A Connectivity-Based Popularity Prediction Approach for Social Networks

Huangmao Quan, Ana Milicic, Slobodan Vucetic, and Jie Wu
Department of Computer and Information Sciences
Temple University, Philadelphia, PA 19122
{homerq,anam,vucetic,jiewu}@temple.edu

***Abstract*—In social media websites, such as Twitter and Digg, certain content will attract much more visitors than others. Predicting which content will become popular is of interest to website owners and market analysts. In this paper, we present a novel technique to predict popularity using the connection features of individuals and their community. Our approach is based on the hypothesis that connection plays a dominant role in spreading content on social media. The resulting predictor is more efficient than approaches which estimate popularity by complex graph properties, and more accurate than approaches that use simple visit counts. We evaluated the proposed approach empirically on several real-life data sets. Results indicate that, compared with the conventional methods, our approach is both accurate and computationally efficient.**

***Index Terms*—Social networks, social media, machine learning, popularity prediction**

I. Introduction

For web content, popularity represents the corresponding amount of attention from users. Unlike traditional media, such as newspapers and television, that have a centralized mode of providing content to users, content in social networks is often spread from one user to another over a period of time, making it difficult to measure and predict the popularity of a particular content. For example, in Twitter, the popularity of a user's content is influenced not only by the number of followers, but also by the number of connections those followers have. A content viewed by followers with more connections will propagate faster, and hence be more popular than content that is viewed by a similar number of followers with less connections. This necessitates new techniques of predicting the popularity of social network content. In this paper, we present and validate the hypothesis that the popularity of social network content can be predicted using the connection structure of the network.

Accurate techniques of determining popularity are important, since popularity can be used to support additional applications. For instance, popularity can be used in performance optimization in the following way: When a content becomes popular on a social network site, it will generate an increase in traffic, which in turn can lead to a slowing down, or even a shutting down of a web site. An accurate prediction of popular content can allow the website to deploy additional resources, such as web servers, to manage the load. An inaccurate prediction, on the other hand, can lead to misallocated resources. Popularity is also an important measure for online marketing.

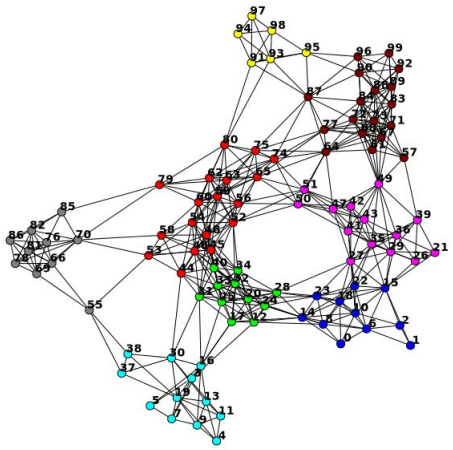

Fig. 1. The high connection coefficient in a social graph

If the popularity can be predicted accurately and early, better marketing strategies can be adopted or different advertising rates can be charged.

The goal of this paper is to develop an effective and general method to predict popularity of a given content. The key assumption was that each individual has a different influence in spreading popularity of a given content. It is a significant challenge to measure influence of an individual. Kemp and Kleinberg [1] proposed an approximated solution for selection of the most influential nodes in a social network to maximize the spread of influence. Our approach predicts popularity based on connections, which is based on the hypothesis that network connections play a dominant role in the spreading of content over the social network. Our approach distinguishes itself by making predictions based on features of the connection, rather than features of content. Our experiments show that predicting popularity using connections is accurate. As will be shown, prediction using connection features can achieve 0.9 AUC accuracy, significantly higher than the 0.8 AUC when connection features are not used. The data sets in our experiment were collected from real social media, namely Digg and Metafilter. The experimental evaluation on these data sets allowed us to measure the importance of connection features in determining long-term popularity of content.

Our method is appealing for a number of reasons: First, it can be applied to any social media with connection information; Second, it is computationally light, since connection

features are easy to calculate; Third, it measures both inter-community and intra-community connections; Finally, by approximation and using linear-time clustering algorithms, our method could be implemented on large-scale social media.

The paper is organized as follows: In Section II, we briefly sketch the related research. In Section III, we describe the hypothesis and data set extraction and cleaning. In Section IV, our approach is evaluated on several data sets and the results of experiments are presented and analyzed. Finally, Section V concludes the paper.

II. Related work

Popularity measurement is one of the most fundamental and practical objectives in social network analysis (SNA). Most of the existing methods predict popularity by counting the number of visitors [2]–[4]. Previous research [5] showed that the key to understanding information flow and popularity is how actors are connected: Individual adoption is more likely when participants receive social reinforcement from multiple neighbors in the social network. The influence spreads farther and faster across clustered-lattice networks than across the corresponding random networks [5]. As a result, Kemp and Kleinberg [1] proposed an approximated solution for selecting the most influential nodes in the social network to maximize the spread of influence.

A significant research effort has been made on the topic of measuring relationships between actors connected through a network represented by a directed graph [6]–[10]. The previous research has found [11] that the networks with a high degree of separation are less effective for content diffusion. Granovetter studied connections between communities, which are called *weak ties* [12], and argued that weak ties bind groups together into the larger society and are crucial for the spread of information. If content can seize more weak ties, it has more of a chance of spreading broadly. It is worth noting that the structure of a social graph can also be studied in the context of contagious disease outbreaks [13]. For example, if a contagion spreads from person to person in a network, centrally located individuals are more likely to be infected than peripheral individuals.

III. Methodology

A. Connectivity of individuals and communities

To predict popularity, only measuring connectivity of individuals is not sufficient. Although it appears that individuals' connectivity is correlated to popularity [11], the usefulness of these features in determining long-term popularity is still uncertain. For example, prior research showed that the popular stories exhibit an extreme diversity in connections of the individuals that created them [14].

Our starting point is an insight that people tend to be organized in communities. A community represents a large group of people who are connected to one another through paths in the social graph. Within a community, the nodes are closely connected. The number of connections each node has is called the connection coefficient. The popularity of content can be observed by how social interactions directly or indirectly related to the connection coefficient: The higher the coefficient, the greater the reach. Also, within a community, people have stronger connections and more frequent interaction. Fig 1 shows a typical connection coefficient in social networks.

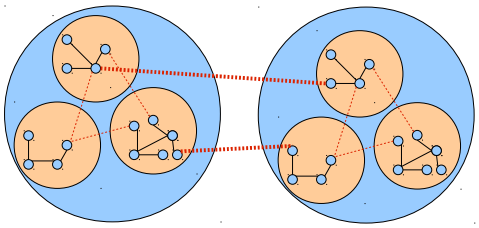

Fig. 2. Nested communities: Demonstration of cascaded communities

The closely connected structure of a sub-community ensures reaching everyone in a sub-community. However, it does not guarantee reaching people outside the community. Therefore, to estimate content popularity, the principle of spreading between communities also has to be addressed. Since most propagations in social networks are within 2 to 3 hops [8], we model networks as 2 level nested communities whose individuals are reachable within 2 hops (Fig 2).

Finding a community is costly and slow for large scale social networks. To address this, we use greedy optimization of modularity [8] to find communities. This algorithm merges individual nodes into communities in a way that greedily maximizes the modularity score of the graph. Considering that a large social network can easily have over one million edges, this algorithm is very appealing because it can run in almost linear time on sparse graphs of typical social networks.

B. Digg and Metafilter

Our approach was tested on data sets from two types of social media. In this section, we briefly introduce how they work, with an emphasis on how information is spread along their social networks.

Digg provides a social bookmarking service to over 3 million registered users. Its core service is Digg, which allows users to share interesting content with their friends. Digg works as follows: Users can choose to browse the front page, which shows recently promoted stories, or the upcoming stories page, which shows recently submitted stories. In addition, users can also view the stories their friends recently submitted or voted for. They can select one of the stories to read and, depending on whether they consider it interesting, vote for it. Alternatively, after visiting Digg's pages, users may choose to leave it. The user's environment is changing in time depending on actions of all other users. All the posts are recorded together with their time stamp and if the story is popular enough, it can receive the status of a "top" story. The criterion that determines whether the story becomes a top story and when is proprietary and unknown to the public.

MetaFilter is a community of web bloggers and includes several sub-sites, which share the same database. Unlike

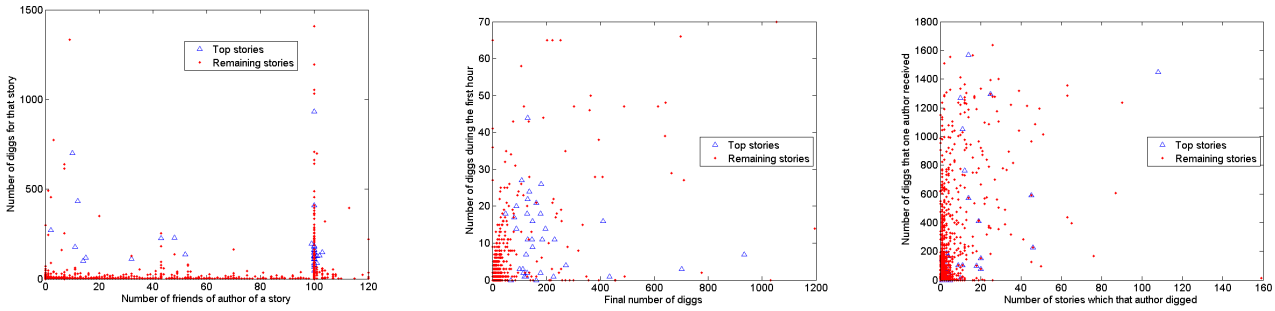

Fig. 3. Correlations between features and target variable for the Digg data set

other social network sites, MetaFilter has developed a fairly stable community, mostly because of member fees. Like Digg, MetaFilter also allows users to share interested stories with friends by marking them as favorites. The number of comments is used as the sole measurement of popularity.

C. Data collection

The Digg data set was acquired from its API 1.1. The API allowed open access to Digg data by remote HTTP callback of its web services. We wrote a set of Python scripts to send requests and then parsed the returned XML data. The final data that we collected had $5,000$ stories with $2,684$ authors of those stories, 2-level social graph with $117,926$ users and $1,164,613$ edges, and $19,645$ posts.

The MetaFilter data set was directly downloaded from database. It included statistics about posts and user relationships. Friendships were given in a form of pairs of user IDs with the time and date of when the friendship was established. The data also contained the whole social graph of its community containing a list of friendships, list of votes, and four lists of stories.

D. Data cleaning

Since the MetaFilter data set was extracted from the database, it was easy to preprocess it. However, a substantial effort was needed to clean and preprocess the Digg data set. First, we did not have an access to the complete social graph. Second, due to the restrictions at the Digg server, we were able to download only the first 100 friends of a user and the first 100 votes of a story.

In the case of Digg data set, the number of posts for each story can be found in table *stories* and also can be obtained by counting the number of appearances of that story in the table *diggs*. By comparing those two numbers we concluded that not all the posts were reported. We also found that the total number of posts was not sufficient to distinguish popular stories from the rest.

There were 4 types of missing data: 1. user is listed in the user table, but missing in the post table; 2. user is listed in the post table but not in the user table; 3. missing votes for a post; 4. missing friends for a user.

For missing data of type 1 and 2, if we decided to remove all the points with missing Digg reports, a large number of top stories would be discarded. Therefore, we decided to keep as many of the stories as possible. Since we were interested in following each story for one hour after its publishing and then predicting whether it will become a top story at some time in the future, missing data of type 3 (stories without reported votes) could have detrimental effect on accuracy. We found that there are two top stories without recorded votes, so we decided to remove them. For missing data of type 4, we found that 91.6% of all users had no reported friends. In addition, 10.2% of authors did not have reported friends. Regardless of this, we decided to retain users without friends and authors without friends and and their stories. Once the collected data set was cleaned from the missing data, we randomly divided it into training and test sets.

E. Data characteristics

In our preliminary study, we found that initial and future popularity were correlated. In our data, the number votes during the first hour after publishing was correlated with the total number of votes during the whole life time of the story.

Some of the results of an exploratory analysis for the Digg data set are shown in Figure 3. The Metafilter data set showed similar characteristics. In the scatter plot at the top, we observe that although a large number of authors of the top stories have 100 or more friends, this is not an exclusive rule. A considerable number of top stories came from the authors with very small number of friends. Scatter plot in the middle shows that there is no strictly linear correlation between number of votes during the first hour after publishing and the final number of votes. Also, it can be observed that a large number of votes does not necessarily mean that the story is going to be a *top story*. The plot at the bottom shows that there is no obvious "returning of favor" behavior, where authors who vote favorably for other stories would receive larger number of votes for their own stories.

For Digg data set, there are 176 clusters with more than 10 users. There are four clusters with more than $5,000$ users, the largest of them having $81,000$ people. Metafilter is a much bigger data set and there is an even larger number of clusters of all sizes. The clusters are build from bottom up using the greedy optimization of modularity [15]. This algorithm merges individual nodes into communities in a way that greedily maximizes the modularity score of the graph. It can be proven that if no merge can increase the current modularity score, the algorithm can be stopped since no further increase can

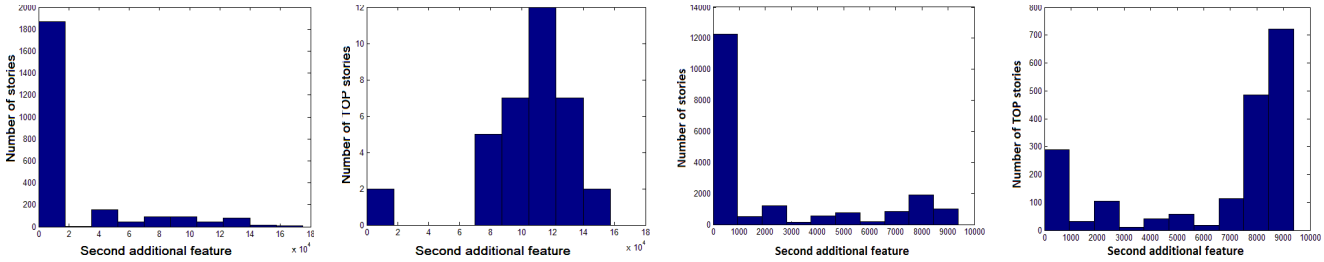

Fig. 4. (a) Digg connection features for all stories (b) Digg connection features on popular stories (c) Metafilter connection features for all stories (d) Metafilter connection features for all stories

be achieved.Since we have almost one million edges, this algorithm can run almost in linear time on sparse graphs.

F. Features

We used data from Digg and Metafilter to evaluate importance of community structure on content popularity. Our hypothesis was that the knowledge of the author's position in a network is helpful in prediction of popularity of his or her stories. Specifically, the task was to predict the popularity of posted stories based on the response of other users to the post (the number of digs that story receives) and information about the author. Having this in mind, features describing each story during the first hour of its publishing ware extracted. The features we assumed to be relevant for prediction were:

1. Number of friends of author of that story
2. Number of digs of that author for other people's stories (for detecting voting as "returning of a favor")
3. Number of digs of all friends of that author for all the other stories (describes low or high activity of friends of this user)
4. Number of stories posted by that author
5. Number of digs that occurred after one hour of publishing the story
6. Topic (binary feature that is 1 for more popular topics and 1 for remaining)

All of these features were expected to have a positive correlation with popularity of the story. The remaining question was how to measure popularity of a post. With Digg, this was not an issue considering that it implements a service called *top story*. Periodically, a certain number of stories are given the status of *top story*, according to an unknown criterion by Digg. This status was used as a target binary variable to be predicted. With the Metafilter social network, the preprocessing was slightly more complicated, as it does not explicitly decide which story is a *top story*. Here, the story was labeled as *top story* if it had more than a specified number of votes. We used the threshold such that 10% of stories are considered to be a *top story*.

IV. Evaluation

Our goal was to show that popularity of content is determined by the characteristics of the cluster rather than by characteristics of the author or the story. For this reason, experiments were conducted in two stages. First, we trained a prediction model using only features that describe the author and the story. Second, using the same learning algorithm, we trained the prediction model using all the features, including the two based on knowledge of clusters. By comparing the accuracy of those two models, we were able to draw conclusions about significance of the connection-based features.

After some preliminary exploration, we observed that linear regression, which is able to capture the linear correlation between features and target, is very appropriate for popularity prediction. The outcome of a linear regression predictor is a real-valued number, where a large prediction can be interpreted as large confidence that a story is a top story. For evaluation of prediction accuracy, we used the Area Under the Curve (AUC) accuracy, representing the area under the ROC curve. The ROC curve is a plot of True positive (TP) rate versus the False Positive (FP) rate as the decision threshold varies from small to large values. For a small decision threshold, almost all stories are predicted as top stories, while for a large decision threshold, almost no stories are predicted as top stories. The larger the area under the ROC, the better the quality of the predictor. The results for the first set of experiments that did not use connectivity-based features are shown in the first column of Table I.

A key objective of our experiments was to observe how much is accuracy influenced by addition of the connectivity-based features. In particular, we added the following two features to our linear regression prediction model. (1) Number of users in an author's cluster, (2) The cumulative size of all clusters from which at least one user read the story within the first hour of publishing

The first feature added was the size of the cluster to which the author belongs. The second was created by summing the sizes of all clusters whose users accessed the story. Both features turned out to be very interesting as its distribution in top stories was very different from its distribution in all stories. By a further examination, we observed that users from the second largest cluster posted most of the stories and that this cluster posted the most of the top stories. From Fig. III-Ea and Fig. III-Eb, we can observe that almost all of the top stories were spread to big clusters very soon after they were published. On the other hand, the majority of the stories that did not become popular never reached members of these big clusters.

Similar behavior was observed on the Metafilter data set

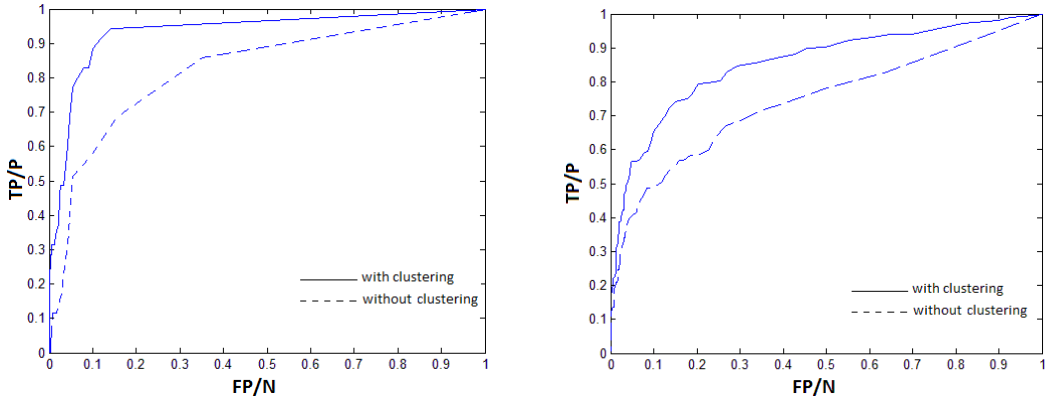

Fig. 5. ROC curves for Digg data set and Metafilter data set

Fig. III-E. A significant difference in the shapes of these two histograms implies the potential usefulness of the second feature in popularity prediction. As can be seen from Table I, the AUC was significantly improved once the two connectivity-based features were added. To gain a further insight into the performance, in Fig. 5 we show the ROC accuracy curves for the two data sets.

The large improvement of accuracy indicates that story popularity does not depend too much on the characteristics of its author or the overall number of votes the story receives during the first hour since its publishing. A more relevant factor is whether the author is a member of a big or small cluster, and how many users from other communities voted for his story during the first hour. To verify the conclusions, the experiments were repeated using only the two connectivity-based features (last column in Table I). The results show that the accuracy upon removing the non-connectivity-based features did not change nuch, thus confirming our observation. This result has some interesting potential consequences in social network analysis.

V. Conclusion

Our work has shown that popularity is weakly related to inherent content quality. Instead, the network connectivity seems to play a dominant role in deciding a fate of a published content. We claim that social dynamics of users on a social media site allows us to quantitatively characterize evolution of popularity of items on that site and study the extent of connection in effecting popularity. Specifically, we proposed a hypothesis that network connections play a dominant role in the spreading of content over the social network. After comparing the predicated breakout with the real-life popularity, we know that connection features can be used to predict long-term popularity. By comparing prediction results from connection features and non-connection features, we found connection-based prediction is more effective.

Data set	w/o connection features	All features	2 connection features
Digg	0.8248	0.9344	0.9327
MetaFilter	0.8439	0.8644	0.8325

TABLE I
AUC COMPARATION OF CONNECTION AND NON-CONNECTION FEATURES.

Acknowledgements

This research was supported in part by NSF grants CCF 1028167, CNS 0948184, ECCS 1128209, CNS 1065444, and CCF 0830289.

References

[1] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, (New York, NY, USA), pp. 137–146, ACM, 2003.

[2] A. Kaltenbrunner, V. Gomez, and V. Lopez, "Description and prediction of slashdot activity," in *Proceedings of the 2007 Latin American Web Conference*, (Washington, DC, USA), pp. 57–66, IEEE Computer Society, 2007.

[3] K. Lerman, "Social information processing in news aggregation," *IEEE Internet Computing*, vol. 11, pp. 16–28, November 2007.

[4] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, pp. 661–703, November 2009.

[5] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, (New York, NY, USA), pp. 491–501, ACM, 2004.

[6] J. Wu and W. Lou, "Forward-node-set-based broadcast in clustered mobile ad hoc networks," *Wireless Communication and Mobile Computing*, vol. 3, pp. 155–173, 2003.

[7] F. Li and J. Wu, "Localcom: a community-based epidemic forwarding scheme in disruption-tolerant networks," in *Proceedings of the 6th Annual IEEE communications society conference on Sensor, Mesh and Ad Hoc Communications and Networks*, SECON '09, pp. 574–582, IEEE Press, 2009.

[8] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, pp. 1– 6, 2004.

[9] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Arxiv Preprint Physics*, vol. 74, 2006.

[10] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E*, vol. 76, Sep 2007.

[11] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," *Science*, vol. 329, no. 5996, pp. 1194–1197, 2010.

[12] M. Granovetter, "The Strength of Weak Ties," *The American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[13] N. A. Christakis and J. H. Fowler, "Social network sensors for early detection of contagious outbreaks," *PloS ONE*, vol. 5, September 2010.

[14] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, pp. 80–88, August 2010.

[15] http://igraph.sourceforge.net/doc/R/fastgreedy.community.html, "iGraph Library."