Influence and Passivity in Social Media

Daniel M. Romero
Center for Applied Mathematics
Cornell University
Ithaca, New York, USA
dmr239@cornell.edu

Wojciech Galuba
Distributed Information Systems Lab
EPFL
Lausanne, Switzerland
wojciech.galuba@epfl.ch

Sitaram Asur
Social Computing Lab
HP Labs
Palo Alto, California, USA
sitaram.asur@hp.com

Bernardo A. Huberman
Social Computing Lab
HP Labs
Palo Alto, California, USA
bernardo.huberman@hp.com

ABSTRACT

The ever-increasing amount of information flowing through Social Media forces the members of these networks to compete for attention and influence by relying on other people to spread their message. A large study of information propagation within Twitter reveals that the majority of users act as passive information consumers and do not forward the content to the network. Therefore, in order for individuals to become influential they must not only obtain attention and thus be popular, but also overcome user passivity. We propose an algorithm that determines the influence and passivity of users based on their information forwarding activity. An evaluation performed with a 2.5 million user dataset shows that our influence measure is a good predictor of URL clicks, outperforming several other measures that do not explicitly take user passivity into account. We demonstrate that high popularity does not necessarily imply high influence and vice-versa.

Categories and Subject Descriptors

K.4.0 [**Computers and Society**]: General

General Terms

Algorithms, Measurement

1. INTRODUCTION

The explosive growth of Social Media has provided millions of people the opportunity to create and share content on a scale barely imaginable a few years ago. Given the widespread generation and consumption of content, it is natural to target one's messages to highly connected people who will propagate them further in the social network. This is particularly the case in Twitter, which is one of the fastest growing social networks on the Internet, and thus the focus of advertising companies and celebrities eager to exploit this vast new medium. As a result, ideas, opinions, and products compete with all other content for the scarce attention of the user community. Given this level of competition and chaos, there is considerable consensus that two aspects of information transmission seem to be important in determining which content receives attention.

One aspect is the popularity and status of members of these social networks, which is measured by the attention they receive from the consumers of their content. The other aspect is the influence that these individuals wield, which is determined by the actual propagation of their content through the network. This influence is determined by many factors, such as the novelty and resonance of their messages with those of their followers and the quality and frequency of the content they generate. Equally important is the passivity of members of the network which provides a barrier to propagation that is often hard to overcome. Thus gaining knowledge of the identity of influential and least passive people in a network can be extremely useful from the perspectives of viral marketing, propagating one's point of view, as well as setting which topics dominate the public agenda.

In this paper, we analyze the propagation of web links on Twitter over time to understand how attention to given users and their influence is determined. We devise a general model for influence using the concept of passivity in a social network and develop an efficient algorithm similar to the HITS algorithm [3] to quantify the influence of all the users in the network. Our influence measure utilizes both the structural properties of the network as well as the diffusion behavior among users. The influence of a user thus depends not only on the size of the influenced audience, but also on their passivity. This differentiates our measure of influence from earlier ones, which were primarily based on individual statistical properties such as the number of followers or retweets [1].

2. THE IP ALGORITHM

The Twitter data set. On Twitter, each user submits periodic status updates, known as *tweets*, which are short messages limited to 140 characters. Each user independently decides what other users to follow. An important Twitter phenomenon that is central to this paper is retweeting. A *retweet* is a post originally made by one user that is forwarded by another user. They are useful for propa-

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

gating interesting posts and links through the Twitter community. Each retweet explicitly credits the author of the original tweet [4] and is an important influence signal.

To obtain the dataset for this study, we continuously queried the Twitter Search API for a period of 300 hours starting on 10 Sep 2009 for all tweets containing URLs (the string `http`). The dataset consists of approximately 22 million tweets mentioning unique 15 million URLs [1].

Algorithm outline. An important question is whether it is possible to identify users who are very good at spreading their content to a large part of the network. While pairwise influence between users can be easily determined, it is not very clear how to accurately obtain information about the relative influence each user has on the whole network. To answer this question, we design an algorithm (IP) that assigns a relative *influence* score and a *passivity* score to every user. The passivity of a user is a measure of how difficult it is for other users to influence him. The algorithm takes into account the passivity of all the people influenced by a user, when determining the user's influence. In other words, we assume that the influence of a user depends on both the quantity and the quality of the audience she influences.

Algorithm operation. The algorithm operates iteratively, computing both the passivity and influence scores simultaneously in the following way:

Consider a weighted directed graph $G = (N, E, W)$ with nodes N, arcs E, and arc weights W, where the weights w_{ij} on arc $e = (i, j)$ are computed as follows: The arc (i, j) exists if user j retweeted a URL posted by user i at least once. The arc $e = (i, j)$ has weight $w_e = \frac{S_{ij}}{Q_{ij}}$ where Q_i is the number of URLs that i mentioned and S_{ij} is the number of URLs mentioned by i and retweeted by j.

For every arc $e = (i, j) \in E$, we define the *acceptance rate* by $u_{ij} = \frac{w_{i,j}}{\sum_{k:(k,j)\in E} w_{kj}}$, which the influence of user i over j normalized by the sum of influence from all the nodes that affect j. Similarly we define the *rejection rate* by $v_{ji} = \frac{1 - w_{ji}}{\sum_{k:(j,k)\in E} (1 - w_{jk})}$. Since the value $1 - w_{ji}$ is amount of influence that user i rejected from j, then the value v_{ji} represents the influence that user i rejected from user j normalized by the total influence rejected from j by all users in the network.

The algorithm is based on the following operations performed iteratively:

$$I_i \leftarrow \sum_{j:(i,j)\in E} u_{ij} P_j \quad (1)$$

$$P_i \leftarrow \sum_{j:(j,i)\in E} v_{ji} I_j \quad (2)$$

3. EVALUATION

To validate whether our algorithm is a good predictor of the attention URLs get on Twitter, we use click data from Bit.ly, a URL shortening service that keeps track of how many times a shortened URL has been accessed. For the 3.2M unique Bit.ly URLs in our dataset, we queried the Bit.ly API for the number of clicks on them. Since one URL can have several Bit.ly shortenings, we sum the clicks over all observed Bit.ly shortenings for each URL.

[1]The URLs shortened via the services such as `bit.ly` or `tinyurl.com` were expanded into their original form by following the HTTP redirects.

Measure	R^2
Number of followers	0.59
Number of retweets	0.02
PageRank	0.84
Hirsch Index	0.05
IP-Influence	**0.95**

Table 1: Comparison of Influence Measures

URL traffic Prediction. Using the URL click data, we take several different user attributes and test how well they can predict the attention the URLs posted by the users receive. It is important to note that none of the influence measures are capable of predicting the exact number of clicks. The main reason for this is that the amount of attention a URL gets is not only a function of the influence of the users mentioning it, but also of many other factors including whether the URL was mentioned anywhere outside of Twitter. In view of that, we look at how each influence metric predicts the maximum number (or rather the $99.9th$ percentile, to eliminate the outliers) of clicks a user can get on the links they post.

As shown in Table 1, we observe that the average IP-influence of those who tweeted a certain URL can determine the maximum number of clicks for a URL with good accuracy, achieving an R^2 score of 0.95, and significantly outperform other measures of influence. Since the URL clicks are not used by the IP algorithm to compute the user's influence, the fact that we find a very clear connection between average IP-influence and the eventual popularity of the URLs (measured by clicks) serves as an unbiased evaluation of the algorithm and demonstrates the utility of IP-influence.

4. CONCLUSIONS

Our study shows that the correlation between popularity and influence is weaker than expected. This is a reflection of the fact that for information to propagate in a network, individuals need to forward it to the other members, thus having to actively engage rather than passively read it and rarely act on it. An evaluation performed with a 2.5 million user dataset shows that our influence measure is a good predictor of URL clicks, outperforming several other measures that do not explicitly take user passivity into account.

5. REFERENCES

[1] Meeyoung Cha and Hamed Haddadi and Fabricio Benevenuto and Krishna P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, Vol 30, 1-7, 1998.

[3] Jon Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46 (5): 604 -632, 1999.

[4] Boyd Danah, Scott Golder, and Gilad Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. HICSS-43. IEEE 2010.