Assessing the Quality of Diffusion Models using Real-World Social Network Data

Tsung-Ting Kuo[1]*, San-Chuan Hung[1], Wei-Shih Lin[1], Shou-De Lin[1], Ting-Chun Peng[2], Chia-Chun Shih[2]

[1]Graduate Institute of Networking and Multimedia
National Taiwan University
Taipei, Taiwan
*d97944007@csie.ntu.edu.tw

[2]Innovative DigiTech-Enabled Applications & Services Institute
Institute for Information Industry
Taipei, Taiwan

Abstract—**Recently, there has been growing interest in understanding information cascading phenomenon on popular social networks such as Facebook, Twitter and Plurk. The numerous diffusion events indicate huge governmental and commercial potential. People have proposed several diffusion and cascading models based on certain assumption, but until now we do not know which one is better in predicting information propagation. In this paper, we propose a novel framework that utilizes the micro-blog data to evaluate which model is better under different circumstances. In our framework, we devise two schemes for evaluation: the direct and the indirect schemes. We conduct experiments using three diffusion models on Plurk data. The results show Independent Cascade model outperforms other diffusion models using direct scheme, while Linear Threshold model, Degree, In-Degree and PageRank perform best using indirect scheme. The main contribution is to provide a general evaluation framework for various diffusion models.**

Keywords – social network analysis; diffusion model evaluation; information diffusion

I. INTRODUCTION

Nowadays, social networks such as Facebook, Twitter and Plurk become more and more popular. On these social networks, millions of users spend lot of time every day, billions of information diffuses every hour, and huge amount of people affected by the messages every minute. The phenomenon is known as diffusion, cascade, propagation, influence, infect, or activation based on different context.

Many recent studies have focused on devising models to simulate diffusion behavior and are evaluated using certain characteristics of real diffusion phenomenon [3, 5, 7, 11]. For example, the Independent Cascade and the Linear Threshold models [5] are evaluated using the target-set-size to coverage curve, while the Greedy and the Courtesy models [3] are evaluated using the user-threshold to coverage curve. The performance of the diffusion models cannot be easily compared because the diffusion models do not exploit the same set of characteristics for evaluation (ex. target-set-size to coverage curve and user-threshold to coverage curve). Furthermore, for different data and diffusion information, it is difficult to decide the most fitted diffusion model and corresponding parameters. Although some recent literatures define evaluation metrics for the influence of users [4], a general evaluation mechanism for different diffusion models is still lacking.

In this paper, we propose a novel framework, *EPIC* (*Evaluation for Predicting Information Cascades*), that utilizes the micro-blog data to evaluate which model is better under different circumstances. Given a set of diffusion models, a social network, and the diffusion records, EPIC evaluates the performance for each model using two different schemes: direct (i.e., the performance of diffusion link prediction) and indirect (i.e., the performance of top influential user prediction). The reason to apply indirect scheme is that the top influential users are critical in most propagation applications; the direct scheme does not distinguish such users with the others. EPIC then selects the most suitable model to estimate the diffusions (or edge activations) which will occur in the next time point. For the indirect scheme, we further design two evaluation flows: one-by-one and leave-one-out. To summarize, using the EPIC framework, we can evaluate different models for deciding the most appropriate model to estimate the diffusion phenomenon on a social network with specific diffusing information.

To demonstrate the effectiveness of the EPIC framework, we conducted experiments to compare three models, Independent Cascade (IC) [5], Linear Threshold (LT) [5], and Susceptible-Infected-Recovered (SIR) [7]. We compare our results with five centrality algorithms: Degree, In-Degree, Out-Degree, Closeness, and PageRank [1] using the indirect scheme. We use the diffusion record of the "Facebook" and the "Earthquake" topics in Plurk micro-blog for our experiments. In the preliminary results, we find that IC model performs best for both datasets using direct scheme. Using indirect scheme, LT model, Degree, In-Degree and PageRank outperform other methods.

By understanding the effectiveness of the models under different circumstances, we can explore the hidden behavior of diffusion effect change, identify the influential users on the network, and simulate the diffusion path of specific information. Also, we can carry out essential applications such as broadcasting emergency governmental announcements and advertising products or services to potential customers.

The contribution of this paper is two-fold.

1. We devise EPIC, a general evaluation framework, to evaluate the performance of the models and select the

This study is conducted under the "Social Media Analysis Platform (1/1)" of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

most feasible model to estimate information diffusion behavior.

2. We design direct and indirect schemes (the latter one consists of one-by-one and leave-one-out flows) in the EPIC framework to assess the performance of the diffusion models under different circumstances.

The remainder of the paper is organized as follows. In the next section, we describe the EPIC framework; in Section 3, we present the experiment results. Section 4 contains a literature survey. Then, in Section 5, we provide some concluding remarks.

II. THE EPIC FRAMEWORK

The EPIC framework is shown in Figure 1. The input of EPIC is a set of diffusion models, a social network, and the diffusion records for specific information; the output is the most suitable model. The EPIC framework consists of two schemes: direct and indirect. The indirect scheme contains two flows: one-by-one and leave-one-out.

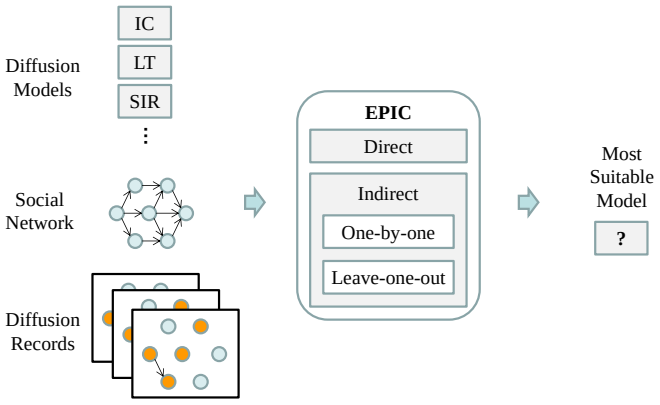

Figure 1. The EPIC framework. Given a set of diffusion models, a social network, and diffusion records, we evaluate the models using direct and indirect schemes to decide the most suitable model.

The direct scheme is to compare the links of predicted diffusions and the links of ground diffusions. In this work, we use static social network with directed links and binary diffused information to evaluate the models. The performance of the diffusion models is determined using F-score of the predicted and the ground links.

The direct scheme is straightforward as it assumes that the influence of all users is the same. However, the top influential users are usually more interesting in real-world. Therefore, we design the indirect scheme which considers the top influential users only.

In order to illustrate the indirect scheme, we first introduce *scale of propagation* [4], a metric to evaluate the influence of the users. Two definitions of propagation is described in [4]: loose (or upper-bound influence) and rigid (or lower-bound influence). In our work we apply the loose definition because it can capture the information diffusion behavior better (i.e., a user replies or reposts a post means that the user knows the information). To evaluate the influence of a seed user, we firstly construct the upper-bound influence tree using BFS-like search method on the diffusion record. It should be noted that if a user replies to multiple posts, only the *first* reply is counted; this is because it is the first time that user received the information. Then, we count the total number of the users (i.e., nodes) in the corresponding influence tree. This number is the scale of propagation for the seed user. Note that the seed node itself is also counted. For example, a diffusion record is shown in Figure 2(a). In Figure 2(b), the scale of propagation of seed node "*a*" is s_a = 5 ("*a*", "*c*", "*e*", "*f*", "*g*"); in Figure 2(c), the scale of propagation of seed node "*b*" is s_b = 4 ("*b*", "*c*", "*d*", "*e*").

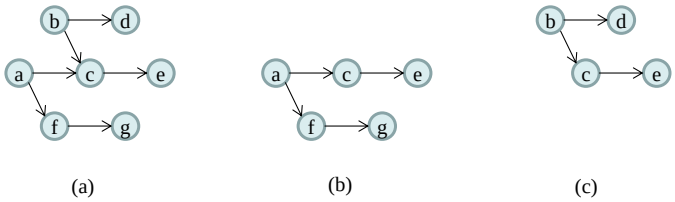

Figure 2. The indirect scheme using one-by-one flow. (a) An example of the diffusion record, assuming node "*a*" and "*b*" are seed nodes. (b) Using node "*a*" as the seed node, the one-by-one scale of propagation is 5. (c) Using node "*b*" as the seed node, the one-by-one scale of propagation is 4.

Using scales of propagation, the proposed indirect scheme is as follows:

Step 1: Identify root users in the ground diffusions. A root user is a user that diffuses the information to others but no one diffuses the information to that user.

Step 2: Compute the scales of propagation of the root users using the ground diffusions. It should be noted that the scales of propagation are computed *one-by-one*. That is, for each root user, we construct a corresponding influence tree and compute the scale of propagation. The scales of propagation are denoted as S_{ground}.

Step 3: Calculate the scales of propagation of the root users one-by-one using the diffusions predicted with the diffusion model. The scales of propagation are denoted as $S_{prediction}$.

Step 4: Select top k scored root users from S_{ground} and $S_{prediction}$ and then compute the F-score to determine the performance of the prediction.

However, evaluating the influences of the users one-by-one might *overestimate* the diffusion power of each user. As shown in Figure 2(b) and Figure 2(c), the nodes "*c*" and "*e*" are actually double-counted for the scales of propagation of root node "*a*" and "*b*", thus we overestimated the diffusion power of "*a*" and "*b*". Therefore, we design another flow,

leave-one-out, to evaluate the prediction results. Suppose we want to compute the influence of node *v*. First, we compute the scale of propagation for the whole diffusion record s_{all}. Then, we compute the scale of propagation *without* node *v*, which is s_{all-v}. Finally, the leave-one-out scale of propagation of node v is

$$s_v' = s_{all} - s_{all-v}$$

The intuition behind the leave-one-out flow is that we want to know "*if a user were not present, how much influencial power will the whole diffusion record lose*?". For example, to compute the influence of the root node "*a*" shown in Figure 3(a), we first compute the whole scale of propagation $s_{all} = 7$. Then, the influence without node "*a*" is $s_{all-a} = 4$, which is the same as the one-by-one scale of propagation s_b shown in Figure 2(c). Finally, the leave-one-out scale of propagation $s_a' = s_{all} - s_{all-a} = 7 - 4 = 3$, as shown in Figure 3(b). Similarly, we can compute $s_b' = 7 - 5 = 2$, as shown in Figure 3(c). Note that while the one-by-one flow tends to overestimate the influencial power of the nodes, the leave-one-out flow tends to *underestimate* the influencial power. We believe the real influencial power lies between the scales of propagations computed using two flows.

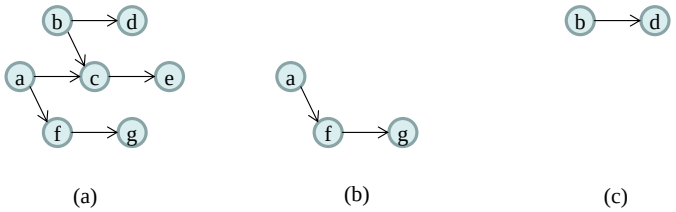

Figure 3. The indirect scheme using leave-one-out flow. (a) The same example of the diffusion record as shown in Figure 2. (b) Using node "*a*" as the seed node, the leave-one-out scale of propagation is 3. (c) Using node "*b*" as the seed node, the leave-one-out scale of propagation is 2.

III. EXPERIMENT

In this section, we describe the dataset, diffusion models, baseline methods, and results. We use a machine with AMD Opteron 2350 2.0GHz Quad-core CPU and 32GB RAM to run the experiments.

A. Datasets

In our experiment, we collect data from the Plurk micro-blog system, a popular micro-blog service in Asia. According to a previous analysis, more than 5 million users formed a giant graph in Plurk in 2009 [9]. We gather social network and diffusion records from Plurk using two different methods, and form the following two datasets respectively:

- *Friend-Facebook*. We gather social network using the largest connected component from Plurk. Then we collect messages and response from the users in the social network. In this dataset, the number of nodes (users) = 857,869, the number of links (friendship relations) = 15,751,810, the number of messages = 20,307,490, and the number of responses = 91,688,014. The duration of the messages and responses is from 2009/02/01 to 2009/05/24. We select the most frequent topic (after removing stop words), "Facebook", which contains 73,078 messages and 279,822 responses. We then remove messages without response, and use only the *first* response as diffusion. Thus, there are 16,561 diffusions. We use the newest 10% of diffusion records, total 1,656 diffusions, for testing.
- *Topic-Earthquake*. We first identify 100 hot topics from Plurk, and then search the whole Plurk to collect the users who post or reply related articles. Then, we collect the two-hop neighbors of the users. In this dataset, the number of nodes = 940,070 and the number of links = 7,660,770. The duration of the messages and responses is from 2011/01/01 to 2011/05/15. The most frequent topic is "Earthquake", which contains 80,336 messages and 303,523 responses. There are 38,453 diffusions, and among them there are 3,845 diffusions for testing.

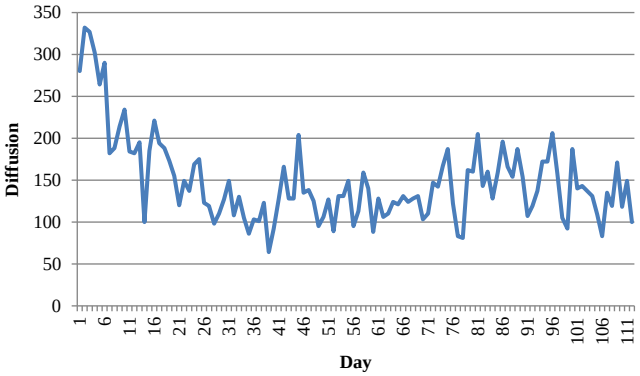

Figure 4. The diffusion-versus-day plot for the Friend-Facebook dataset.

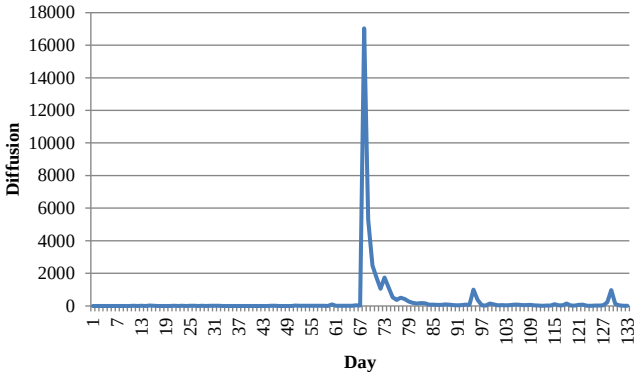

Figure 5. The diffusion-versus-day plot for the Topic-Earthquake dataset.

We select k=50 for the indirect scheme. Note that we removed the responses from users who are not part of the selected social network. The diffusion-versus-day plots for the two datasets are shown in Figure 4 and Figure 5; the Plurk-Large-Facebook dataset contains fluctuating diffusions, while the Plurk-Hot-Earthquake dataset contains single spike diffusions (the spike is at the day after the disastrous Japanese earthquake)

B. Diffusion Models

We utilize three diffusion models in our experiments. The diffusion models we apply are:

- *Independent Cascade (IC)*. When node v becomes active, it has a single chance of activating any currently inactive neighbor w. The activation attempt succeeds with probability p_{vw}. Whether or not v succeeds, it cannot make any further attempts to activate w in subsequent rounds. The process runs until no more activation is possible. The single parameter of IC model in our experiment is p_{vw}.

- *Linear Threshold (LT)*. A node v is influenced by each neighbor w according to a weight b_{vw} such that

  $$\sum_{w \text{ neighbor of } v} b_{vw} \leqq 1$$

  Each node v has a threshold θ_v. A node v is diffused if

  $$\sum_{w \text{ neighbor of } v} b_{vw} \geqq \theta_v$$

  The process continues until no more activation is possible. In our experiment, we set $b_{vw} = 1 / degree_v$, thus the single parameters of the LT model in our experiment is θ_v.

- *Susceptible-Infected-Recovered (SIR)*. Each node can be one of the three states: susceptible (is healthy but can catch diseases if exposed to some infective), infected (has a certain disease and can pass it on), and recovered (has recovered from the disease and has permanent immunity). The birth rate β is defined as the probability a node catches the disease from an infective one and the death rate γ is defined as the probability that an infected individual recovers. The mathematical formulation of the SIR model at time t is then defined as

  $$ds/dt = -\beta_{is}$$
  $$di/dt = \beta_{is} - \gamma_i$$
  $$dr/dt = \gamma_i$$

  In our experiment, we set γ_i = 0.01, thus the single parameter of the SIR model in our experiment is β_{is}.

For the single parameter for each diffusion model, we arbitrarily select three values in our experiment.

C. Baseline Methods

For direct scheme, we use simple random walk algorithm as baseline method. In each step, we randomly select an activated user and then activate an inactive neighbor of that user. The diffusion process continues until a specific number of nodes are activated. However, the number to stop walking is not known in prior. Therefore, we use average activated edges of all model and parameter combinations as the stopping number. It should be noted that the activated edges includes both originally and newly activated ones.

For indirect scheme, we exploit five centrality algorithms as baseline methods which are all well-known algorithms.

- *Degree, In-Degree, Out-Degree*. The intuition is that the users who have many friends are influential nodes. There is no parameter for the Degree, In-Degree and Out-Degree algorithms.

- *Closeness*. Closeness is based on the length of average shortest path between a node and all nodes in the graph. The intuition is that those who are close to others are influential nodes. There is no parameter for the Closeness algorithm.

- *PageRank*. PageRank [1] is a well-known centrality measure. In PageRank, the importance of a node v is measured as the probability of a random visit to the node

  $$PR_v = \alpha/T + (1 - \alpha) * \Sigma(PR_w / L_w)$$

  Where T is the total number of nodes, α is the probability of leaving a node (usually set as 0.15), each w is a node that contains a link to v, and L_w is the number of outgoing links in w. The intuition is that a node is more influential if it is connected to some influential nodes.

It should be noted that the centrality algorithms can only be compared using the indirect scheme because these algorithms cannot explicitly predict the diffusions. Also, because these algorithms compute the scales of user influence directly, using one-by-one or leave-one-out flow makes no difference.

D. Results

The results for two Plurk datasets are as follows:

- *Friend-Facebook*. The results using direct scheme are shown in Table 1. The baseline random walk method

uses average activated edges (667,537) as stopping number and performs poorly (0.0146). Comparing all model and parameter combination, IC model with p_{vw} = 0.50 performs best. In general, IC and SIR model outperform LT model. Note that the activated edges count of the IC model with p_{vw} = 0.80 and SIR model with β_{is} = 0.20 are close to that of the random walk baseline but the results of diffusion models is significantly better. Although the best F-score does not seem high (0.0479), it should be noted that we are predicting 1,656 ground truth diffusions (which is unseen in the historical records) from a large number of candidate friendship links (about 15.8M), which is an extremely difficult task. As for the indirect scheme, the results are shown in Table 2. LT model with parameter θ_v =0.10 and SIR model with β_{is} = 0.20 perform best using on-by-one flow. For leave-one-out flow, degree centrality performs best.

TABLE I. RESULTS FOR DIRECT SCHEME IN FRIEND-FACEBOOK DATASET

Method	Parameter	Activated Edges	F-Score
Random Walk	N/A	667,537	0.0146
IC Model	p_{vw} = 0.20	114,809	0.0273
	p_{vw} = 0.50	417,265	**0.0479**
	p_{vw} = 0.80	639,931	0.0236
LT Model	θ_v = 0.05	840,305	0.0034
	θ_v = 0.10	839,398	0.0206
	θ_v = 0.20	837,662	0.0129
SIR Model	β_{is} = 0.20	663,920	0.0196
	β_{is} = 0.50	816,891	0.0332
	β_{is} = 0.80	837,648	0.0456

TABLE II. F-SCORE RESULTS FOR INDIRECT SCHEME IN FRIEND-FACEBOOK DATASET

Method	Parameter	One-by-one	Leave-one-out
Degree	N/A	0.2000	**0.2000**
In-Degree	N/A	0.1600	0.1600
Out-Degree	N/A	0.1600	0.1600
Closeness	N/A	0.1800	0.1800
PageRank	α = 0.15	0.1400	0.1400
IC Model	p_{vw} = 0.20	0.2000	0.0800
	p_{vw} = 0.50	0.2000	0.1000
	p_{vw} = 0.80	0.2000	0.1200
LT Model	θ_v = 0.05	0.2000	0.1000
	θ_v = 0.10	**0.2200**	0.1000
	θ_v = 0.20	0.1800	0.1000
SIR Model	β_{is} = 0.20	**0.2200**	0.1200
	β_{is} = 0.50	0.2000	0.1600
	β_{is} = 0.80	0.2000	0.0600

- *Topic-Earthquake*. The results using direct scheme are shown in Table 3. IC model with p_{vw} = 0.80 performs best. In general, three models preforms almost equally good. Although the best F-score does not seem high (0.0332), it should be noted that we are predicting 3,845 ground truth diffusions (which is unseen in the historical records) from a large number of candidate friendship links (about 7.7M), which is an extremely difficult task. As for the indirect scheme, the results are shown in Table 4. The In-Degree and PageRank baseline preform best, while the diffusion models do not perform well.

TABLE III. RESULTS FOR DIRECT SCHEME IN TOPIC-EARTHQUAKE DATASET

Method	Parameter	Activated Edges	F-Score
Random Walk	N/A	664,643	0.0150
IC Model	p_{vw} = 0.20	103,940	0.0114
	p_{vw} = 0.50	354,454	0.0225
	p_{vw} = 0.80	662,588	**0.0314**
LT Model	θ_v = 0.05	890,634	0.0303
	θ_v = 0.10	890,516	0.0255
	θ_v = 0.20	882,833	0.0140
SIR Model	β_{is} = 0.20	570,016	0.0117
	β_{is} = 0.50	770,598	0.0200
	β_{is} = 0.80	856,212	0.0279

TABLE IV. F-SCORE RESULTS FOR INDIRECT SCHEME IN TOPIC-EARTHQUAKE DATASET

Method	Parameter	One-by-one	Leave-one-out
Degree	N/A	0.1200	0.1200
In-Degree	N/A	**0.1600**	**0.1600**
Out-Degree	N/A	0.0000	0.0000
Closeness	N/A	0.0000	0.0000
PageRank	α = 0.15	**0.1600**	**0.1600**
IC Model	p_{vw} = 0.20	0.0200	0.0800
	p_{vw} = 0.50	0.0200	0.0400
	p_{vw} = 0.80	0.0200	0.0400
LT Model	θ_v = 0.05	0.0400	0.0400
	θ_v = 0.10	0.0000	0.0400
	θ_v = 0.20	0.0400	0.0400
SIR Model	β_{is} = 0.20	0.0400	0.0400
	β_{is} = 0.50	0.0000	0.0200
	β_{is} = 0.80	0.0200	0.0400

IV. RELATED WORK

Many recent studies have focused on devising diffusion models to simulate diffusion behavior, such as Independent Cascade model [5], Linear Threshold model [5], Susceptible-Infected-Recovered model [7], Susceptible-Infected-Susceptible model [7], and Heat Diffusion model [11]. Usually, these models are evaluated using certain characteristics of real diffusion phenomenon. For example, the Independent Cascade and the Linear Threshold models [5] are evaluated using the target-set-size to coverage curve while the Greedy and the Courtesy models [3] are evaluated using the user-threshold to coverage curve. The performance of the diffusion models cannot be easily compared because the diffusion models do not exploit the same set of characteristics for evaluation (ex. target-set-size to coverage curve and user-threshold to coverage

curve). Furthermore, for different data and diffusion information, it is difficult to decide the most fitting diffusion model and corresponding parameters. Although some recent literatures define evaluation metrics for the influence of users, such as scale of propagation [4], a general evaluation mechanism remains lacking.

On the other hand, there are several brilliant works in domains related to the diffusion phenomenon such as identifying influential users [2, 14, 16, 17], selecting an initial user set with maximum influences given a diffusion model [5, 6, 8], mining diffusion pattern [10], recognizing diffusion sequences [15], sampling networks to identify influential users [12], and constructing the underlying diffusion network given historical diffusion records [13]. But among these researches, few attempts to evaluate the capabilities of different diffusion models.

V. CONCLUSION

In this paper, we propose the EPIC framework to evaluate the prediction capabilities of different diffusion models. The EPIC framework consists of direct and indirect schemes to assess the capability of diffusion prediction and influential user prediction respectively. For the indirect scheme, we further devise two flows: one-by-one and leave-one-out to give the bounds of the predicting power. Also, we conduct preliminary experiments to compare three famous diffusion models (IC, LT, and SIR). We test on two Plurk datasets: Friend-Facebook which contains fluctuating diffusions, and Topic-Earthquake which contains single spike diffusions. We find that IC model performs best for both datasets using direct scheme. Using indirect scheme, LT model and Degree perform good for Friend-Facebook dataset, while In-Degree and PageRank centrality perform good for Topic-Earthquake dataset. Although the direct prediction results still has room for improvement, we believe that for this difficult problem (ex. in the Friend-Facebook dataset, predict 1,656 diffusions from 15.8M candidate links, none of the 1,656 diffusions are seen in historical records), we provide an initial research direction. On the other hand, from the indirect prediction results we show that diffusion models perform equally well comparing to centrality algorithms.

In the future we think there are three plausible studies. First, the model parameters might be fine-tuned or new diffusion models might be adopted to provide better prediction results based on the EPIC framework. Second, new evaluation flow which can estimate the prediction power between the bounds provided by EPIC (i.e., from the one-by-one and leave-one-out flows) might be devised. Finally, more experiments on different diffusion models, social networks and diffusing information might be conducted using EPIC framework.

REFERENCES

[1] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems,* vol. 30, pp. 107-117, 1998.

[2] M. Cha, *et al.*, "Measuring User Influence in Twitter: The Million Follower Fallacy," in *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[3] M. Gupte, *et al.*, "News Posting by Strategic Users in a Social Network," presented at the Proceedings of the 5th International Workshop on Internet and Network Economics, Rome, Italy, 2009.

[4] C.-T. Ho, *et al.*, "Modeling and Visualizing Information Propagation in a Micro-blogging Platform," in *ASONAM*, 2011.

[5] D. Kempe, *et al.*, "Maximizing the spread of influence through a social network," presented at the Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C., 2003.

[6] D. Kempe, *et al.*, "Influential Nodes in a Diffusion Model for Social Networks," in *Automata, Languages and Programming*. vol. 3580, L. Caires, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2005, pp. 1127-1138.

[7] W. O. Kermack and A. G. McKendrick, "A Contribution to the Mathematical Theory of Epidemics," in *Roy. Soc. Lond.*, 1927.

[8] M. Kimura, *et al.*, "Extracting influential nodes on a social network for information diffusion," *Data Min. Knowl. Discov.*, vol. 20, pp. 70-97, 2010.

[9] H.-C. Lai, *et al.*, "Exploiting Cloud Computing for Social Network Analysis – Exemplified in Plurk Network Analysis," in *TAAI*, 2009.

[10] J. Leskovec, *et al.*, "Patterns of Influence in a Recommendation Network," in *Advances in Knowledge Discovery and Data Mining*. vol. 3918, W.-K. Ng, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2006, pp. 380-389.

[11] H. Ma, *et al.*, "Mining social networks using heat diffusion processes for marketing candidates selection," presented at the Proceeding of the 17th ACM conference on Information and knowledge management, Napa Valley, California, USA, 2008.

[12] A. S. Maiya and T. Y. Berger-Wolf, "Online Sampling of High Centrality Individuals in Social Networks," in *PAKDD*, 2010.

[13] M. G. Rodriguez, *et al.*, "Inferring networks of diffusion and influence," presented at the Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, DC, USA, 2010.

[14] X. Song, *et al.*, "Identifying opinion leaders in the blogosphere," presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, Lisbon, Portugal, 2007.

[15] A. Stewart, *et al.*, "Discovering information diffusion paths from blogosphere for online advertising," presented at the Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising, San Jose, California, 2007.

[16] J. Tang, *et al.*, "Analysing information flows and key mediators through temporal centrality metrics," presented at the Proceedings of the 3rd Workshop on Social Network Systems, Paris, France, 2010.

[17] C. C. Yang, *et al.*, "An analysis of user influence ranking algorithms on Dark Web forums," presented at the ACM SIGKDD Workshop on Intelligence and Security Informatics, Washington, D.C., 2010.