Computational Experiments Based on Competitive Influence Diffusion Model

Kainan Cui
The School of Electronic and Information Engineering
Xi'an Jiaotong University
Xi'an, China
kainan.cui@live.cn

Xiaolong Zheng
The State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences
Beijing, China
xiaolong.zheng@ia.ac.cn

***Abstract*—Understanding the strategies to optimize/suppress information spreads under intense competition could provide important insights in a broad range of settings including viral marketing, emergency response and information system design. However, most of existing studies about competitive influence diffusion mainly focus on two-information competition mechanism. To date, the competitive influence maximization problem considering the mechanism of multi-information competition is still not well studied. In this paper, we conducted computational experiments to study the competitive influence maximization with multi-information competition mechanism. By applying an information diffusion model called limited attention model (LAM), we carried on two computational experiments to validate the model and investigate the relation between seed selection methods and the properties of information cascades. Our experimental results show that 1) the LAM model could reproduce the features of empirical distribution in Chinese social media; 2) the eigenvector centrality-based heuristic is a reasonable seed selection method for competitive influence maximization problem. The results of this paper can provide significant potential implications for information system design and management.**

Keywords—computational experiments; information diffusion; competitive influence maximization;social media

I. Introduction

With the advent of social media, the cost of information propagation has been reduced significantly, changing the traditional paradigm of information production, propagation and consumption fundamentally [1]. The boundary between information producers and information consumers has been blurred. While this transition boosts the potential influence of each messages, the abundance of information to which we are exposed increased the competition among information [2]. The online behavior is becoming more frequent and complex.

Competitive influence maximization problem (CIM problem) is one of the crucial research problems in the area of information diffusion [3, 4], which has received more and more attention in recent year. Understanding the strategies to optimize/suppress information spreads under intense competition could provide important insights in a broad range of settings, from viral marketing, emergency response [5] to information system management. Although a lot of work has been done to study how to intervene the information diffusion process, most of the existing works about competitive influence mainly focus on the two-information competition mechanism [6,7] and rely on a fundamental assumption that the users could only participate one piece of information diffusion, which is obviously far from the real situation. Furthermore the models used in these studies have seldom been verified through empirical data. The CIM problem considering the mechanisms of multi-information competition is largely unexplored and lack of understanding.

In order to fill the important research gap, we conducted computational experiments [8,9,10] to study the CIM problem with multi-information competition mechanisms. In particular we validated the limited attention model (LAM) using the Chinese social media data. Based on the LAM model, we further investigated the relation between different seed selection methods and properties of information cascades. The results of this paper can provide significant potential implications for information system design and management.

The rest of this paper is organized as follows: Section II introduces the information diffusion model and the parameter estimation procedure used in this study. Section III presents the two computational experiments designed to carry on in this paper. Section IV shows the results of two computational experiments described in Section III and provides some discussions. Section V draws conclusions with remarks on future works and directions.

II. Information Diffusion Model

Aiming to study how information spreads in the social media, we need a model able to create information cascades with properties similar to those find in real environments. In this section, we first introduce an information diffusion considering the multi-information competition mechanism called limited attention model. Then the parameter estimation procedure used in this study is presented.

A. Limited Attention Model

Limited attention model (LAM) is proposed by L. Weng et al [2], which is an agent-based probabilistic model. The model is based on the scenario of micro-blogging platform, which allows massive users to post short messages through social connections. Users can "follow" interesting people, by which a directed social network is formed. Messages ("tweets") appear on the screen of followers. People can forward ("retweet") selected messages from their screen to their followers.

One of the key features of the LAM model is that finite list were used to model the users' attention. Agents interact on a directed social network of friends/followers. Each agent owns two important lists called memory and screen, which stored the messages posted by the agents and the messages posted by the agents' followers respectively. Messages could survive in the agents' list only for a finite amount of time. The main procedure of the LAM model is listed as follows:

- At each step, an agent is selected randomly to post messages to neighbors.
- The agent may post about a new message with probability P_n.
- Otherwise, the agent reads messages from the screen. Each message may attract the user's attention with probability P_r.
- Then the agent either retweets the messages with probability $1-P_m$, or post about a message chosen from memory with probability P_m.
- The screen list and memory list work like first-in-first-out and will be cleared periodically.

B. Parameter Estimation

There are three parameters needed to be estimated: P_n regulates the amount of new messages entering the system (number of cascades), P_r determines the overall retweet activity (size of cascades), and P_m accounts for individual focus (diversity of user attention).

We selected a public-available dataset of Chinese micro-blogging [11] to estimate the parameters of the LAM model. The empirical data contains 1.7 million users and 0.4 billion following relationships among them, with average 200 followees per user. For each user, The 1,000 most recent microblogs (including tweets and retweets) were saved. Among those messages (in totally 1 billion microblogs), 232978 popular diffusion episodes was sampled. Each diffusion episode contains the original messages and all its retweets. The basic data statistics of the dataset was showed in table 1.

TABLE I. EMPIRICAL DATA STATISTICS

Dataset	***#Users***	***#Follow-relationships***	***#Original messages***	***#Retweets***
Weibo	1776950	308489739	232978	3307189

The parameter P_n characterizes the probability of sending a new original message. To estimate this parameter from the empirical data, we calculate the ratio of original messages to all messages as shown in (1).

$$P_n = \frac{original\_post\_number}{total\_post\_number} \tag{1}$$

According to the empirical data, the total number of messages is 33540167. Among those messages 232978 are original messages. The proportion of original messages is approximately 00694623. We thus set P_n. = 0.069 for all the experiments.

The parameter P_m represents the proportion of messages tweeted by an individual more than once. To estimate it from the empirical data, we compute the ratio of messages that have been retweeted multiple times by one user, which we called memory post as shown in (2). Using the average value across all users, which is 0.06248. We set $P_m = 0.062$.

$$P_m = \frac{memory\_post\_number}{total\_post\_number} \tag{2}$$

Finally, the parameter P_r is tuned to capture the average number of posted messages per user. According to the empirical data the average number of posted messages per user is 25.014743. We set P_r =0.01785. Under this setting the average number of posted messages per user of the LAM model is 26.43, which is closed to the empirical result.

In sum, the parameters of LAM model and corresponding value are shown in table 2.

TABLE II. A SUMMARY OF PARAMETERS OF LAM MODEL

Parameter	***Meaning***	***Value***
P_n	Probability of sending a new message	0.069
P_m	Probability of chosing a message from memory	0.062
P_r	Probability of replying each message in screen	0.01785

III. COMPUTATIONAL EXPERIMENTS

In this section, two computational experiments were designed and conducted based on the LAM model. In the first experiment, we try to reproduce the statistical patterns of empirical information diffusion data by applying LAM model and the parameters estimated in the previous section. Then in the second experiment, we try to investigate and compare the performances of different network structure based heuristics in competitive influence maximization problem.

Due to the size of the empirical follower network, we sampled a manageable subset for our experiments. The sampling procedure was a random walk with occasional restarts from random locations (teleportation factor 0.15). The sampled network has 100000 nodes and about 127418 edges.

A. *Model Validation Experiment*

Aiming to obtain comparable results with the real data, we have selected a set of measures that describe important statistical features of both real diffusion episodes and user behavior. We compare experiment results with real data to verify the LAM model. We have considered two characteristics of the information cascades, 1) the message lifetime defined as the maximum number of consecutive time units in which posts about the messages are observed and 2) the message popularity, defined as the number of users per day who tweet about a messages.

For user behavior, we considered the user activity and user attention. User activity was defined as the number of messages per day and User attention was quantified though Shannon entropy as shown in (3)

$$S = -\sum_i f(i) \log f(i) \qquad (3)$$

Where *f(i)* is the proportion of tweets generated by the user about message *i*. The experiments will run 50 times and all of the quantities will be averaged.

B. *Seed Node Selection Experiment*

In order to explore the performance of different seed node selection methods, we also selected a set of measures that describe some important properties of information cascades. We have considered three characteristics of the information cascades, 1) the user attendance, defined as the number of users retweeted the information during the diffusion process, 2) the message popularity, defined as the number of the retweets of the messages and 3) the message lifetime, defined as the maximum number of consecutive time units in which posts about the messages are observed. Since the user could retweet messages for multiple times, the user attendance number is different from the message popularity.

We choose several widely-used structural measures of influence in social networks as seed node selection method. In particular, we consider five different heuristics, which are listed as follows:

- Degree centrality
- Eigenvector centrality
- Closeness centrality
- Betweeness centrality
- Random selection.

In this experiment, we activate five seed nodes to post one specific message at the beginning of model running. The top five nodes and corresponding values of these quantities are presented in table 3. For each set of seeds, we ran 50 independent simulations. The characteristics of this piece of information will be saved and averaged, including the message lifetime, message popularity and the user attendance of the information cascades.

TABLE III. TOP FIVE NODES OF DIFFERENT SEED SELECTION METHODS

Seed node selection methods		***Rank 1***	***Rank 2***	***Rank 3***	***Rank 4***	***Rank 5***
Degree centrality	Node ID	10767	57700	84581	21858	96672
	Value	159	158	153	139	138
Eigenvector centrality	Node ID	26961	85291	98100	98096	9309
	Value	1	0.968	0.94	0.924	0.915
Closeness centrality	Node ID	35962	74239	38981	17460	73930
	Value	2.14E-5	2.14E-5	2.14E-5	2.14E-5	2.14E-5
Betweeness centrality	Node ID	10767	86896	84581	21858	57700
	Value	2.58E+08	2.33E+08	1.85E+08	1.74E+08	1.63E+08
Random selection	Node ID	96438	2139	1460	19733	1734
	Value	N/A	N/A	N/A	N/A	N/A

IV. RESULTS AND DISCUSSIONS

The results of the model validation experiment are illustrated in Fig. 1. All of the empirical data point to extremely heterogeneous behaviors. Some messages are extremely successful (popular and persistent), while the great majority die quickly. A small fraction of messages therefore account for the great majority of all posts. Likewise, a small fraction of users account for most of the traffic. The present findings demonstrate that the LAM model captures the heterogeneity of the empirical distributions of message lifetime, message popularity, user activity, and user attention which indicated the consistency between the simulations and the reality. Therefore, the feasibility of applying the LAM model is verified.

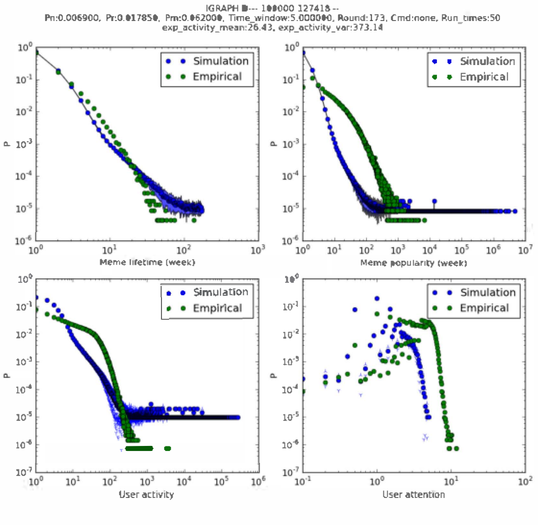

Fig. 1. Evaluation of model by comparison of simulations (blue circles) with empirical data (green circles) (a) Probability distribution of the lifetime of messages (b) Probability distribution of the popularity of messages. c) Probability distribution of user activity, measured by the number of messages. (d) Probability distribution of the user attention (entropy), based on the message tweeted by a user.

Table 4 shows the performance of different seed selection methods in the LAM model. From the table we could conclude that the eigenvector centrality-based heuristic outperforms other heuristics obviously.

The statistical test results support the intuition that the eigenvector centrality-based heuristic outperforms other heuristics as shown in table 5. For all of the measures, there are significant difference between the diffusion results of eigenvector centrality-based heuristic and other heuristics. The result shows that by using eigenvector centrality as the structural properties to identify influential node could obtain significantly better diffusion results which means the eigenvector centrality-based heuristic is a reasonable seed node selection methods in CIM problem based on LAM model.

TABLE IV. COMPARISON OF PERFORMANCE OF DIFFERENT HEURISTICS

Heuristic	***Message popularity***	***Message life time***	***User attendance***
Betweeness centrality	6153.70	98.58	1705.28
Closeness centrality	8.00	3.94	5.36
Degree centrality	11903.98	118.74	2157.68
Eigenvector centrality	925046.22	165.5	10623.26
Random selection	16.22	5.78	12.16

TABLE V. STATISTICAL TEST RESULTS OF PERFORMANCE DIFFERENCES BETWEEN EIGENVECTOR CENTRALITY-BASED HEURISTIC AND OTHER HEURISTICS

	Performance measures	***Statistic (Wilcoxon rank sum test)***
Eigenvector centrality vs. Degree centrality	Message popularity	W = 7127, p-value = 2.024e-07*
	Message life time	W = 6940, p-value = 9.534e-07*
	User attendance	W = 6497, p-value = 0.0002544*
Eigenvector centrality vs. Closeness centrality	Message popularity	W = 9948.5, p-value < 2.2e-16*
	Message life time	W = 9537.5, p-value < 2.2e-16*
	User attendance	W = 9995, p-value < 2.2e-16*
Eigenvector centrality vs. Betweeness centrality	Message popularity	W = 7213.5, p-value = 6.357e-08*
	Message life time	W = 7296.5, p-value = 9.294e-09*
	User attendance	W = 6531, p-value = 0.0001834*
Eigenvector centrality vs. Random selection	Message popularity	W = 9868, p-value < 2.2e-16*
	Message life time	W = 9478.5, p-value < 2.2e-16*
	User attendance	W = 9881.5, p-value < 2.2e-16*

* means statistical significant (p=0.05)

Fig. 2 shows the information diffusion results of eigenvector centrality-based heuristic in LAM model. As shown in Fig.2 (a) the mean value of message popularity increased gradually during the diffusion process. At the same time, the variation of the message popularity also increased significantly, which emphasize the random nature of diffusion results under intense competition. As shown in Fig.2 (b), the user attendance displays long-tailed distributions. The distribution also demonstrates the obvious heterogeneity of diffusion results with the same set of seeds.

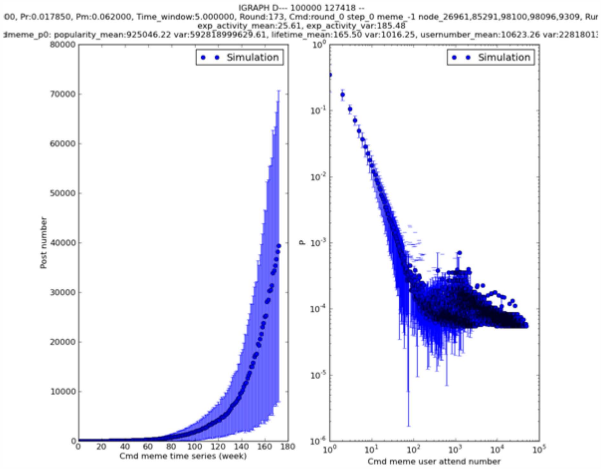

Fig. 2. Diffusion results of eigenvector centrality-based heuristic in LAM model (a) Temporal patterns of message popularity (b) Probability distribution of the user attendance

V. CONCLUSION

The development of information diffusion theory and availability of data from online social media provides us an opportunity to investigate the competitive influence maximization problem from new perspectives. In this paper, we have presented our works and results of applying computational experiments to analyze the CIM problem considering the multi-information competition mechanism. Two computational experiments have been carried out based on the limited attention model. The experimental results show that 1) the LAM model could reproduce the features of empirical distribution in Chinese social media; 2) the eigenvector centrality-based heuristic is a reasonable seed selection method for competitive influence maximization problem. However the intense competition among information increases the random nature of the diffusion results.

ACKNOWLEDGMENT

This work was supported in part by the following grants: The National Natural Science Foundation of China, Grant No. 71103180, 91124001, 71025001,and 91024030, by the Early Career Development Award of SKLMCCS and by the Ministry of Health under Grant No. 2012ZX10004801.

REFERENCES

[1] D. Zeng, H. Chen, L. R., and S.-H. Li, "Social Media Analytics and Intelligence," Intell. Syst. IEEE, vol. 25, pp. 13–16, 2010.

[2] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, "Competition among memes in a world with limited attention," Sci. Rep., vol. 2, Mar. 2012.

[3] C. Budak, D. Agrawal, and A. El Abbadi, "Limiting the Spread of Misinformation in Social Networks," in Proceedings of the 20th

International Conference on World Wide Web, New York, NY, USA, 2011, pp. 665–674.

[4] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the Spread of Influence Through a Social Network,” in Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2003, pp. 137–146.

[5] K. Cui, X. Zheng, D. D. Zeng, Z. Zhang, C. Luo, and S. He, “An Empirical Study of Information Diffusion in Micro-blogging Systems during Emergency Events,” in Web-Age Information Management, Y. Gao, K. Shim, Z. Ding, P. Jin, Z. Ren, Y. Xiao, A. Liu, and S. Qiao, Eds. Springer Berlin Heidelberg, 2013, pp. 140–151.

[6] S. Bharathi, D. Kempe, and M. Salek, “Competitive Influence Maximization in Social Networks,” in Internet and Network Economics, X. Deng and F. C. Graham, Eds. Springer Berlin Heidelberg, 2007, pp. 306–311.

[7] A. Borodin, Y. Filmus, and J. Oren, “Threshold Models for Competitive Influence in Social Networks,” in Internet and Network Economics, A. Saberi, Ed. Springer Berlin Heidelberg, 2010, pp. 539–550.

[8] F.-Y. Wang, “A Computational Framework for Decision Analysis and Support in ISI: Artificial Societies, Computational Experiments, and Parallel Systems,” in Intelligence and Security Informatics, H. Chen, F.-Y. Wang, C. C. Yang, D. Zeng, M. Chau, and K. Chang, Eds. Springer Berlin Heidelberg, 2006, pp. 183–184.

[9] W. Duan, X. Qiu, Z. Cao, X. Zheng, and K. Cui, “Heterogeneous and Stochastic Agent-Based Models for Analyzing Infectious Diseases' Super Spreaders,” IEEE Intelligent Systems, vol. 28, no. 4, pp. 18–25, 2013.

[10] F. Zhu, D. Wen, and S. Chen, “Computational Traffic Experiments Based on Artificial Transportation Systems: An Application of ACP Approach,” IEEE Trans. Intell. Transp. Syst., vol. 14, no. 1, pp. 189–198, Mar. 2013.

[11] “InfluenceLocality.” [Online]. Available: http://arnetminer.org/Influencelocality. [Accessed: 31-Jul-2014].