Information Diffusion Mechanisms in Online Social Networks

Shushen Fu
School of Computer and Communication Engineering
University of Science and Technology Beijing
Beijing, China
Email:fushushen@sina.com

Chungjin Hu
School of Computer and Communication Engineering
University of Science and Technology Beijing
Beijing, China
Email:hucj0313345@163.com

Ying Hu
School of Computer and Communication Engineering
University of Science and Technology Beijing
Beijing, China
Email:yinghu@xs.ustb.edu.cn

Bo Sun
School of Computer and Communication Engineering
University of Science and Technology Beijing
Beijing, China
Email:s654241016@sina.com

Wenrui Ying
School of Computer and Communication Engineering
University of Science and Technology Beijing
Beijing, China
Email:yingwenrui@163.com

Peng Shi
National Center for Materials Service Safety
University of Science and Technology Beijing
Email:shipengustb@sina.com

***Abstract*—A great deal of research interest has arisen in the area of information diffusion since the emergence of online social networks (OSN). Many models for describing how information diffuses in OSN have been proposed. But so far, the mechanisms of information diffusion remain largely unexplored. We try to find the mechanisms from the perspective of both network structure (micro-scale level) and popularity of information (macro-scale level). We focus on how these two levels interact with each other impacting information diffusion.**

Based on a Twitter data set which contains 196 million tweets, 10 million users, and 3 million hashtags, we perform a temporal analysis by calculating the time-evolving properties of network structure, including node degree, global clustering coefficient, the number of the nodes in the largest cluster. We then investigate how popularity evolves over time as the network structural properties change. We finally find that the moment when hashtag enter into bursting period is relevant to the largest cluster of users discussing the hashtag. When largest cluster changes from a closely connected small cluster to an emanative large cluster, hashtag enters into bursting period. We find this moment by calculating global clustering coefficient and the number of the nodes in the largest cluster. Our study on the interactive mechanisms between micro- and macro-scale levels reveals the essence of information diffusion and provides a necessary theoretical basis for predicting popularity or public sentiment warning.

***Keywords*—Information Diffusion Mechanisms; Information Diffusion Model; Social Networks Structure; Global Clustering Coefficient**

I. Introduction

Because OSN has become more and more popular, more and more researchers pay attention to information diffusion in OSN. With the results of this research, government can understand public psychology and do network public sentiment warning. Advertising agency can adverse more effectively.

As a representative of OSN, Twitter has accumulated more and more users since built in March, 2006. Hashtag is a kind of label to distinguish topics in Twitter. Tweets with the same hashtag are about the same topic. The diffusion of hashtag is the diffusion of a topic. It is a typical kind of information diffusion in OSN. So our work on the diffusion of hashtag is also a research on information diffusion in OSN.

The diffusion of hashtag is inseparable from retweeting by users. Its macro-scale level is the growth of total popularity which can be represented by the number of users who discuss about the hashtag. The trend of the popularity changes is a part of macro-scale level, too. And its micro-scale level is users tweeting or retweeting which make the hashtag spread more widely. This paper is to find interactive mechanisms between micro- and macro-scale levels.

In this paper, we researched on both micro- and macro-scale levels by analyzing a real data set of the diffusions of many hashtags. For any hashtag, we not only researched micro- and macro-scale levels at a given time, but also researched them in the whole period of the diffusion. The reason of this scheme is that we think micro- and macro-scale levels of a given hashtag at a given time are different from them at an earlier time or a later time. Researching micro- and macro-scale levels during the whole diffusion period can help us to find the different characteristics of them in different times.

Our work in this paper can be summarized as follows: we worked hard to find interactive mechanisms between micro-

978-1-5090-1192-6/16 $31.00 © 2016 IEEE
DOI 10.1109/DSC.2016.46

and macro-scale levels by observing a large amount of real data. We proposed a reasonable hypothesis according to our observation. Then we proved our hypothesis by making many experiments on the data set.

The above is the main contribution of this paper. We review related work in Section 2. We explain the data set and propose our hypothesis in Section 3. We establish our hypothesis in Section 4. We reiterate our conclusion and propose future work in Section 5.

II. BACKGROUND AND RELATED WORK

In recent years, macro-scale level of information diffusion has been done a detailed study by a lot of researchers. Popularity has been studied by many researchers.

Study on the popularity of information diffusion in OSN started at Wu and Huberman s work[1]. They proved Digg posts follow log normal distribution. Soon after, Cha et al. [2] proved the popularity of YouTube video follows a power law distribution. Later, other kinds of information in OSN have also been proved that they also follow these two distributions. Such as Tweets in Twitter[3], answers in Essembly[4], article in Wikipedia[5].

Many researchers work hard to find why the distribution of popularity follow these distributions. Studies have shown that power law distribution is formed in preferential attachment process[6]. This process is also known as the rich-get-richer process or Yule process[7]. On the other hand, Newman mentioned that the preferential attachment process of information in OSN can be explained as follows: the popularity which a video or a post will get is positive proportional to the popularity it has gotten[8].

Just like [9] discussed, the potential mechanisms leading to these two distributions are very similar. So sometimes it is hard to tell the popularity follow a power law distribution or log normal distribution.

The popularity of different information is different. The popularity of the same information at different time is also different. The popularity is always changing with the process of information diffusion. This is called as popularity evolution.

There are many ways of popularity evolution. A comprehensive study of the entire period of diffusion of every information is difficult. Thus, part of researchers do research on how popularity grow and fade and the speed of these changes. They used peak [10,11,12] and trends [12,13,14] to capture the features of popularity on the sequence.

Another part of researchers are concerned that the evolution model of popularity. Some of them studied on videos in OSN. Crane and Sornette [11,15] make YouTube videos divided into three categories according to the ratio of the popularity at the peak time to the total popularity during whole period. But these videos can not spread widely in OSN. Figueriedo [12] found another evolution mode of popularity. Its popularity maintain sustained growth in a long period of time. Figueriedo et al. [16] also studied the evolution of three different types of videos. They are top list video, copyrighted video and video which is selected randomly by keywords. Their results suggest that top video will have a sudden sharp rise to form a peak.

There is another part of the people are concerned about the evolution model of popularity of topics in OSN. Lin et al. [17] proposed a statistical method of modeling the sequential evolution of topic, considering users interests and network structure. Romero et al. [18] proposed a generating model based on simulation. They found that topic become popular or not depends not only on whether it was exposed, but also on the persistence of this topic. Lehmann et al. [19] also observed hashtag to study the peak of popularity evolution of topics. They found that the peak of popularity is mainly caused by external factors, such as the involvement of media. Ardon et al. [20] made a strict sequential, spatial analysis on topics. They found that those users having many followers have a larger impact on the popularity of the topic. Hu et al. [14] studied on the popularity evolution model of hot topics in Tianya. They proposed a method to analyze popularity evolution model based on time series considering trend, period and average value. They found that non-bursting topics trend and average value change more stability than bursting topic.

For the micro-scale level of information diffusion, there are many researchers have done experiments, too. Their study is detailed.

Cascade structure [21,22,23,24] is a kind of micro performance of information diffusion. Kupavskii et al. [25] studied on the retweet cascade in Twitter. They expressed the cascade by directed graph. Wang et al. [29,30] made a deep research on cascades and information diffusion.

There are some fundamental models in this category, namely Independent Cascades Model (IC) [22,23] and Linear Threshold Model (LT) [26]. Guille A and Hacid H [27] proposed the Susceptible-Infected-Removal (SIR) model.

There are scholars studying on the relationship between micro- and macro-scale levels of information diffusion. They focus on predicting popularity according to micro-scale levels.

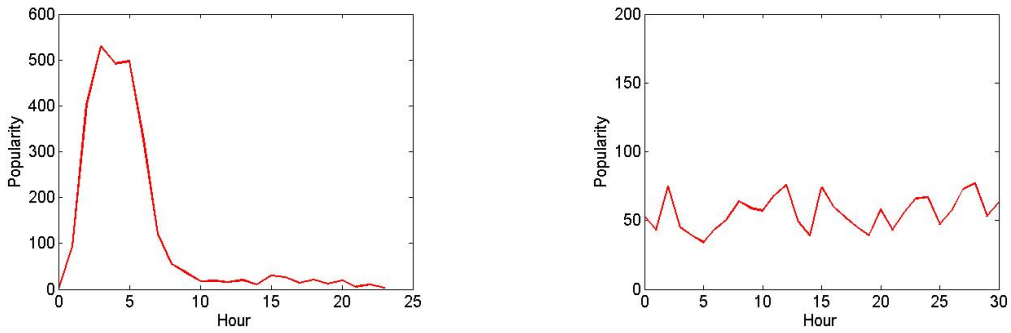
Kupavskii et al. [25] model the growth process of retweets by infectious diseases model. Li et al [28] studied the popularity of the videos which were shared from video sites to renren site by users of renren site. And they proposed Social network assisted Video Prediction (SoVP) method which is based on the cascade structure.

III. HYPOTHESIS

In this section, we describe the data set used in the experiments and describe the observations we found from data set. At the end of this section, we propose our hypothesis.

A. Data Set Description

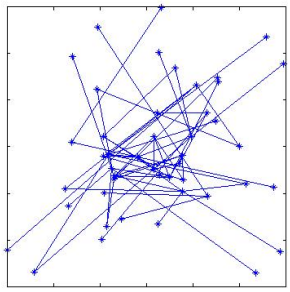
We choose our dataset as the tweet7 dataset which is collected by Yang and Lescovec[10]. Many research work has



(a) Bursting hashtag: #BringBackRachelle# (b) Non-bursting hashtag: #book#

Fig. 1. The popularity evolution of a bursting hashtag and a non-bursting hashtag.

(a) Hour 1 popularity:29 (b) Hour 2 popularity:63



(c) Hour3 popularity:158

Fig. 2. Evolution graph of the max cluster. #clubrules#

been done by using this dataset. The dataset collected 196 million tweets and 3 million hashtag from 10 million user from June 2009 to December 2009. We do some pre-filtering on the dataset and find most of the hashtags only have few users involved, which means there is limited value on conducting research on those hashtags. We then filter these hashtags. As showed in TABLE 1, there are 2594 hashtags having more than 1000 users discussed with it. Popularity of an hour is the number of users discussed with it in this hour.

Hashtag has its own life period. Considering about the life period of hashtags, we use one hour as time basis and make a popularity evolution graph as Fig 1. This graph shows popularity's change in its whole life period.

By observing the diffusion process of hashtags, we found some hashtags has explicit bursting growth (as showed in Fig 1.a). If peak hour get more than 20% of all popularity the hashtag get in whole life period, we call it as bursting hashtags. Contrary to these hashtags, other hashtags dont have explicit bursting growth (as showed in Fig 1.b). They are steady during the whole period. These can be seen from their popularity evolution line. We call them non-bursting hashtags. The period when the hashtag get more than 10% of its all popularity per

TABLE I
THE DISTRIBUTION OF POPULARITY

Total Popularity	10 - 100	100 - 1000	more than 1000
Count	35,694	10,492	2,594

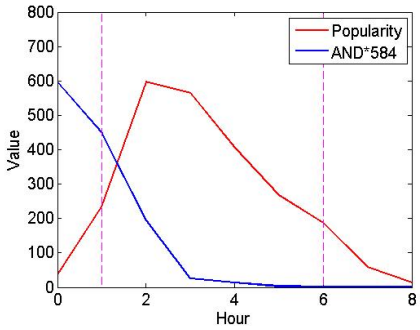


Fig. 3. Average node degree evolves over time. #bestHead#

hour is called as bursting period.

Since our research interest is more about what is the differences of interactive mechanism between micro- and macro-scale levels in different periods, the non-bursting hashtags are not very important to our interest. Therefore, most of our observation focus on the 700 bursting hashtags.

B. Network structure evolving graph

Hashtags' popularity is the macro-scale level of its diffusion. Diffusion's micro-scale level is network structures evolution. We use network structure evolving graph to describe network structure. It is undirected graph. We make an evolving graph for every hashtag per hour.

If a user tweet about a hashtag at T hour, we add this user into nodes of the hashtag's evolving graph of T hour. If his friends have tweeted about this hashtag before T hour, we add edges between this user and all of these friends into edges of the hashtag's evolving graph of T hour. We can not determine which friend influence the user, so we assume that the user tweet about the hashtag because of the influence of all his friends.

There is a maximal connected subgraph in evolving graph of each hour. It is the largest cluster of users, too. We pay more attention to this cluster. We observe this clusters evolution of all these 700 hashtags. We find that more than 71% percent of these hashtags have a similar evolution process as shown in Fig 2. The largest cluster is a closely connected small cluster at Hour 1 as shown in Fig 2a. And the largest cluster is change to an emanative large cluster at Hour 2 and Hour 3 as shown in Fig 2b and Fig 2c. Its popularity is 29 at Hour 1. But its popularity is 158 at Hour 3.

C. Proposed hypothesis

According to this phenomenon, we propose our hypothesis: the moment when hashtag enters into bursting period is relevant to the largest cluster of users discussing the hashtag.

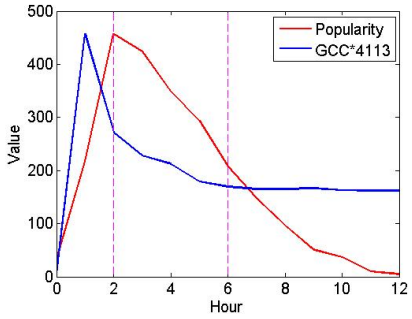


Fig. 4. Global clustering coefficient evolves over time. #alesanputus#

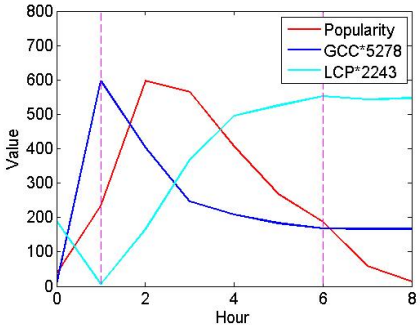


Fig. 5. Largest cluster proportion,gcc and popularity. #bestlovesong#

When largest cluster changes from a closely connected small cluster to an emanative large cluster, hashtag enters into bursting period.

IV. HYPOTHESIS ESTABLISHMENT

In this section, we conduct experiments to establish our hypothesis. We calculate the time-evolving properties of network structure and investigate how popularity evolves over time as the network structural properties change.

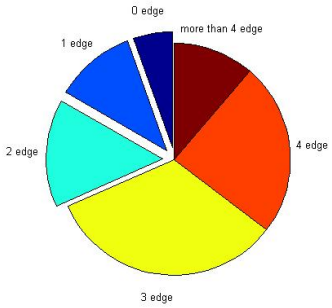
A. Average node degree

Degree of a node is the number of edges connected to this node. Average node degree (AND) is the mean of the number of edges connected to nodes in network structure. We calculate AND as follows.

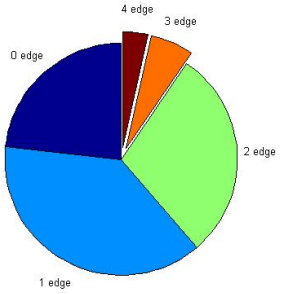
$$AND = \frac{N_{edge}}{N_{node}} \tag{1}$$

where N_{edge} is the number of edges in whole network structure. N_{node} is the number of nodes in whole network structure.

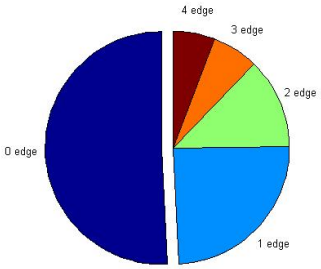
We find that 94% of hashtags' AND evolves over time similar to the mode shown in Fig 3. To make figure more intuitive, we enlarge AND as shown in legend of Fig 3. The period between two dotted lines is bursting period. AND maintain in a high level before bursting period and fall quickly in bursting period. This experiment shows that new nodes have



(a) Before bursting period



(b) During bursting period



(c) After bursting period

Fig. 6. Edges distribution beween the new node and the largest cluster. #Brooklyn#

less edges connected to old nodes when information enters into a bursting period.

B. Global clustering coefficient

Global clustering coefficient (GCC) describes tightness of whole network structure. A high value of GCC means a closely connected network structure. We calculate global clustering coefficient as follows.

$$GCC = \frac{Number\ of\ closed\ triplets}{Number\ of\ connected\ triplets} \tag{2}$$

71% of hashtags GCC evolves over time similar to the mode shown in Fig 4. We enlarge GCC because of the same reason as AND. The period between two dotted lines is bursting period. GCC arise before bursting period and fall in bursting period. It maintains in a low level after bursting period. This experiment shows that network become more tightly before bursting period. And network structure become more emanative in bursting period.

C. Largest cluster proportion

We call the ratio of the number of nodes in largest cluster to the number of all nodes in network structure as largest cluster proportion (LCP). It can describe the importance of the largest cluster in whole network structure. A high LCP means largest cluster plays an important role in network structure. We calculate LCP as follows.

$$LCP = \frac{N_{largest\ cluster}}{N_{whole}} \tag{3}$$

where $N_{largest\ cluster}$ is the number of nodes in largest cluster. N_{whole} is the number of all nodes in whole network structure.

83% of hashtag's LCP evolves over time similar to the mode shown in Fig 5. We enlarge GCC and LCP because of the same reason as AND. The period between two dotted lines is bursting period. LCP arises in bursting period. It means largest cluster plays a more important role in bursting period. According to GCC's falling, Fig 5 shows that the network structure become more emanative because of largest cluster. Largest cluster changing from a closely connected small cluster to an emanative large cluster make GCC fall and make information enter into bursting period. This moment is the moment when GCC begins to fall and LCP begins to arise.

D. Connect beween the new node and the largest cluster

We calculate the number of edges between new nodes and largest cluster in different periods. As shown in Fig 6, most of new nodes have more than two edges connected to largest cluster before bursting period. In bursting period, most of new nodes have one or two edges connected to largest

cluster. These new nodes make largest cluster become a more emanative and larger cluster. After busting period, most of new nodes haven't any edge connected to largest cluster. 74% of hashtags have this evolution mode as shown in Fig 6. This experiment establish our hypothesis from a more intuitive perspective.

In summary, experiments show that most of hashtags are consistent with our hypothesis. According to these results of experiments, our hypothesis is established. Interactive mechanisms between micro- and macro-scale levels can be described as follows. The popularity of a hashtag enters into a bursting period when the largest cluster of users discussing the hashtag changes from a closely connected small cluster to an emanative large cluster.

V. Conclusion

In this paper, we perform a temporal analysis by calculating the time-evolving properties of network structure on data set. We find that the moment when hashtag enter into bursting period is relevant to the largest cluster of users discussing the hashtag. When largest cluster changes from a closely connected small cluster to an emanative large cluster, hashtag enters into bursting period.

However, we find some outliers based on our experiments. In our future work, we will examine those outliers and see what factors affect our hypothesis. We think Hashtag is possibly related to the content of hashtags. So we will focus on the content of hashtag and the content of tweets in our future work.

Acknowledgment

The authors would like to thank Jure Leskovec and Sebastien Ardon for providing us Twitter data set.This work was supported by the National Grand Fundamental Research 973 Program of China under Grant No. 2013CB329605.

References

[1] F. Wu, BA. Huberman. Novelty and collective attention. Proceedings of the National Academy of Sciences, 2007, 10445 17599-17601.

[2] M. Cha, H. Kwak, P.Rodriguez, et al. I tube, you tube, everybody tubes analyzing the world's largest user generated content video system. Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, 2007 1-14

[3] K. Lerman, R. Ghosh. Information contagion an empirical study of the spread of news on Digg and Twitter social networks. ICWSM, 2010, 10 90-97.

[4] T. Hogg, G. Szabo. Diversity of user activity and content quality in online communities. ICWSM. 2009, 58-65.

[5] DM. Wilkinson. Strong regularities in online peer production. Proceedings of the 9th ACM Conference on Electronic Commerce, 2008, 302-309.

[6] A. L. Barabasi, R. Albert. Emergence of scaling in random networks. Science, 1999, 2865439 509-512

[7] GU. Yule. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character, 1925, 21-87.

[8] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 2005, 465 323-351.

[9] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. Internet Mathematics, 2004, 12 226-251.

[10] J. Yang, J. Leskovec. Patterns of temporal variation in online media. Proceedings of the 4th ACM International Conference on Web Search and Data Mining, 2011, 177-186.

[11] R. Crane, D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. Proceedings of the National Academy of Sciences, 2008, 10541 15649-15653.

[12] F. Figueiredo. On the prediction of popularity of trends and hits for user generated videos. Proceedings of the 6th ACM International Conference on Web Search and Data Mining, 2013, 741-746.

[13] S. Asur, B. A. Huberman, G. Szabo, et al. Trends in social media persistence and decay. Available at SSRN 1755748, 2011.

[14] C. Hu, Y. Hu, W. Xu, et al. Understanding popularity evolution patterns of hot topics based on time series features. Web Technologies and Applications, 2014, 58-68.

[15] R. Crane, D. Sornette. Viral, Quality, Junk. Videos on YouTube separating content from noise in an information-rich environment. AAAI Spring Symposium Social Information Processing, 2008, 18-20.

[16] F. Figueiredo, F. Benevenuto, JM. Almeida. The tube over time characterizing popularity growth of YouTube videos. Proceedings of the 4th ACM International Conference on Web Search and Data Mining, 2011, 745-754.

[17] C. Lin, B. Zhao, Q. Mei, et al. PET a statistical model for popular events tracking in social communities. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, 929-938.

[18] DM. Romero, B. Meeder, J. Kleinberg. Differences in the mechanics of information diffusion across topics idioms, political hashtags, and complex contagion on twitter. Proceedings of the 20th International Conference on World Wide Web, 2011, 695-704.

[19] J. Lehmann, B. Gonalves, J. J. Ramasco, et al. Dynamical classes of collective attention in twitter. Proceedings of the 21st International Conference on World Wide Web, 2012, 251-260.

[20] S. Ardon, A. Bagchi, A. Mahanti, et al. Spatio-temporal and events based analysis of topic popularity in twitter. Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, 2013, 219-228.

[21] K. Saito, R. Nakano, M. Kimura. Prediction of information diffusion probabilities for independent cascade model. Knowledge-Based Intelligent Information and Engineering Systems, 2008, 67-75.

[22] J. Goldenberg, B. Libai, E. Muller. Talk of the network A complex systems look at the underlying process of word-of-mouth. Marketing letters, 2001, 123 211-223.

[23] J. Goldenberg, B. Libai, E. Muller. Using complex systems analysis to advance marketing theory development Modeling heterogeneity effects on new product growth through stochastic cellular automata. Academy of Marketing Science Review, 2001, 93 1-18.

[24] FD. Sahneh, C. Scoglio, P. Van Mieghem. Generalized epidemic mean-field model for spreading processes over multilayer complex networks. IEEE/ACM Transactions on Networking, 2013, 215 1609-1620.

[25] A. Kupavskii, L. Ostroumova, A. Umnov, et al. Prediction of retweet cascade size over time. Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012, 2335-2338.

[26] M. Granovetter. Threshold models of collective behavior[J]. American journal of sociolo-gy, 1978, 83(6): 1420.

[27] A. Guille, H. Hacid. A predictive model for the temporal dynamics of information diffusion in online social networks[C]//Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012: 1145-1152.

[28] H. Li, X. Ma, F. Wang, et al. On popularity prediction of videos shared in online social networks. Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, 2013, 169-178.

[29] S. Wang, Z. Yan, X. Hu, P. S. Yu, Z. Li: Burst Time Prediction in Cascades. AAAI 2015: 325-331

[30] S. Wang, X. Hu, P. S. Yu, Z. Li: MMRate: inferring multi-aspect diffusion networks with multi-pattern cascades. KDD 2014: 1246-1255