

Social Influence, Popularity and Interestingness of Online Contents

Hang M. Ung

Hewlett-Packard Laboratories, Palo Alto, CA
and Ecole Polytechnique, Palaiseau, France
hang.ung@polytechnique.edu

Abstract

From user-generated content platforms to shopping portals, an important feature of websites is to identify which contents or items are the most popular and list them on the front page – assuming these contents are also the most interesting. In this paper, we investigate how social influence effect on popularity can violate this assumption, by studying data from a major news aggregator website where stories are contributed by users. We first observe that the popularity of stories is correlated with the visibility of their authors, and can thus be explained by social influence. We then provide evidence of a feedback loop between the visibility of contributors and the popularity of their stories, which increases the importance of social influence over time. Finally we quantify the effect of social influence in making each story popular and show that this effect tends to favor less interesting stories.

1 Introduction

In a wide range of situations, people have a tendency to make choices based on popularity. In the absence of other information, most people would rather queue at a crowded restaurant than try an empty one. The attractiveness of popularity relies notably on the belief that popularity signals quality of product or interestingness of information.

Thus on the Internet, products and information are often ranked according to simple metrics such as the number of views, downloads or purchases. The objective is to help users find what is the most interesting or relevant, and relies on the assumption that popularity can effectively measure interestingness. In large and dynamic websites such as user-generated content platforms or e-commerce websites, identifying what is popular involves not only counting clicks, but also aggregating explicit votes and ratings from users (Lerman 2006). Thus, defining and identifying popularity is not a trivial problem, especially when new contents are added continuously. Moreover, that “hot stories” and “top products” will receive most of the attention of visitors because they are displayed on the front page, leads to some reinforcement effects and creates incentives for people to influence or game the ranking (Tran et al. 2009).

Copyright © 2011, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Another important aspect of attention allocation on the Web is the one of online social networks, which connect millions of Internet users and allow them to share what they are doing, reading or watching, on a real-time basis. Online social networks enables “word-of-mouth”-type of social influence and are known to be involved in e-commerce decision-making (Kim and Srivastava 2007), as well as in adoption of user-created contents (Cha et al. 2008; Bakshy, Karrer, and Adamic 2009).

In this paper, using data from the news aggregator website Digg (<http://www.digg.com>), we study the relation between popularity and interestingness of online contents. Specifically, we focus on characterizing how social influence, by impacting attention allocation, also affects popularity and thus creates a bias in popularity rankings with respect to the actual interestingness of contents. In this study, popularity is defined as the attribute of news stories which have entered the *popular stories* section of Digg, and is (potentially) prone to social influence. On the other hand, interestingness is a scalar and refers to the average interest that independent users would express for a story, in the absence social influence. Note that in the literature, *e.g.* in Szabo and Huberman (2010), this later value is also called popularity, but we use interestingness instead to avoid a possible confusion.

Our work is motivated by the need to understand to what extent popularity metrics are relevant to identify interesting contents. We also aim at describing the social dynamics by which some contents on the Web succeed in attracting significant collective attention, while most of them fail.

2 Related Work

Social influence, or the fact that individuals do not make decisions independently, but rather are influenced by the behavior of others, has long been studied by psychology (Cialdini and Goldstein 2004). Social ties viewed as the vehicles of word-of-mouth referrals have received much interest from marketing research (Brown and Reingen 1987). More recently, by making use artificial markets, it has been shown that social influence creates inequality and unpredictability of success (Salganik, Dodds, and Watts 2006).

The emergence of online social networks has drawn a lot of interest from computer scientists, notably because of the availability of observational data. Empirical studies have highlighted their importance in e-commerce decision-

making (Kim and Srivastava 2007) or adoption of user-created contents such as pictures (Cha et al. 2008), or objects and places in virtual worlds (Bakshy, Karrer, and Adamic 2009). Theoretical approaches have also been explored to study information flow or adoption behaviors in networks, often with models borrowed from the study of epidemics (e.g., Wu et al. 2004).

Both empirical and theoretical works consider the social networks as a static conveyors of information, whereas we adopt the view that in fact, links are created according to the dynamics of information. From that perspective, our work is also related to the broader physics literature on networks dynamics, for instance Barabasi and Albert (1999).

With respect to the issue of popularity, a previous result is that the popularity of contributed contents is related to the visibility (i.e. the number of links in the social network) of their authors (Lerman 2006; Wu, Wilkinson, and Huberman 2009). Our work clearly builds upon this observation and the existence of a “tyranny of the minority” whereby a small group of contributors controls *de facto* what is popular (Lerman 2006).

3 Methods

3.1 Digg, dataset and data processing

The purpose of the Digg website is to provide a digest of the most interesting news-like content published on the Internet. Digg aggregates and displays lists of *stories*, which are URL to news articles, blog posts, videos or photos. Digg’s crowdsourcing scheme involves users both to *submit* (contribute) stories and to *digg* (vote for) stories they liked. Following their submission, all stories are displayed in the *upcoming* page in a reverse-chronological order. Some stories receive sufficient number of diggs in a limited period of time and are then selected (*promoted*, in the Digg jargon) to be displayed on the Digg homepage, where they are called *popular* stories.

Since hundreds of stories are submitted every hour, users do not read them all. Instead, they mainly use two ways of browsing stories. One is to browse popular stories displayed on the Digg front page and another is to follow what other users are digging (Szabo and Huberman 2010). Indeed, Digg implements its own social network such that if user A becomes a *fan* of user B, A will then see what B has “dugg” recently. Note these links are directed.

When a story has become popular, it undergoes a fast accumulation of diggs, the total amount of which we define as its interestingness. That this number reflects the actual interestingness of the story is supported by results from Szabo and Huberman (2010) showing that most diggs on popular stories are not prone to social influence, because they come mainly from users who are browsing popular stories listed on the Digg front page.

Digg was launched in December 2004 but the social filtering features were released only in August 2005. Our dataset consists of downloaded records of 108 million diggs on 6.1 million stories, from 3 distinct time periods: a) January 2007 - November 2007, b) June 2008 - July 2008, and c) December 2008 - May 2009. Each story record contains its unique

identifier, the username of the submitter and timestamps of submission and promotion time (if ever promoted). Each digg record contains the username of the digger, a timestamp and the identifier of the story it is related to. Thus, we are able to track all diggs from any individual users, as well as to count all the diggs received by a particular story. We also retrieved all the fan links between the users with the timestamps of their creation.

3.2 Social influence analysis

To measure the effect of social influence, we constructed for each popular story, a directed graph of diggs where vertices are diggs in the upcoming phase. An edge pointing from A to B means that B occurred after A and user who dugg B is a fan of user who dugg A. A is a *parent* digg and B a *child* digg. A first approach to quantify social influence is to count the proportion μ_{up} of child diggs.

However, a certain proportion $\mu_{0,up}$ of edges is in fact explained by randomness, and is related to the average degree of the social network (i.e. the average number of links per user). Intuitively, the more people are connected, the more likely independent diggs will appear as parent-child related. Thus we offset this randomness by taking the social effect ν to be equal to the difference: $\nu = \mu_{up} - \mu_{0,up}$

To approximate $\mu_{0,up}$, we observe that the same equation holds for the popular phase during which the number of diggs that can be attributed to the social network is negligible (Szabo and Huberman 2010), and therefore the social effect is zero. In addition, the network average degree varies only very little during the lifetime of a story, and so does the randomness term. Thus, $\mu_{0,up} \approx \mu_{0,pop} = \mu_{pop}$ and $\nu = \mu_{up} - \mu_{pop}$

To reduce stochastic noise, we actually averaged μ_{pop} over all stories submitted in a month and used that single value $\langle \mu_{pop} \rangle$. During the period under consideration, $\langle \mu_{pop} \rangle$ increased from 0.05 to 0.3, reflecting an increase in the network average degree.

4 Results

As pointed out in Wu, Wilkinson, and Huberman (2009), attention received by contributed contents increases with the network of their authors. In the case of Digg, this translates into the fact that the users who contribute popular stories (call them *successful* users) also those who have a higher number of fans (Lerman 2006).

We first verify this correlation by comparing for each month in our dataset the average number of fans of successful users, to the average number of fans of all users who have submitted at least one story during the month. As expected, the ratio ρ of these two numbers is far greater than 1 (Fig. 1a), highlighting the correlation between having a large number of fans and contributing popular stories. Although our observation does not demonstrate causality, previous empirical (Lerman 2006) and theoretical (Lerman 2009) studies on Digg have suggested there is a direct cause effect.

We also observe that ρ increases until December, 2008. To interpret this increase, one can view ρ as a measure of inequality in number of fans, between users who contributed

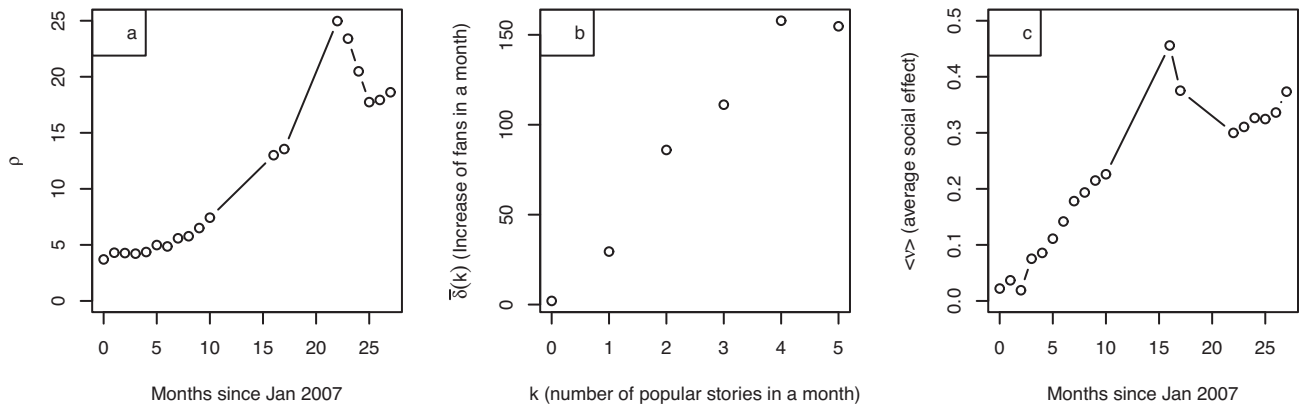


Figure 1: *a*) Relative number of fans of successful submitters vs. all submitters (ρ). Users who submitted a story that became popular are distinguished by a higher number of fans, as observed by $\rho > 1$. *b*) Acquiring new fans and submitting popular stories. The increase in the number of fans during a month is correlated to the number of popular stories submitted in that month. *c*) Evolution of social effect. The monthly averaged value $\langle \nu \rangle$, which measures the effect of social filtering in making stories popular, has substantially increased until June, 2008 ($x = 17$).

popular stories and all users who contributed. Thus, the increase in ρ suggests that fan links are getting more unequally distributed. Indeed, we observe that the 80% “poorest” users (in number of fans) have had their share of the total number of fans decreasing from 24% to 7% between March, 2007 and March, 2009 (not shown).

Thus, on Digg, popularity is explained by social influence. Besides, links in the social network are unequally distributed around successful users, and this inequality has increased. An explanation for this evolution is that success itself brings new fans. This would result in a positive feedback loop: users having many fans easily gain success which in return, brings them new fans. To support this hypothesis, we measured for each user i and each month j , the increase in number of fans $\delta_{i,j}$ between the start and the end of the month. We restricted our analysis to the months between October 2007 and July 2008, within which most of the network growth is concentrated. Within each month, we considered only the users who submitted at least one story. This gave us over 300 thousand $\delta_{i,j}$ values. We then plotted $\bar{\delta}(k)$, the average value of $\delta_{i,j}$ where users i had k stories promoted to the front page during month j . We can observe on Fig. 1b that $\bar{\delta}(k)$ grows with k , suggesting that success indeed plays a role in *acquiring* new fan links. We did also control the variable k by the initial number of fans, and this did not significantly change our result. This effect saturates quickly, which is not surprising given that the number of new links one can make in a month is finite.

Put together, these results bring evidence of a feedback mechanism between visibility of contributors (number of fans) and the popularity of their contributions. It is straightforward to see how this mechanism can explain the highly unequal fans distribution in the Digg social network.

Next, we formulate the hypothesis that social influence not only has an effect on which stories become popular but also affects the interestingness of popular stories. As de-

tailed in the Methods section, we quantify the effect of social influence ν on a per-story basis, while interestingness is approximated by the total number of diggs received by a story.

For each month we calculated $\langle \nu \rangle$, the average value of ν of all stories that became popular during the month. We first verified that $\langle \nu \rangle$ has increased in the given period, from 0.00 to 0.37, as shown on Figure 1c, reflecting a growing effect of social influence in acquiring the diggs during the upcoming phase and therefore in becoming popular.

To investigate the relationship between social influence ν and interestingness, we then considered within each month window the stories in the first 20% (Q_1) and the last 20% (Q_5) quantile with respect to ν values. Stories in Q_1 and Q_5 are those with the highest and lowest ν , respectively. We compared their interestingness by calculating:

$$N_\nu = \frac{\text{avg. interestingness of stories in } Q_1}{\text{avg. interestingness of stories in } Q_5}$$

Note that although $\langle \nu \rangle$ is increasing, we cannot expect to observe a variation in the monthly average interestingness of promoted stories. Indeed, for any month, the average interestingness of promoted stories is roughly proportional to the site traffic divided by the number of promoted stories, rather than related to the interestingness of stories.

A value of N_ν below 1 means that the stories with the highest ν values are (on average) less interesting than those with the lowest ν . In order to assess the significance of N_ν , we define N_{rand} as the value of N_ν below which the null hypothesis “the expected value of N_ν is equal to 1” can be rejected at a significance level of 99%. In other words, values of N_ν below N_{rand} cannot be explained by randomness alone. We calculated N_{rand} by numerical simulations. We also introduce N_{min} , by looking at the first 20% (R_1), and the last 20% (R_5) quantile with respect to interestingness.

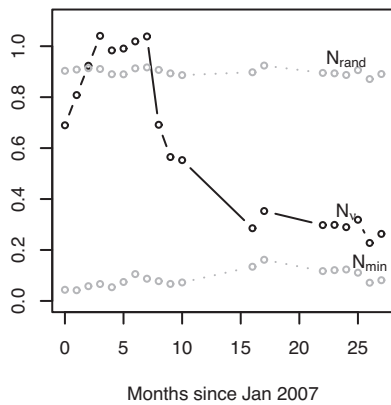


Figure 2: Relative interestingness of popular stories. N_{ν} measures how interesting are stories with higher ν values compared to stories with lower ν . Thus, $N_{\nu} < 1$ indicates that higher social effect correlates with lower interestingness.

$$N_{\min} = \frac{\text{avg. interestingness of stories in } R_1}{\text{avg. interestingness of stories in } R_5}$$

It is immediate to observe that N_{\min} is a lower bound of N_{ν} , with $N_{\nu} = N_{\min}$ only when the less interesting stories coincide perfectly with the stories with the higher social effect.

One can see that from month 0 to 7, N_{ν} fluctuates around 0.8, and is slightly below N_{rand} , but then rapidly drops down to 0.3 and fluctuates around this value (figure 2). First, this confirms that the more a story benefited from social effect to become popular, the less it is likely to be interesting. Second, Figure 2 shows that this phenomenon has been drastically accentuated, with N_{ν} having values close to N_{\min} .

5 Conclusion

We have provided evidence that social influence affects the popularity of content, and that there is a feedback loop between the visibility of contributors and the popularity of their contributions. Finally, by quantifying the effect that social influence had in making each story popular, we show that stories which have benefited from social influence are on average less interesting than those which have not.

While our study focuses on Digg only, almost every social filtering website (e.g., stumbleupon.com, reddit.com) utilize popularity in similar manners. When popularity-based rankings are not a feature of the website itself, the availability of API allows for third-party websites to spot popular stories. For instance, twitterly.com and tweetmeme.com compute lists of popular stories from Twitter. Again, this is justified by the attractiveness of what can be labeled as popular. In all these systems, it seems inevitable that social influence also contributes to make stories popular since, by nature, social filtering captures attention. The 100 most-connected Twitter

users have over 1 million followers¹ and each of their stories are likely to be popular. Then the question is, will users connect to contributors of popular stories, as we observe in Digg? In this respect, it would be valuable to study how network links are created in other social filtering systems, and investigate whether our results hold.

References

- Bakshy, E.; Karrer, B.; and Adamic, L. A. 2009. Social influence and the diffusion of user-created content. In *Proceedings of the tenth ACM conference on Electronic commerce*, 325–334. Stanford, California, USA: ACM.
- Barabasi, A. L., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286(5439):509.
- Brown, J. J., and Reingen, P. H. 1987. Social ties and Word-of-Mouth referral behavior. *The Journal of Consumer Research* 14(3):350–362. ArticleType: research-article / Full publication date: Dec., 1987 / Copyright © 1987 Journal of Consumer Research Inc.
- Cha, M.; Mislove, A.; Adams, B.; and Gummadi, K. P. 2008. Characterizing social cascades in flickr. In *Proceedings of the first workshop on Online social networks*, 13–18. Seattle, WA, USA: ACM.
- Cialdini, R. B., and Goldstein, N. J. 2004. Social influence: Compliance and conformity. *Annual Review of Psychology* 55(1):591–621.
- Kim, Y. A., and Srivastava, J. 2007. Impact of social influence in e-commerce decision making. In *Proceedings of the ninth international conference on Electronic commerce*, 293–302. Minneapolis, MN, USA: ACM.
- Lerman, K. 2006. Social networks and social information filtering on digg. *cs/0612046*.
- Lerman, K. 2009. Dynamics of a collaborative rating system. In *Advances in Web Mining and Web Usage Analysis*. 77–96.
- Salganik, M. J.; Dodds, P. S.; and Watts, D. J. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762):854–856.
- Szabo, G., and Huberman, B. A. 2010. Predicting the popularity of online content. *Communications of the ACM* 53:80–88. ACM ID: 1787254.
- Tran, N.; Min, B.; Li, J.; and Subramanian, L. 2009. Sybil-resilient online content voting. In *Proceedings of the 6th USENIX symposium on Networked systems design and implementation*, 15–28. Boston, Massachusetts: USENIX Association.
- Wu, F.; Huberman, B. A.; Adamic, L. A.; and Tyler, J. R. 2004. Information flow in social groups. *Physica A: Statistical and Theoretical Physics* 337(1-2):327–335.
- Wu, F.; Wilkinson, D.; and Huberman, B. 2009. Feedback loops of attention in peer production. In *Computational Science and Engineering, 2009. CSE '09. International Conference on*, volume 4, 409–415.

¹Ranking and statistics from <http://twitterholic.com/>.