times, this data set allowed us to classify each user by how early they generally posted links, within their activity level. We assigned each user a percentile based on their average “earliness score” relative to other users at their activity level. We then compared the median and average number of followers of the top 10% earliest users vs. the entire population, within each activity level. As you can see in figure 7, users who post content earlier generally have more followers (note that these are log-log plots, so the difference in the average/median number of users between these two groups is substantial).

These results show that certain sets of bloggers have a disproportianate number of followers. Certain traits, like posting frequently, selectively, and quickly, are correlated with success in our dataset. Investigating whether these relationships are causal is an interesting direction for future work.



(a) Mean # followers (TW Bin Laden)

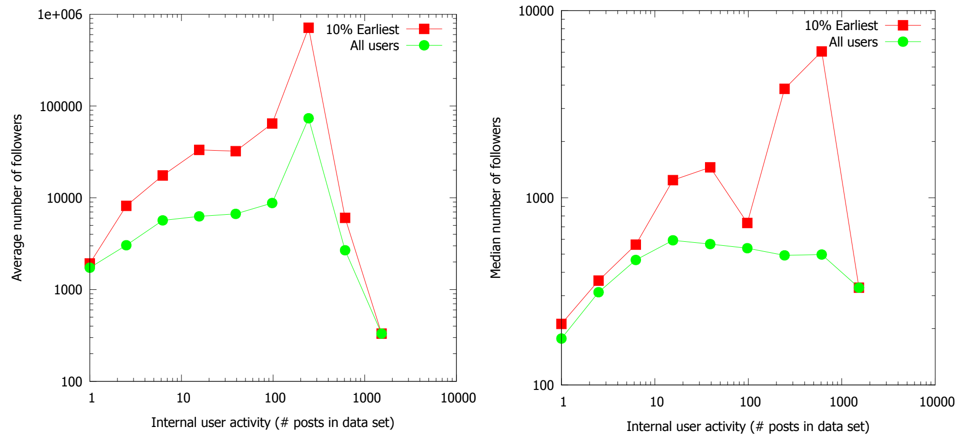
(b) Median # followers (TW Bin Laden)

Fig. 6. More selective users are generally more successful. (x-axis) user internal activity in number of URLs, (y-axis) mean/median number of followers of each group of users, bucketed by internal activity. Each line in these plots represents a different group of users with a certain minimum ‘selectivity percentile’.

4. INFORMATION FILTERING MODEL

In the previous section, we observed two main trends in our datasets:

(1) Less active users generally post more popular content on social media.
(2) Across activity levels, users who on average post more popular content within that activity level tend to have more followers.



(a) Mean # followers (TW nyti.ms) (b) Median # followers (TW nyti.ms)

Fig. 7. Users who post articles earlier are generally more successful. (x-axis) user internal activity in number of URLs, (y-axis) mean/median number of followers of each group of users, bucketed by internal activity. The top line represents the 10% earliest users, and the bottom line represents all users.

Here, we focus on analyzing a simple theoretical model based on this simple two observations that capture the dynamics of the blog-sphere and on studying its efficiency.

In particular, we describe a model to study content consumption via blogs, and in the following section, we will consider its theoretical consequences. We believe this model to be a natural choice given the empirical observations mentioned above. Additionally, we believe it is important for the model to capture the competition for user attention which occurs between content intermediaries, given the following observations: 1) There is limited time within a day which users give to consuming online content, 2) Not all users follow all content intermediaries, and thus can be thought to have a limited budget of "follower relationships" which they can establish and sustain, and 3) As we saw in the previous section, there are certain intermediary behaviors which are associated with larger follower counts, indicating that a user's choice of who to follow depends (at least in part) on the content being shared by each intermediary.

We assume that every item of content lives in an arbitrary multi-dimensional space Ω. Every day d, news items S_d are generated from Ω following an arbitrary, possibly probabilistic, but time-invariant process. For instance, a special case of this model is that all news pertains to the same topic, where each news article within that topic has a certain uni-dimensional importance/quality which is drawn according to some distribution; the full model is considerably more general.

Blog Filtering: How does a blogger decide which items to post? We assume that each blogger b corresponds to a filtering function $I_b\colon \Omega \rightarrow \{0,1\}$, that indicates for each possible point in the space, whether or not the blogger will post an item corresponding to this point. For news item s corresponding to point $\omega(s) \in \Omega$, we use $I_b(s)$ to denote $I_b(\omega(s))$. We call the filtering function I_b the posting strategy for blogger b. The strategic choice for each blogger is what posting strategy she will use in order to maximize her objective of amassing a large number of users. A simple class of blogger filtering functions, corresponding once again to the single-topic scenario mentioned above, is that a blogger only posts news articles above a certain quality threshold within a single topic.

As we discussed below, our game can be defined in general, but we are primarily interested in *pure strategies*. In other words, we model blogs that apply a deterministic and invariant filter to content items that are generated randomly everyday. Mixed strategy are more general, but seems impractical: it is not reasonable to think of a blogger applying a filtering function to select all political news items on one day, the next day, picking all the sports news, and on the day after, picking science and fashion. Similarly, knowing that her audience – or the audience this blogger wants to attract – has a deterministic taste, why would she decides randomly not to include a given item that is in their interest? Notice, however, that even with this assumption the precise articles a blogger posts on a given day are still random, as a result of the random process whereby news content is generated.

User Utility: How much utility does a user u get from following blogger b with filtering function I_b? In its most general form, we can assume that a user's utility is an arbitrary real-valued function U of the filtering function I_b, and the identity of the blogger b. However, we will discuss a more concrete version of a user's utility below.

We assume that each user u has an interest function $i_u \colon \Omega \to \mathbb{R}$ that assigns a real number in $[-\infty, \infty]$ to every item, according to its position in Ω, where this number corresponds to how much the user would enjoy seeing this item. For news item s corresponding to a point $\omega(s) \in \Omega$, we use $i_u(s)$ to denote $i_u(\omega(s))$. This function i_u could be arbitrary: for example, i_u could be $+1$ for items that a user is interested in, and -1 otherwise; the function could also be more complex, with a score quantifying exactly how interested the user is in the item. A simple class of user interest functions, corresponding to the scenario mentioned above in which all news pertains to the same topic, but has varying quality, is that each user has a quality threshold and gets positive utility from consuming content above that threshold, and negative utility for content below their threshold.

Now, we can ask what utility a user u gets from consuming a set S of content. We define this utility $\sum_{s\in S} i_u(s)$ as the total interest of the user in this set (or it could be a function of this sum, with diminishing returns). With this definition, how much utility would user u derive from following blogger b on day d? Recall that S_d denotes the set of news items of that day; let $I_b(S_d) = \{s \in S_d : I_b(s) = 1\}$. Since the blogger posts $I_b(S_d)$, the utility of u from following b will be $\sum_{s\in I_b(S_d)} i_u(s)$. This utility might vary from day to day, in which case, one could define the utility of u from following b to be the expected value (over the randomness of the daily news S_d) of the daily utility. In general, one can interpret the utility of u from following b as a measure of how well b's strategy I_b matches u's interest function i_u. For our model, all that matters is that one can define a measure of the utility obtained by u from following b when b chooses the posting strategy I_b. We also assume that a user has always the option to not follow any blogger and thus to get utility 0.

Note that though we have defined the blogger's incentives, and how readers obtain utility from following blogs, we have not described how exactly readers choose which blog to follow. Does user u simply pick one blog b that gives her the highest utility ? Are the dynamics more complex? Interestingly, in Section 5, we show that the outcomes of this game (equilibria, social welfare, etc.) are strongly affected by subtle differences in the precise way a user picks which blog to follow.

5. STABILITY OF BLOG COMPETITION

As mentioned above, so far we did not specify how the users select which bloggers to follow? We now show that answering this question has unexpected important consequences on the nature of equilibrium of the game.

We focus on two properties: On the one hand, ideally we wish to prove that any equilibrium in which this game stabilizes is *efficient*, where efficiency is defined in relation to the satisfaction of the users. This is not in general the most difficult property to obtain since we can usually write this game as a *valid utility system* as defined in [Vetta 2002]. This allows us to obtain that *any* Nash equilibrium is within a factor 2 of efficiency of a social optimum (see below for an exact definition), and we show that this result indeed is tight and cannot be improved.

On the other hand, this method has one serious limitation. Although an equilibrium where bloggers follow mixed strategies exists, as proved by a classical result, there is no guarantee that one exists where bloggers all follow a deterministic filtering function applied to the content. In fact, we start by showing an example where *no* such pure strategy Nash Equilibrium exists. In other words, even once content gets created according to a distribution, some bloggers in that case are *required* to randomly filter some items, or it means that a blogger can change strategy and benefits from it. It is hence possible that some users, whatever be the choices of blogs they follow, will always miss items of their interest. In contrast, in a pure strategy, a set of deterministic filters exist that achieve a robust outcome of the game, and the users can decide based on the set of news that they inherently wish to receive.

Our main result is to show that, under a very general condition on the behavior of users and bloggers, this limitation can be overcome. In fact, we prove more: we show that our model in this case lies at the intersection of potential games and valid utility system. We note that a pure strategy deriving from a potential is also a more satisfactory solution concept from the modeling standpoint. It simplifies a blogger's strategy and, in contrast to general results, is a first step towards obtaining an equilibrium in a constructive way.

5.1. Greedy user model

We first prove that a simple greedy user selection model is not sufficient to obtain the properties above. Following a simple intuition, if a user had complete knowledge of the system, among the bloggers who give her positive utility, she would pick the one that maximizes it (breaking ties randomly).

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 even simple cases where only two bloggers are present may be pathological.

PROPOSITION 5.1. *Suppose that the generative process of news creation creates one item daily in each of* 3 *points of* Ω*:* p_1, p_2 *and* p_3*. Furthermore, assume that there are* 30 *users who are interested in all the news and give them score* $+1$*. There are* 20 *users who are interested in news items corresponding to points* p_2 *and* p_3*, and give them score* $+1$ *and score* -1 *to news at point* p_1*. Finally,* 16 *users are interested only in news items at point* p_3 *and give them score* $+1$ *and score* -1 *to all other items. Then with the* greedy user model *, there is no pure strategy equilibrium in this scenario even if there are only two bloggers.*

See the appendix for a detailed proof.

We note that though this proposition applies to a specific news generation process and scoring function, it is in fact very easy to come up with many such examples, particularly in continuous spaces. The greedy strategy is very unstable. Although we do not do it here, the game played by bloggers in this case can be written as a valid utility system, but it will still suffer from this pathology.

In retrospect, the greedy model, however simple, seems a bit too restrictive. In reality, there are search costs / frictions / a lack of perfect information that prevent users from finding this "optimal" blogger. Indeed, the blogosphere, though chaotic, seems to be a reasonably stable environment, with blogs following fairly consistent strategies. How is this consistent with the lack of equilibria? We posit that this is because users *satisfice* (that is, 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 to entirely change the expected behavior of the game.

5.2. Satisficing user model

In the *satisficing* user model, each user sets her own utility requirement (as high or low as she prefers), and either selects a blogger uniformly at random among those who give her utility above this value, or selects all the bloggers that give her utility above this value and then shares her attention between them.[4] Let $S_u(s)$ denote the set of bloggers who provide user u with at least the required utility under strategy vector s; when the strategy vector s is clear or irrelevant, we simply use S_u. If blogger $i \in S_u$, in expectation i receives a fraction $\frac{1}{|S_u|}$ of u's attention. In this new setting, we assume that bloggers want to maximize the attention that they receive from all the users.

THEOREM 5.2. *The blog-positioning game in the satisficing user model is an exact potential game: It has a pure strategy equilibrium which can be found via best-response dynamics.*

PROOF. To prove the statement, we will associate a potential function with the system, and show that this function is upper bounded and its value increases every time a blogger moves to optimize his value. Thus the bloggers cannot move forever and so a pure strategy equilibrium exists.

Consider the the following potential function where we assign a score $H_{|S_u|}$ to each user u, where $H_{|S_u|}$ is the harmonic sum computed on the size of the set S_u of bloggers giving u at least her required utility; we then sum the scores over all the users. Now we analyze what happens when a blogger b_i changes his strategy. Let $b_i(t)$ be the strategy of blogger i at time t and let $b_i(t+1)$ be the strategy of blogger i at time $t+1$. Recall that the strategies of all the other bloggers, b_{-i}, remain unchanged, so $b_{-i}(t) = b_{-i}(t+1)$. If b_i changes his posting strategy , this implies that she will get more user attention. In particular, she will receive $\sum_{u:b_i \in S_u(t+1)-S_u(t)} \frac{1}{|S_u(t+1)|} - \sum_{u:b_i \in S_u(t)-S_u(t+1)} \frac{1}{|S_u(t)|}$, where the first sum collects how much b_i gains from new users who follow him, and the second sum collects how much he loses from old users who no longer follow him. But each user counted in the first sum has her score increase by exactly her contribution to the sum, and each user counted in the second sum has her score decrease by this amount.

Thus the change in the overall potential function is exactly the change in blogger i's utility. But if i moved, this change is strictly positive, so the potential function increases in each time step, and there exists an $\epsilon > 0$ lower bounding this increase. The value of the potential function is upper bounded by $|U| \log B$, where U is the number of users and B is the number of bloggers, which completes the proof. □

We define the *efficiency* of the blogs' filtering system as the fraction or number of users who end up receiving at least a utility equal to their requirement. As mentionned above, we can expect to obtain a good bound on the efficiency of *any* equilibrium as the game can be rewritten as a valid utility system. To see why, note that this efficiency or social welfare coincides with the total user attention that the bloggers get in that model. We recall the concept of price of anarchy.

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

Definition 5.3. For a measure of efficiency defined for each outcome, or welfare function $W : \Sigma \rightarrow R^{+}$. Let us define $E \subset S$ to be the set of equilibrium strategies, then we can define the price of anarchy as

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

This concept captures the efficiency of a system in *any* set of strategy that ends up being robust to bloggers' selfish decisions. We prove that the system is indeed efficient. In fact, the proof ends up being simpler than in the general valid utility case, so we prove the result directly.

THEOREM 5.4. *The price of anarchy of the system in the satisficing user model is upper bounded by* 2*.*

PROOF. Let s^* be the strategy in Σ that maximize $W(s^*)$ and let $s \in E$ be the strategy in E that minimize $W(s)$. Recall that the welfare function coincides also with the number of users who have at least one blogger giving them utility over their threshold. Thus this implies that the number of users who are following at least one blogger in s^* is greater than or equal to the number of users that are following at least one blogger in s. We will now show that the difference is at most a multiplicative factor of 2.

For each user u that follows at least one blogger in s^* and no blogger in s^E, select one blogger who is in $S_u(s^*)|$ and add it to the set B'. For each blogger in B' let $I^*_{b_1}, \cdots I^*_{b_r}$ be their posting strategies in s^* and $I_{b_1}, \cdots I_{b_r}$ be the set of posting strategies in s^E. Let $U^*(b_i)$ denote the set of all users who placed b_i in B'. Let $g(I_{b_i}, I_{b_{-i}})$ be the gain that blogger i gets when he uses strategy I_{b_i} and all the other players use the strategies described in $I_{b_{-i}}$. By definition of equilibrium, we have that for all $1 \leq i \leq r$, $g(I_{b_i}, I_{b_{-i}}) \geq g(I^*_{b_i}, I_{b_{-i}})$. Furthermore $g(I^*_{b_i}, I_{b_{-i}})$ is at least $|U^*(b_i)|$ because by moving, i would cover all the users in this set. Thus it is also true that $W(s) \geq \sum_{i=1}^{r} g(I_{b_i}, I_{b_{-i}}) \geq \sum_{i=1}^{r} g(I^*_{b_i}, I_{b_{-i}}) \geq |\bigcup_{i=1}^{r} U^*(b_i)|$. But by definition of $U^*(b_i)$, we have that $|\bigcup_{i=1}^{r} U^*(b_i)| = W(s^*) - W(s)$, thus $W(s^*) \leq 2W(s)$ □

Next we prove that the theorem above is essentially tight.

PROPOSITION 5.5. *The price of anarchy of the system in the satisficing user model 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 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. □

In this new model, we have proved that we have both equilibrium and a good price of anarchy, but it is still missing important aspects of reality. In particular, note that it implies that the maximum relative difference in the share of attention received by two bloggers is 2, but in reality, some bloggers have considerably larger audiences than others, as they may be faster or present a particular news item in a more compelling manner. To solve this problem, in the next section we introduce the concept of *ability* of a blogger. We show that, although this model is more realistic and allows one to account for many differences among blogs, all the results above generalize, and a robust set of deterministic filter always emerge, and satisfies at least half the fraction of users that any system creates.

5.3. Model with blogger abilities

Our last modeling ingredient would be to add the concept of the 'ability' of a blogger in the system. The notion of ability of a blogger captures various attributes of a blogger and his blog: the quality of his writing, his fame, the time since the blog was created / first-mover advantage, etc.

We model bloggers abilites 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 (by posting a set of items that is close to what she wants), 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 that the user shares her attention equally among all bloggers who provide her sufficient utility, and for whom the coin comes up heads. (Equivalently, she may pick uniformly at random from one of the bloggers for whom the coin comes up heads.)

Having defined the model, we move to our theoretical results. In this more general model, there is a pure-strategy equilibrium, and the price of anarchy of the model is upper bounded by 2. Furthermore, note that there is also a lower bound of 2, following the result of the previous subsection, as that setting was a special case (when $a_b = 1$ for all bloggers) of this more general case. Our proof of existence of an equilibrium is similar to the proof of existence of an equilibrium in [Kleinberg and Oren 2011]. In particular, we use the following theorem (modifying terminology to our setting) adapted from [Monderer and Shapley 1996].

THEOREM 5.6. *A game is an (exact) potential game if for every two bloggers i, j, posting strategies $I_i \neq I_j$, I'_i, I'_j and strategy vector $s_{-i,j}$:*

$$u_i(I'_i, I_j, s_{-i,j}) - u_i(I_i, I_j, s_{-i,j}) + u_j(I'_i, I'_j, s_{-i,j}) - u_j(I'_i, I_j, s_{-i,j})$$
$$+u_i(I_i, I'_j, s_{-i,j}) - u_i(I'_i, I'_j, s_{-i,j}) + u_j(I_i, I_j, s_{-i,j}) - u_j(I_i, I'_j, s_{-i,j}) = 0$$

where $u_i(*)$, is the utility of blogger i under a specific strategy vector and $s_{-i,j}$ is a vector containing the strategies of all bloggers except i and j.

Now we can prove our equilibrium result:

THEOREM 5.7. *The game in the satisficing user with bloggers' quality model is an exact potential game and so it has a pure strategy equilibrium.*

PROOF. To prove the statement we use theorem 5.6. First, let compute the utility of a blogger i when she select a posting strategy I_i; we abuse notation and write $I_i \in S_u$ if blogger i is in S_u when she plays I_i:

$$u_i(I_i, I_j, s_{-i,j}) = \sum_{u: I_i \in S_u, I_j \notin S_u} a_i \sum_l \frac{1}{l+1} \Pr(F(u)_{-i,j} = l) +$$
$$\sum_{u: I_i \in S_u, I_j \in S_u} a_i \overline{a_j} \sum_l \frac{1}{l+1} \Pr(F(u)_{-i,j} = l) + \sum_{u: I_i \in S_u, I_j \in S_u} a_i a_j \sum_l \frac{1}{l+2} \Pr(F(u)_{-i,j} = l)$$

Where $F(u)_{-i,j}$ is the number of bloggers followed by u not including i, j, and $\overline{a_j} = 1 - a_j$. Subsequently, we use U_i to denote $\{u \colon I_i \in S_u\}$, the users to whom a blogger would provide their required utility if she picked a posting set equal to I_i. U_j is defined similarly. We also use the following notation:

$$k(u) = \sum_l \frac{1}{l+1} \Pr(F(u)_{-i,j} = l) \qquad\qquad k'(u) = \sum_l \frac{1}{l+2} \Pr(F(u)_{-i,j} = l)$$

Now, we can rewrite the utility of a blogger who picks q_i as:

$$a_i \sum_{u \in U_i - U_j} k(u) + a_i \overline{a_j} \sum_{u \in U_i \cap U_j} k(u) + a_i a_j \sum_{u \in U_i \cap U_j} k'(u)$$

Now we can write the formula of Theorem 5.6:

$$\begin{aligned}
&u_i(I'_i, I_j, s_{-i,j}) - u_i(I_i, I_j, s_{-i,j}) + u_j(I'_i, I'_j, s_{-i,j}) - u_j(I'_i, I_j, s_{-i,j}) + \\
&u_i(I_i, I'_j, s_{-i,j}) - u_i(I'_i, I'_j, s_{-i,j}) + u_j(I_i, I_j, s_{-i,j}) - u_j(I_i, I'_j, s_{-i,j}) = \\
&\quad a_i \sum_{u \in U'_i - U_j} k(u) + a_i \overline{a_j} \sum_{u \in U'_i \cap U_j} k(u) + a_i a_j \sum_{u \in U'_i \cap U_j} k'(u) - \\
&\quad (a_i \sum_{u \in U_i - U_j} k(u) + a_i \overline{a_j} \sum_{u \in U_i \cap U_j} k(u) + a_i a_j \sum_{u \in U_i \cap U_j} k'(u)) + \\
&\quad a_j \sum_{u \in U'_j - U'_i} k(u) + a_j \overline{a_i} \sum_{u \in U'_i \cap U'_j} k(u) + a_j a_i \sum_{u \in U'_i \cap U'_j} k'(u) - \\
&\quad (a_j \sum_{u \in U_j - U'_i} k(u) + a_j \overline{a_i} \sum_{u \in U'_i \cap U_j} k(u) + a_j a_i \sum_{u \in U'_i \cap U_j} k'(u)) + \\
&\quad a_i \sum_{u \in U_i - U'_j} k(u) + a_i \overline{a_j} \sum_{u \in U_i \cap U'_j} k(u) + a_i a_j \sum_{u \in U_i \cap U'_j} k'(u) - \\
&\quad (a_i \sum_{u \in U'_i - U'_j} k(u) + a_i \overline{a_j} \sum_{u \in U'_i \cap U'_j} k(u) + a_i a_j \sum_{u \in U'_i \cap U'_j} k'(u)) + \\
&\quad a_j \sum_{u \in U_j - U_i} k(u) + a_j \overline{a_i} \sum_{u \in U_i \cap U_j} k(u) + a_j a_i \sum_{u \in U_i \cap U_j} k'(u) - \\
&\quad (a_j \sum_{u \in U'_j - U_i} k(u) + a_j \overline{a_i} \sum_{u \in U_i \cap U'_j} k(u) + a_j a_i \sum_{u \in U_i \cap U'_j} k'(u))
\end{aligned}$$

with one line for each of the 8 terms of the formula. First, consider all the terms containing U_i and U_j (Lines 2 and 7). Substituting $\overline{a}_i = (1 - a_i)$ and $\overline{a}_j = (1 - a_j)$, and cancelling terms, these two lines yield:

$$-a_i \sum_{u \in U_i} k(u) + a_j \sum_{u \in U_j} k(u)$$

From lines 5 and 8 (terms containing U_i and U'_j), we get:

$$a_i \sum_{u \in U_i} k(u) - a_j \sum_{u \in U'_j} k(u).$$

Similarly, from lines 1 and 4, and lines 3 and 6, we get:

$$a_i \sum_{u \in U'_i} k(u) - a_j \sum_{u \in U_j} k(u) + a_j \sum_{u \in U_j} k(u) - a_i \sum_{u \in U'_i} k(u).$$

Adding these all together, we obtain a sum of 0, completing the proof. $\square$

Having proved that the game has a pure strategy equilibrium, we show that it has also bounded price of anarchy.

THEOREM 5.8. *The game in the model of satisficing users with bloggers' abilities has price of anarchy at most 2.*

The proof of this theorem is similar in spirit to the proof that we use to upper bound the price of anarchy in the simple satisfying game, however the last steps here are more mathematically involved; please see the appendix for details.

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, and it generally impacts the success of information intermediaries. 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 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 [Goel et al. 2012]. 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*. Each of these topics – 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.

REFERENCES

AHMED, A., HO, Q., EISENSTEIN, J., XING, E., SMOLA, A. J., AND TEO, C. H. 2011. Unified analysis of streaming news. In *WWW '11: Proceedings of the 20th international conference on World wide web*.

AN, J., CHA, M., GUMMADI, K., AND CROWCROFT, J. 2011. Media landscape in Twitter: A world of new conventions and political diversity. *Proceedings of 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

ATHEY, S., CALVANO, E., AND GANS, J. S. 2012. The Impact of the Internet on Advertising Markets for News Media. *SSRN Electronic Journal*.

ATHEY, S. AND MOBIUS, M. 2012. The impact of news aggregators on internet news consumption: The case of localization. *working paper*.

BACKSTROM, L., KLEINBERG, J. M., AND KUMAR, R. 2009. 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*.

BENEVENUTO, F., CHA, M., GUMMADI, K., AND ALMEIDA, V. 2011. On word-of-mouth based discovery of the web. *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*.

BURTON, K., KASCH, N., AND SOBOROFF, I. 2011. The ICWSM 2011 spinn3r dataset. *Proceedings of the Fifth Annual Conference on Weblogs and Social Media*.

CHA, M., BENEVENUTO, F., HADDADI, H., AND GUMMADI, K. 2012. The World of Connections and Information Flow in Twitter. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*.

DAS, A. S., DATAR, M., GARG, A., AND RAJARAM, S. 2007. Google news personalization: scalable online collaborative filtering. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*.

GHOSH, A. AND MCAFEE, P. 2011. Incentivizing high-quality user-generated content. In *WWW '11: Proceedings of the 20th international conference on World wide web*.

GOEL, S., WATTS, D. J., AND GOLDSTEIN, D. G. 2012. The structure of online diffusion networks. In *Proceedings of ACM EC*.

GUPTE, M., HAJIAGHAYI, M., HAN, L., IFTODE, L., SHANKAR, P., AND URSU, R. 2009. News posting by strategic users in a social network. *Proceedings of WINE*.

JORDAN, P. R., NADAV, U., PUNERA, K., SKRZYPACZ, A., AND VARGHESE, G. 2012. Lattice games and the economics of aggregators. *WWW '12: Proceedings of the 21st international conference on World Wide Web*.

KATZ, E. 1957. The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis. *Public Opinion Quarterly*.

KLEINBERG, J. M. AND OREN, S. 2011. Mechanisms for (mis)allocating scientific credit. In *Proceedings of ACM STOC*.

KRAKOVSKY, M. 2013. Just the facts. *Communications of the ACM*.

KWAK, H., LEE, C., PARK, H., AND MOON, S. 2010. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*.

LESKOVEC, J., BACKSTROM, L., AND KLEINBERG, J. M. 2009. 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*.

MONDERER, D. AND SHAPLEY, L. S. 1996. Potential games. *Games and Economic Behavior*.

SHAHAF, D. AND GUESTRIN, C. 2012. Connecting Two (or Less) Dots: Discovering Structure in News Articles. *ACM Transactions on Knowledge Discovery from Data (TKDD)*.

VETTA, A. 2002. Nash equilibria in competitive societies, with applications to facility location, traffic routing and auctions. In *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*.

WALDMAN, S. 2011. The Information Need of Communities. *Report to the Federal Communications Commission (FCC)*.

WU, F. AND HUBERMAN, B. A. 2008. Popularity, novelty and attention. *EC '08: Proceedings of the 9th ACM conference on Electronic commerce*.

WU, S., HOFMAN, J. M., MASON, W. A., AND WATTS, D. J. 2011. Who says what to whom on twitter. *WWW '11: Proceedings of the 20th international conference on World wide web*.

YANG, J. AND LESKOVEC, J. 2011. Patterns of temporal variation in online media. In *WSDM '11: Proceedings of the fourth ACM international conference on Web search and data mining*.

A. OMITTED PROOFS

Proof of Proposition 5.1: First note that we can assume without loss of generality that the posting strategy of the bloggers is to always post all the news, or post news in p_2 and p_3, or post news only in p_3. In fact, other strategies such as posting p_1, p_2, or only p_1 and p_2 are dominated.

Let now call the two bloggers A and B, and without loss of generality, we assume that blogger B posts at most as much news than blogger A in equilibrium. In this case there are only 6 feasible configurations. A simple analysis of each case shows that none of them is an equilibrium. We present one case here, and omit the straightforward calculations for the others.

Suppose A and B are posting all 3 news items every day. In this setting all users interested in all the news or interested in news at points p_2 and p_3 follow A or B with equal probability, while users interested only in news at point p_3 follow neither because they would get negative utility. Thus the expected number of followers for A and B is 25. But now if B posts only news at p_3, all the people interested in only p_3 would follow her, plus (in expectation) half of the people interested in news at p_2 and p_3 would follow her (because they would get utility 1 both from A and from B).Thus, her expected number of followers would be $26 > 25$, and so this configuration is not an equilibrium. □

Proof of Theorem 5.8: This proof is similar in spirit to the proof that we use to upper bound the price of anarchy in the simple satisfying game, however the last steps here are more mathematically involved. As in the previous section let s^* be the strategy in Σ that maximizes $W(s^*)$ and let $s \in E$ be the strategy in E that minimizes $W(s)$. Note that in this new setting the social welfare is equal to the sum over all users u, of the probability that u follows at least one blogger.

Let us consider the set of users Y who have higher probability of following at least one blogger in s^* than in s. For each user $u \in Y$, select the bloggers who provide her with utility above her requirement in s^*, but not in s and add them to the set B'_u. Let the posting strategies of the bloggers in $B' = \bigcup_u B'_u$ be equal to $I^*_{b_1}, \cdots I^*_{b_r}$ in s^* and $I_{b_1}, \cdots I_{b_r}$ in s. Note that by definition we have $(v_Y(s^*) - v_Y(s)) \geq W(s^*) - W(s)$, where $v_Y(s)$ is the utility of bloggers restricted to the users in Y when they play strategy s.

Now by definition of equilibrium we have that $u(b_i, b_{-i}) \geq u(b^*_i, b_{-i})$ for all $b_i \in B'$. Now consider any user $u \in Y$; recall that $v_u(s)$ is equivalent to the probability that u follows at least one blogger under strategy profile s, and that a_i denotes the ability of blogger i. We have that:

$$\begin{aligned} v_u(s) + \sum_{i \in B'_u} v_u(b^*_i, b_{-i}) &= v_u(s) + \sum_{i \in B'_u} (1 - v_u(s)) a_i = v_u(s) + (1 - v_u(s)) \sum_{i \in B'_u} a_i \\ &\geq v_u(s) + (1 - v_u(s)) \times \left(1 - \prod_{i \in B'_u} (1 - a_i)\right) \geq v_u(s^*) \end{aligned}$$

where the second-last inequality follows from the fact that for any set X of real numbers between 0 and 1, $\sum_{i \in X} i \geq 1 - \prod_{i \in X} i$ (this can be checked by induction, for example) and the final inequality follows from the fact that the utility of y is equal to the probability that y follows at least one blogger. But now, summing over all users, and using the fact that for each blogger, $u_i(b^*_i, b_{-i}) \leq u_i(b_i, b_{-i})$, we get $2W(s) \geq W(s^*)$. □