On the Users' Efficiency in the Twitter Information Network

Mahmoudreza Babaei Przemyslaw A. Grabowicz Isabel Valera Manuel Gomez-Rodriguez
Max Planck Institute for Software Systems, Saarbrucken and Kaiserslautern, Germany
{babaei, pms, ivalera, manuelgr}@mpi-sws.org

Abstract

Social media systems have increasingly become digital information marketplaces, where users produce, consume and share information and ideas, often of public interest. In this context, social media users are their own curators of information – however, they can only select their information sources, who they follow, but cannot choose the information they are exposed to, which content they receive. A natural question is thus to assess how *efficient* are users at selecting their information sources. In this work, we model social media users as information processing systems whose goal is acquiring a set of (unique) pieces of information. We then define a computational framework, based on minimal set covers, that allows us both to evaluate every user's performance as information curators within the system. Our framework is general and applicable to any social media system where every user *follows* others within the system to receive the information they produce.

We leverage our framework to investigate the efficiency of Twitter users at acquiring information. We find that user's efficiency typically decreases with respect to the number of people she follows. A more efficient user tends to be less overloaded and, as a consequence, any particular piece of information lives longer in the top of her timeline, thus facilitating her to actually read the information. Finally, while most unique information a user receives could have been acquired through a few users, less popular information requires following many different users.

Introduction

In the last decades, the creation, distribution, acquisition and manipulation of information has been a significant economic, political and cultural activity (Castells 2011). In the information society, people have increasingly relied on the Web for finding information of public interest as well as keeping up to date with last breaking news. In other words, the Web has become people's "first reads," *i.e.*, their default source of information (Kohut and Remez 2008). A broad spectrum of websites have emerged to serve that purpose, ranging from online news media outlets such as the New York Times or CNN, specialized blogs and online magazines such as Engadget or Gizmodo, to digital encyclopedias such as Wikipedia. Importantly, in most of these sites, information is often curated by editorial boards, professional journalists, or renowned experts with a recognized reputation.

The advent of social media and social networking sites is changing dramatically the way in which people acquire and consume information. Social media sites such as Twitter, Tumblr or Pinterest have become global platforms for public self-expression and conversation; more formally, we can think of these sites as large information networks where nodes (*i.e.*, users) both create and consume information (Kwak et al. 2010). In this context, nodes play the role of information curators by deciding which information to *post*, which other nodes in the network to *follow*, and which information posted by other nodes to *forward*. This entails serious challenges as well as raises many important questions, which have not been addressed until very recently. For example, users tend to follow too many other users, perhaps afraid of missing out (Hodas and Lerman 2012). As a consequence, they become overloaded and effectively miss the information they were interested in (Gomez-Rodriguez, Gummadi, and Schoelkopf 2014; Lerman and Hogg 2014).

The present work is motivated by a fundamental question that, perhaps surprisingly, is barely understood: How *efficient* are the users of a social media site at selecting which other users to follow to acquire information of their interest? In what follows, we will quantify users' efficiency at acquiring information.

The Present Work

In this paper, we define a intuitive computational framework that allows us to quantify Twitter users' efficiency at acquiring information. Our framework is based on the following key concept: given a set of unique ideas, pieces of information, or more generally, *contagions*, $\mathcal{I}$ spreading through an information network, we think of the minimal set of nodes $\mathcal{U}^*(\mathcal{I})$ that, if followed, would allow us to get to know $\mathcal{I}$, as the most compact representation of $\mathcal{I}$ in the information network. Finding $\mathcal{U}^*(\mathcal{I})$ reduces to a minimum set cover problem, which we can solve using an well-known (efficient) greedy algorithm with provable guarantees (Johnson 1973). We then leverage this idea to define the efficiency of a user, by comparing the number of people she follows, i.e., the number of *followees*, with the size of the minimal set of

users that, if followed, would provide the same unique contagions. High efficiency means that the number of followees is close to the size of the minimal set. Importantly, we can extend this idea to account for partial coverage or additional constraints on the minimal sets, enabling a deeper understanding of the subtleties of information acquisition in an information network.

Our analysis yields several insights that do not only reveal how efficient are users in social media at acquiring information, but also help us in understanding the influence different factors have on this efficiency:

1. We find that users' efficiency decreases with respect to the number of people they follow.
2. Users that are more efficient at choosing the people they follow are proportionally less overloaded. As a consequence, contagions live longer in the top of their timeline, facilitating its discovery.
3. There is a trade-off between information coverage and information efficiency. While most unique information a Twitter user receives could have been acquired through a few users, less popular information requires following many different users.

Most of our empirical findings shed light on how our intuitive notion of efficiency relates to different aspects of a user's timeline (*e.g.*, amount of information a user receives). However, an important remaining question, left as potential future work, is to investigate how users' efficiency relates to their level of activity and engagement within the online social media system.

Dataset Description

We use data gathered from Twitter as reported in previous work (Cha et al. 2010), which comprises the following three types of information: profiles of 52 million users, 1.9 billion directed follow links among these users, and 1.7 billion public tweets posted by the collected users. The follow link information is based on a snapshot taken at the time of data collection, in September 2009. In our work, we limit ourselves to tweets published during one week, from July 1, 2009 to July 7, 2009, and filter out users that did not tweet before July 1, in order to be able to consider the social graph to be *approximately* static. After the preprocessing steps, we have $395,093$ active users, $39,382,666$ directed edges, and $78,202,668$ tweets.

Then, we pick $10{,}000$ users at random out of the $395{,}093$ active users and reconstruct their timelines by collecting all tweets published by the people they follow (among all the $395{,}093$ users), build their ego networks by finding who follows whom among the people they follow, and track all the unique contagions they are exposed to. Following previous work, we consider two different types of contagions: hashtags[1] (Romero, Meeder, and Kleinberg 2011) and websites (Mislove et al. 2007). Our set of active users mention $286{,}219$ and $379{,}424$ unique hashtags and websites, respectively, during the week under study. As one may have expected, the distribution of unique number of mentions for both types of contagions follows a power-law distribution. Our methodology does not depend on the particular choice of contagion, however, it does make two key assumptions. First, it assumes we can distinguish whether two contagions are equal or differ. Distinguishing certain contagions such as hashtags may be trivial but distinguishing others, such as ideas, may be very difficult. Second, it assumes that receiving several copies of the same contagion from different users does not provide additional information, even if different users express different opinions about the contagion. Although it seems difficult, it would be very interesting to relax the second assumption in future work.

Finally, an important characteristic of Twitter in 2009 is that it did not have features such us "Lists" and "Personalized Suggestions" and so the primary way users received and processed information was through their feed, for which we have complete data. However, this comes at the cost of observing a smaller number of users and social activity.

Information Efficiency

In this section, we will first define an intuitive notion of *efficiency*, which can be efficiently approximated with theoretical guarantees. Then, we will compute how efficient Twitter users are, Finally, we will give empirical evidence that more efficient users are proportionally less *overloaded* and, as a consequence, information in their feeds has longer lifetime. We will conclude elaborating on the trade-off between information coverage and information efficiency.

Consider a Twitter user u and the set of unique contagions $\mathcal{I}_u$ (be it in the form of hashtags or websites) she is exposed to through her feed in a given time period, by following $\mathcal{U}_u$ users[2]. Now, we can think of the minimal set of users $\mathcal{U}^*(\mathcal{I}_u)$ that, if followed, would expose the user to, at least, $\mathcal{I}_u$, during the same time period as the most compact representation of $\mathcal{I}_u$ in Twitter. Then, we define the efficiency of a Twitter user u at acquiring a certain type of contagion as:

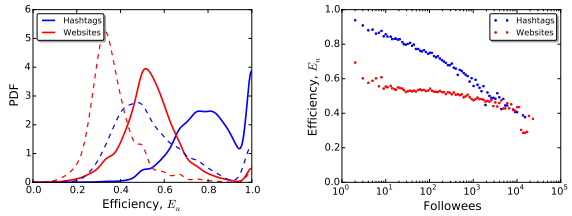
$$E_u = \frac{|\mathcal{U}^*(\mathcal{I}_u)|}{|\mathcal{U}_u|}, \tag{1}$$

where $0 \leq E_u \leq 1$. If the number of users she follows coincides with the number of users in the minimal set, her efficiency is $E_u = 1$, and the larger the relative difference between the size of the set of followees and the size of the minimal set, the smaller the efficiency. Importantly, it is straightforward to extend our definition of efficiency to account for partial coverage, by simply considering minimal set of users that, if followed, would expose the user to, at least, a percentage of the unique contagions $\mathcal{I}_u$.

Our definition captures two types of inefficiency, which we illustrate by two contrasting examples. If a user follows U other users, each of them mentioning different (disjoint)

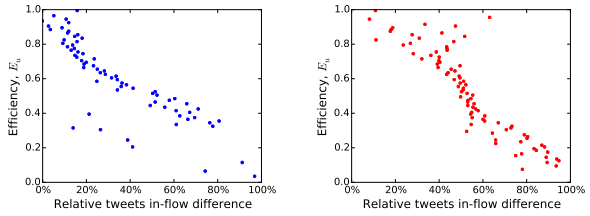
[1] Hashtags are words or phrases inside a tweet which are prefixed with the symbol # (Romero, Meeder, and Kleinberg 2011). They provide a way for a user to generate searchable metadata, keywords or tags, in order to describe her tweet, associate the tweet to a (trending) topic, or express an idea.

[2] We consider only users (followees) that mention a contagion at least once. Considering all followees leads to qualitatively similar results, but lowers absolute values of efficiency.



(a) PDF (b) Average efficiency vs number of followees

Figure 1: Efficiency. Panel (a) shows the empirical probability density function (pdf) of user's efficiency. Panel (b) shows the average efficiency against number of followees



(a) Hashtags (b) Websites

Figure 2: Average efficiency vs. tweet in-flow difference between the original timeline and the timeline induced by the minimal set cover.

set of contagions, and there is another user $v \notin U$ that cover all the contagions the followees cover, then the user's efficiency will be $E_u = 1/U$. If a user follows U other users and all these users mention exactly the same contagions, then the user's efficiency will be $E_u = 1/U$ and $\lim_{U\to\infty} E_u = 0$. The former type is due to following users that cover too few contagions while the latter is due to following redundant users.

In practice, computing E_u, as defined by Eq. 1, reduces to finding the minimal set of users $\mathcal{U}^*(\mathcal{I}_u)$, which can be cast as the classical minimum set cover problem (SCP) (Karp 1972). Although the minimum set cover problem is NP-hard, we can approximate $\mathcal{U}^*(\mathcal{I}_u)$ using a well-known (efficient) greedy algorithm (Johnson 1973), which returns a set cover of cost at most $H(d) \cdot \mathrm{OPT}$, where OPT is the minimum size of any set cover, $d = \max_{s\in\mathcal{S}} |s|$ is the maximum set size and $H(d) \approx 0.58 + \ln d$ is the d-th harmonic number.

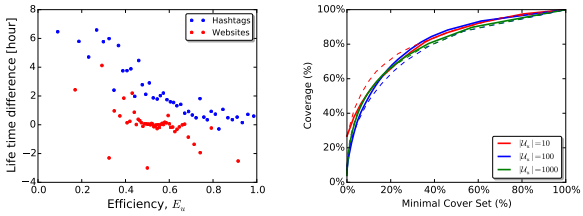
Once we have a meaningful definition of users' efficiency at selecting their information sources, we measure it for Twitter users. Solid lines in Figure 1 represent the empirical probability density function (pdf)[3] of user's efficiency for two types of contagions, as defined by Eq. 1. We find several interesting patterns. First, the pdf exhibits a clear peak around 0.55 for the websites and two clear peaks, around 0.80 and 1.0, for hashtags, with most of the density lying around these peaks. Second, approximately 64% and 6% of the users, respectively for hashtags and websites, yield efficiency above 0.75; or, in other words, users are much more efficient at acquiring hashtags than websites. A plausible explanation is that users often create new short-term hashtags to describe breaking news or ideas, which are used only once (Huang, Thornton, and Efthimiadis 2010). As a consequence, minimal set covers for hashtags may often be close in size to the original set of followees. In contrast, many users only mention well-known websites, such as newspapers or specialized blogs, who can be covered by few users. Now, we compute the empirical pdf's of user's efficiency for a 90% partial coverage for two types of contagions and plot them using dashed lines in Figure 1. The results suggest that covering the last 10% of the contagions requires significantly larger minimal set covers, especially for hashtags, as one may have expected given the short-term nature of some hashtags. However, both efficiency curves, for partial and complete coverage, are qualitative similar. In the remainder of the paper, we compute efficiency using full coverage but, remarkably, we found qualitatively similar results for a 90% partial coverage.

Next, we investigate users' efficiency against the number of people they follow. Figure 1(b) summarizes the results, where we plot the average efficiency against the number of followees. We find that the efficiency decreases with respect to the number of people they follow. This indicates that whenever a user decides to follow one more person, the new content this followee adds to the user's timeline diminishes with the number of people she follows, degrading her efficiency. Importantly, average efficiency is always larger than 0.3, independently of the number of people a user follows. In other words, for any given user, the number of people she follows is typically less than three times the size of the minimal set.

We have defined efficiency in terms of number of people a user follows. However, do the minimal set cover representations result in smaller tweet in-flow rates? If so, our measure of efficiency will be also useful to assess how efficient is a user in terms of in-flow. On the one hand, since tweet in-flow rates are strongly and linearly correlated with the number of followees, one may expect minimal set covers to result in smaller tweet in-flow rates. On the other hand, users in the minimal set covers may be among the most active ones, and thus produce significantly higher tweet in-flow rates than the average set of followees. Figure 2 gives empirical evidence that the minimal set covers do result in a smaller tweet in-flow rates than the original sets of followees, by showing the relative difference between the original in-flow and the in-flow induced by the minimal set cover. One potential explanation is that users in the minimal set cover *specialize* in one type of contagions but together do not actually produce larger tweet in-flow than the original sets. Importantly, the greater the efficiency, the smaller the difference in tweet in-flow between the minimal set covers and the original sets of followees.

If minimal set covers result in smaller tweet in-flow rates, one may expect contagions in the feed induced by the minimal set cover to have longer lifetime. Here, we define life-

[3]The pdf has been empirically estimated using kernel density estimation (Bowman and Azzalini 2004).



(a) Contagion lifetime difference vs efficiency

(b) Coverage percentage against the growing set cover

Figure 3: Panel (a) shows the average difference in contagion lifetime between the feed induced by the minimal set cover and the original feed, plotted against average efficiency. Panel (b) shows the coverage percentage against the growing set cover for three users with different number of followees ($|\mathcal{U}_u|$) for two types of contagions: hashtags (solid lines) and websites (dashed lines).

time as the time that a contagion appears in the top of their feed. In practice, we define lifetime of a contagion in a user's feed as the time that the contagion appears in the top-50 of the feed. However, results are qualitatively similar for other choices such as top-10 or top-20. Figure 3(a) confirms this intuition by showing that the contagion lifetime is longer in the feed induced by the minimal set cover than in the original feed. Here, we find that the more efficient a user is, the smaller the contagion lifetime difference between the feed induced by the minimal set cover and the original feed. A possible interpretation of this finding is that contagions live relatively longer in the feeds of efficient users and therefore are more easily discovered (Gomez-Rodriguez, Gummadi, and Schoelkopf 2014).

At the beginning of this section, we computed the distribution of user's efficiency considering both full and partial (90%) coverage. Now, we investigate further the tradeoff between partial coverage and the size of the minimal set cover. Our hypothesis here is that while most popular contagions may be acquired by following a few active users, acquiring rare contagions falling far into the tail of the distribution may require following proportionally more users. Figure 3(b) supports our hypothesis by showing the coverage percentage provided by a growing minimal set cover for three users with different number of followees ($|\mathcal{U}_u|$), chosen at random (we find a similar trend across all users). The first 25% of the users chosen by the greedy algorithm cover already 65% of the contagions, which are typically the most popular; in contrast, the last 25% of the users cover only 10% of the contagions, which are typically the rarest.

Discussion

We have introduced a framework that allows us to define an intuitive notion user's efficiency in social media, based on minimal set covers, to measure how *good* users are at selecting who to follow within the social media system to acquire non redundant information. Our framework is general and applicable to any social media system where every user *follows* others within the system to receive the information they produce.

Our work also opens interesting venues for future work. For example, we have defined and computed a measure of efficiency for each Twitter user independently. However, one could also think on global notions of efficiency for the Twitter information network as a whole, perhaps using a multi set cover approach. We have evaluated user's efficiency at acquiring two types of contagions (hashtags and websites). However, a systematic comparison of user's efficiency at acquiring many types of contagions appears as a interesting research direction. We have applied our framework to study information efficiency on Twitter. It would be interesting to study information efficiency in other microblogging services (Weibo, Pinterest, Tumblr) and social networking sites (Facebook, G+). Finally, it would be worth to investigate how users' efficiency relates to their levels of activity and engagement within the online social media system.

References

Bowman, A. W., and Azzalini, A. 2004. *Applied smoothing techniques for data analysis*. Clarendon Press.

Castells, M. 2011. *The rise of the network society: The information age: Economy, society, and culture*, volume 1. John Wiley & Sons.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, P. K. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 10–17.

Gomez-Rodriguez, M.; Gummadi, K.; and Schoelkopf, B. 2014. Quantifying Information Overload in Social Media and its Impact on Social Contagions. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 170–179.

Hodas, N., and Lerman, K. 2012. How visibility and divided attention constrain social contagion. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing*, 249–257.

Huang, J.; Thornton, K. M.; and Efthimiadis, E. N. 2010. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, 173–178.

Johnson, D. S. 1973. Approximation algorithms for combinatorial problems. In *Proceedings of the fifth annual ACM symposium on Theory of computing*, 38–49. ACM.

Karp, R. M. 1972. *Reducibility among combinatorial problems*. Springer.

Kohut, A., and Remez, M. 2008. Internet overtakes newspapers as news outlet. *Pew Research Centre*.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, 591–600.

Lerman, K., and Hogg, T. 2014. Leveraging position bias to improve peer recommendation. *PLoS One* 9(6).

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 29–42.

Romero, D. M.; Meeder, B.; and Kleinberg, J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, 695–704.