Following the Follower: Detecting Communities with Common Interests on Twitter

Kwan Hui Lim and Amitava Datta
School of Computer Science and Software Engineering
The University of Western Australia
Crawley, WA 6009, Australia
kwanhui@graduate.uwa.edu.au, datta@csse.uwa.edu.au

ABSTRACT

We propose an efficient approach for detecting communities that share common interests on Twitter, based on linkages among followers of celebrities representing an interest category. This approach differs from existing ones that detects all communities before determining the interest of these communities, a computationally intensive process given the large scale of online social networks. In addition, we also study the characteristics of these communities and the effects of deepening or specialization of interest.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and behavioral sciences

General Terms

Theory

Keywords

Twitter, Social Networks, Community Detection, Graph Mining

1. INTRODUCTION

One important problem in the application of target advertising and viral marketing to online social networks is the efficient identification of communities with common interests. Current approaches involve detecting all communities, then determining the interests of these communities [2, 5]. These approaches involve a lengthy and intensive process of detecting communities for the entire social network and many of the detected communities may not share the interest we are looking for. We propose a method to identify communities comprising like-minded individuals with common interests on Twitter. Also, our method does not unnecessarily detect communities that do not share any specific interest.

2. DATASET AND METHODS

The Twitter dataset collected by Kwak et al. [1] is used for our experimentations. A followership link (i, j) indicates that user i is a follower of user j, while a friendship link $Fr_{i,j}$ indicates $(i, j) = (j, i)$. We define celebrities as users with more than 10,000 followers. The interest of a user, Int_{cat} is inferred by the number of celebrities (of category cat) that the user follows.

Suppose we identify a set of k celebrities $c_1, c_2, ..., c_k$. We next identify all the followership links for the individual celebrities in this set. Consider celebrity $c_j, 1 \leqslant j \leqslant k$, and all the followership links for this celebrity $\bigcup_i link(i, c_j)$. We construct the set:

$$\mathcal{P} = \bigcap(\bigcup_i link(i, c_j)), for\ 1 \leqslant j \leqslant k$$

$\mathcal{P}$ is the set of fans who follow all the k celebrities in the set $\bigcup c_j, for\ 1 \leqslant j \leqslant k$. We consider only friendship links for community detection as friendship links are stronger and more reflective of real-life interactions. Next, we try to detect communities among the members of $\mathcal{P}$ using the Infomap algorithm and Clique Percolation Method (CPM) at a k-value of 3. Refer to [3] and [4] for more details on the CPM and Infomap respectively.

3. INVESTIGATING COMMON INTERESTS

For our study, we selected Film & TV, Music, Hosting, News and Blogging as categories of interest due to their popularity. For each category, we selected the six most popular celebrities based on their number of followers. The categories that these celebrities represent were determined using information from Google and Wikipedia. Following which, we selected users with $Int_{cat} = 6$, for $cat \in \{Film\&TV, Music, Hosting, News, Blogging\}$. As a control group, we randomly chose 200,858 users to represent the group with no shared interest. We now use our approach and compare the detected communities with common interests against the control group in terms of the total number of communities, size of largest community, and average community size.

Fig. 1 and 2 show that users with common interests form larger and more communities than users without a common interest. Similarly, users with common interests form larger communities on average as shown in Fig. 3. The exception is the News category detected using CPM as many cliques of three nodes were detected as communities thus decreasing the average community size.

Table 1: Network statistics of the communities

Category	Control	Film/TV	Music	Hosting	News	Blogging
Path Length	2.83	3.03	2.82	3.09	3.35	3.09
Clustering Coefficient	0.60	0.62	0.63	0.59	0.58	0.62
Diameter	6	7	8	8	8	7
Average Degree	7.81	6.80	7.29	8.17	9.15	7.51

Users with common interests also form communities that are more cohesive than those without common interest. Table 1 shows this trend where the communities with common interest have a higher clustering coefficient than our control group with no common interest, except the Hosting and News categories. However, users interested in Hosting and News have a higher average degree of links which shows that these users are better connected than users in the control group. These results show that our community detection approach finds communities that are larger, more cohesive and share common interests.

HT'12, June 25–28, 2012, Milwaukee, Wisconsin, USA.
ACM 978-1-4503-1335-3/12/06.

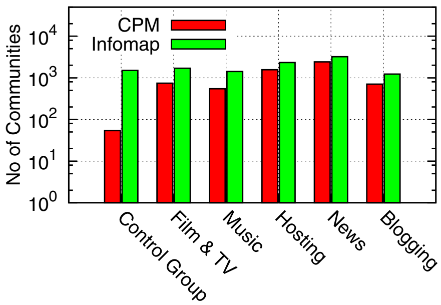

Figure 1: Total Communities

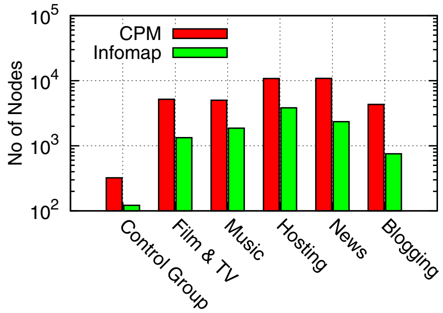

Figure 2: Size of Largest Community

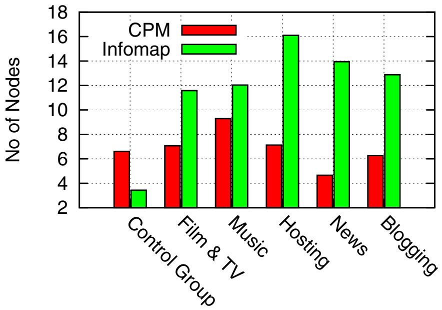

Figure 3: Average Size of Communities

4. SPECIALIZING/DEEPENING INTERESTS

We show that users sharing a specialized interest (i.e. Country Music) form a more tightly-coupled community than users sharing a general interest (i.e. Music). The control group is the users interested in the general Music category as discussed in Section 3. The celebrities representing the Country Music category are seven Country Music singers who have won various awards at the Country Music Awards between 2001 to 2008 and have more than 10,000 followers.

We investigate the changes in community formation as users specialize in their common interest (from Music to Country Music). The results are normalized by the number of users in each respective group to give an accurate representation of the community characteristics of each interest group. This Normalized Average Community Size (NACS) allows us to compare if users with specialized interests form larger communities than users with a general interest, without the biases of the base population size. We observe that the NACS of the $Int_{Country} = 6$ group is 23 and 28 times larger than the $Int_{Music} = 6$ group using CPM and Infomap respectively as shown in Table 2. In addition, users with a lower level of interest in a specialized category are also more likely to form larger communities on average compared to users with a higher level of interest in a general category.

Table 2: Comparison of General and Specialized Interest

Statistic	General (Music)	Specialized (Country)
NACS (CPM)	0.00032	0.00750
NACS (Infomap)	0.00041	0.01151
Path Length	2.82	2.10
Clustering Coefficient	0.63	0.76
Diameter	8	4
Avg. Degree	7.29	5.52

Communities comprising users with a specialized interest are also more cohesive and well-connected than those with a general interest. Table 2 best illustrates this where users with a specialized interest in Country Music form communities with a shorter average path length and diameter but higher clustering coefficient compared to those with a general interest in Music.

Next, we investigate the changes in communities as their interest in a category grows deeper, which is indicated by an increasing Int_{cat} value. Specifically, we report on the changes in number of communities, community size, clustering coefficient and path length among users as their interest deepens. The size and number of communities show how likely users with common interests form communities while clustering coefficient and path length give an indication of connectedness within the communities.

An increase in interest level among users corresponds to an increase in their average community size. We observe an increasing NACS with increasing $Int_{Country}$ values. This result supports our original observation that communities are more likely to be formed among like-minded individuals. In addition, the average size and number of communities formed increases as the interest level of the users increases.

Communities comprising users with a common interest get more tightly coupled as their level of interest increases. We observe a gradual increase in clustering coefficient among the largest communities with increasing $Int_{Country}$ values. Similarly, the largest communities at varying values of $Int_{Country}$ have an average path length of 1.7 to 3.0 hops, illustrating that users sharing common interests form communities that are better connected.

Even considering only friendship links for community detection, the communities detected still display the characteristics of scale-free networks. Upon closer examination, we observe that many individuals with large degree distribution are also country music artists but with less fans than the celebrities we have chosen. The fact that there are other minor country singers among these communities shows that our method effectively detects communities comprising users with a common interest.

In conclusion, we proposed a method to efficiently detect communities comprising individuals with common interests for application in target advertising and viral marketing. As Twitter has no explicit options for users to state their interest, we derived a measurement of interest based on the number of celebrities in an interest category that the user follows. Our approach detects communities that are larger, more cohesive and only comprise users that share a common interest. Also, we observed how their community structures become more connected and cohesive with specializing or deepening of interest in a given category.

5. ACKNOWLEDGMENTS

Kwan Hui Lim was supported by the Australian Government, University of Western Australia (UWA) and School of Computer Science and Software Engineering (CSSE) under the International Postgraduate Research Scholarship, Australian Postgraduate Award, UWA CSSE Ad-hoc Top-up Scholarship and UWA Safety Net Top-Up Scholarship.

6. REFERENCES

[1] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proc. of WWW*, pages 591–600, 2010.

[2] D. Li, B. He, Y. Ding, J. Tang, C. Sugimoto, Z. Qin, E. Yan, J. Li, and T. Dong. Community-based topic modeling for social tagging. In *Proc. of CIKM*, pages 1565–1568, 2010.

[3] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, 2005.

[4] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *PNAS*, 105(4):1118–1123,2008.

[5] S. H. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, and H. Zha. Like like alike - Joint friendship and interest propagation in social networks. In *Proc. of WWW*, pages 537–546, 2011.