Anatomy of News Popularity on Social Media*

"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.

Avner May
avnermay@cs.columbia.edu,

Augustin Chaintreau
augustin@cs.columbia.edu,

Nitish Korula
nitish@google.com

Silvio Lattanzi
silviol@google.com

ABSTRACT

"Is this news?" is a critical and difficult question which had always been answered partly by domain experts and partly by a more informal network of opinion leaders. Recently, the rise of social sharing and information diffusion through blogs, micro-blogs, and social networking, opens this curation process for everyone's participation. It is well known that this apparent level playing field is characterized by sharp contrasts: An active minority of information intermediaries generate most traffic and gather most of the followers, while a minority of items receive most of the attention. But what remains unknown is how these two concentration results relate to each other, and how they may interact to offer the audience a layered offering of news with various level of depths.

Here, and for the first time, we study *jointly* the volume and popularity of URLs received and shared by users. We show that users and bloggers 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 data sets of different media, topics, and domains and various measures of URL popularity, and it leads us to formulate various hypothesis on the nature of information filtering social media permit.

1. INTRODUCTION

Traditional media have historically acknowledged that information reach their audience through a two-step information flow [3]. Opinion leaders who 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 [4, 5], and more recently identifying mass media and intermediaries as critical to information spread [2]. What has changed is the scale of the set of people that can potentially act as intermediaries, since this process is in theory open to any Internet users.

It leads almost all traditional media to embrace users playing the role of referrers and intermediaries. Buttons to "like" or "share" multiply on the web. Mainstream media that are typically reluctant to open all their content free of charge (*e.g.,* the New York Times) make notable exceptions for users accessing it through a social referral. While domain experts and professional journalists are still trusted as sources, the answer to *what* constitutes noteworthy information has become the result of a collaborative process. This distributed curation allows Internet users in theory to be informed from a broader set of news (potentially less biased, as argued for instance in [1]). It also offers a chance for users to see a richer set of news offering: Spontaneous intermediaries may quickly emerge, especially in reaction to major events, or following the intricate web of a hierarchy of niche tastes.

Various claims have been made on the power of social media using the *volume* and *speed* of information diffusion online. They are also showing that in a network where anyone can post and any information can potentially propagate, popularity is highly skewed, with only a few users and URLs creating and receiving most traffic. Here we wish to go beyond these observations to see how they interact together. How do today's information intermediaries selectively post content? How does it relate to content's popularity? And briefly, what consequences can be foreseen to address the need of information of all communities.

Our work makes the following primary contribution:

- Based on 11 data sets characterizing users from different social media platforms with various topics, time periods, and information domains, we demonstrate that social curation obeys a "filtering law". This law predicts that the volume of content received by a user, or posted by an intermediary, are both inversely correlated with the content's popularity. This relationship is robust to different definitions used to identify information intermediaries, and to measure popularity of content.

- This poses a natural question which is what behaviors may explain this filtering law, especially as it seems to be so prevalent. While any causal claims is difficult to validate, some potential candidates may seem natural: It could be a consequence that intermediaries "run out" of popular URLs as they get more active. It could be that the less active simply receive less, and those are particularly popular. Through a statistical comparison with such null hypothesis model, we prove that it cannot be explained this way.

A recent analysis of Twitter suggest that social media can be efficient at giving users information relevant to their interests [6]. Among other theoretical results, this paper

*A preliminary evidence of the "filtering law" was presented during the second SC-UGC workshop at the ACM EC '13 conference, inside a theoretical paper focusing on blogs' incentive for curation. The results shown in this paper have been obtained with new and more complete data, and never appeared in a previous publication.

shows that it is necessary that users select from their reading interests a subset that they post, although this argument only applies for undirected graph. Interestingly it shows that Twitter users behavior align with this need. In contrast, we focus on the behaviors of intermediaries in *directed* graphs, and on the impact of selection on popularity.

2. EVIDENCE OF FILTERING

Prior work proved the existence of information intermediaries, while we explore for the first time how these intermediaries affect what part of the 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?

2.1 Data-sets

Collecting information shared on social media.

To answer the questions above, we gathered and analyzed several large traces from various media; we focus on tweets, FB shares, and blog posts containing URLs. We gathered a mix of *topical* traces (i.e. relating to an event) and *generic* ones (where any topic of current interest is included). Our data set is comprehensive; it contains both blockbuster URLs, as well as niche content, and posts from anyone that could be serving as an information intermediary.

Data sets	Users (% live today)	URLs (% 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 Twitter data sets, we report the fraction of users whose accounts is still active. We also report for all the fraction of URLs with a bit.ly shortener.

- **Microblogging** (TW). Data was gathered using `DiscoverText.com` to harvest tweets from Twitter unrestricted firehose in two ways. First, we collected during two weeks in Dec. 2012 all tweets with a URL from a particular media source (we report here results from the New York Times; we also gathered data from CNN and FoxNews). Second, we collected all URLs associated with a particular event (*e.g.,* TW Bin Laden contains tweets with the words "Osama" or "Bin Laden" posted during 34h 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 also use a set of 1m Facebook comments collected for one event (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. It contains 386m blog posts collected from Spinn3r during the months of January-February 2011.

Additional crawls.

Twitter Crawls: We performed Twitter crawls to gather additional information about the users in our data sets.[1] First, we gathered general information about these users via the Twitter REST API. 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 feasibly, we stopped the crawl after 50k friends.

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 these in each data set by counting how many posts of a given normalized URL or from a given user occur. We complement this by measuring *externally* how popular a given URL is in two ways. If this URL was ever shortened using bit.ly (20-40% of all URLs exchanged in our data sets), we collected its number of clicks via the bit.ly public REST API. We also collected the number of times each URL was "shared" on Facebook using the FB API.

2.2 The Filtering Law

Ideally, depending on how many and which intermediaries a user follows, she may be able to reduce the volume of information while statistically increasing its overall quality. We now show that this is indeed the case. URLs received by users who receive fewer URLs are disproportionately popular: see Fig. 1 (top), where the quality (here, estimated as the number of bit.ly clicks) of typical URLs received decreases as the total number of URLs a user 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 to 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

[1]We focus on accounts that are still live today. This is larger than 75% a year after, and 94% after a couple of months.

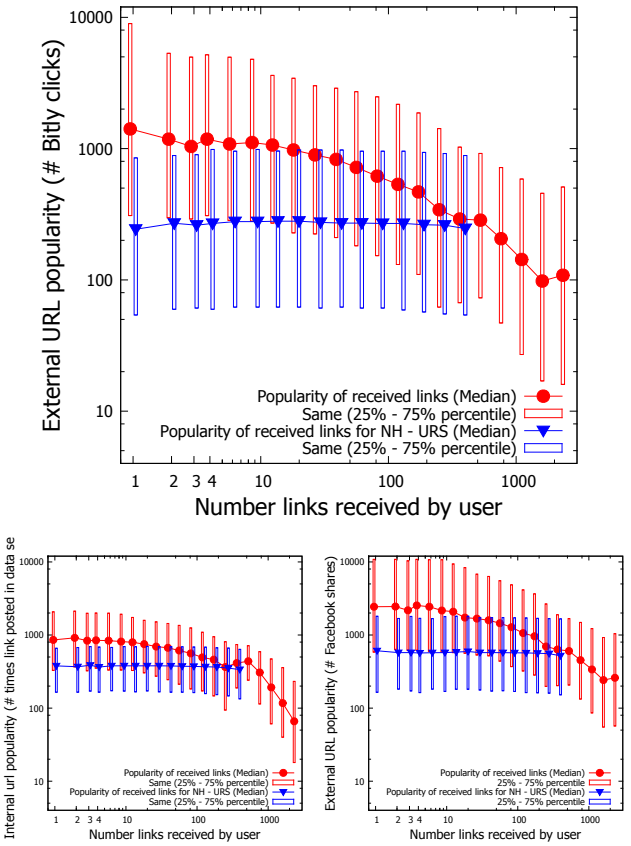

Figure 1: The filtering law. (x-axis) Users grouped by number of URLs received, (y-axis) Popularity distribution of URLs received by these users, shown using median and 25%-75% percentiles. Red circles denote real data, blue triangles denote a random null hypothesis model. The 3 plots correspond to different metrics of URL popularity: (top) number of bit.ly clicks, (bottom left) number of times this URL was posted, (bottom right) number of FB likes.

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, 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 universally found: we observe it for all popularity metrics as shown in Fig. 1 (bottom).

Our results are encouraging because they show that users of social media can navigate the volume/quality trade-off: users who receive less are effectively focusing on the most promising URLs.

These positive results demand additional explanation. 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? In the next section, we will examine the posting behaviors of information intermediaries, and their consequences.

2.3 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.

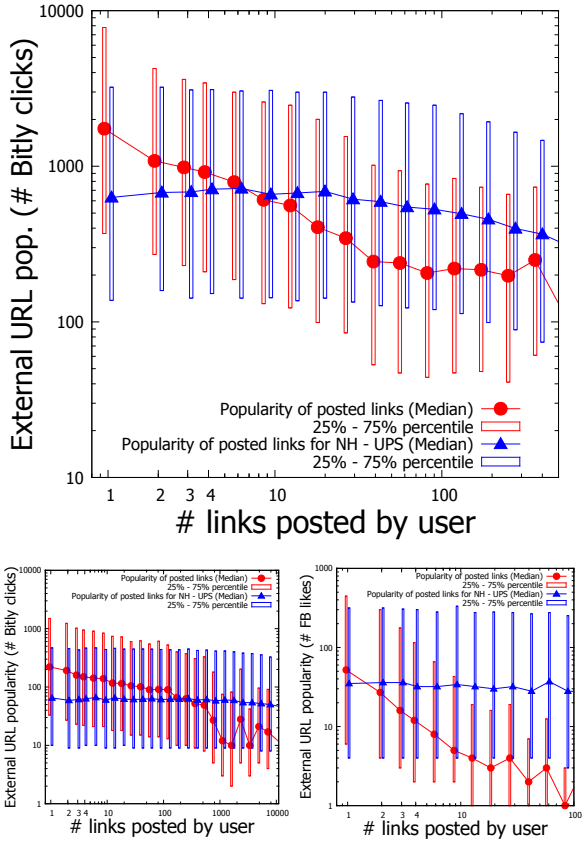

Figure 2: The posting filtering law. (top) [x-axis] Users grouped by number of URLs posted, [y-axis] popularity distribution (by number of bit.ly clicks) of URLs posted by these users shown using median and 25%-75% percentiles. Red circles denote real data, blue triangle denotes a random null hypothesis model. In the bottom 2 figures we see that the posting filtering law is observed across data sets: (bottom left) Blogs All Datasets, (bottom right) FB Iphone5 Data set (y-axis denotes FB likes)

That is, as intermediaries decrease the number of URLs they post, they are more likely to post the most popular URLs. Fig. 2 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, in fact, *filtering* when choosing what content to post (more on that below).

Our trend is remarkably universal. No matter what URL popularity metric is used, the trend is statistically significant. We present a few representative results in Fig. 2.

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 post a given fraction of the content they receive. Intermediaries who receive less, post less; those who follow them receive still less, etc. Figure 3 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. 3 (bottom) shows that the average number

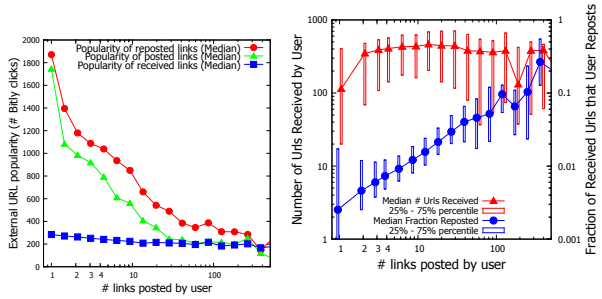

Figure 3: Selection of popular content by intermediaries: (x-axis) User posting activity in number of URLs, (top y-axis) popularity distribution of URLs received by these users, and the popularity distributions of URLs posted or reposted by the users. (bottom left y-axis) distribution of number of URLs received plotted as red triangle, (bottom right y-axis) distribution of reposting fraction plotted as blue circle. (Boxplots show 25%-75% percentiles).

of URLs received by intermediaries is more or less constant across activity levels. However, users 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 a different portion of it, usually composed of selected 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. 3).

In summary, we have observed two important phenomena about the posting behavior of intermediaries: First, they are selective about what they post, being more likely to post the URLs that eventually become popular. Secondly, intermediaries have different posting thresholds. Those who post less content post only a tiny fraction of what they receive, corresponding to the most popular URLs; this accounts for the huge gap between the blue line and the red and green lines on the far left of Fig. 3. Those who post more select a larger fraction of the content they receive to share, though they are still selectively posting more popular URLs.

3. DISCUSSION

The filtering law potentially brings notable consequences, many remain to be explored: On the one hand, it makes it easier for users to filter for popular URLs (by following less active intermediaries). But the concentration may also exacerbate the dominance of topics of broad appeal, leaving genuinely interesting topics that are secondary with even less space in the blogosphere.

The available body of evidence that we gather points to an active information selection made by intermediaries, which requires better understanding. For instance, in a previous paper we assume that this behavior is endogenous to the media itself: bloggers are implicitly incentivized to filter a subset of most popular items of various sizes, in order to adapt to varying level of interests in the audience. More generally, our findings call for a renewed study of the follower graph. Since users in the audience pick blogs selectively as well, a filtering effect should be present in this graph as well, which may relate or complement some of the effects we found, potentially connecting the two filtering laws we observe. Another important direction to consider is how activity and popularity are composed of different topics themselves. So far we ignored this dimension, especially in the datasets focusing on a single topic. Intermediaries may exhibit specific filtering behaviors with regard to those compositions.

4. REFERENCES

[1] J. An, M. Cha, K. Gummadi, and J. Crowcroft. Media landscape in Twitter: A world of new conventions and political diversity. *Proceedings of 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[2] 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*, 2012.

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

[4] 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.

[5] 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.

[6] 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.