

# Diffusion of “Following” Links in Microblogging Networks

Jing Zhang, Zhanpeng Fang, Wei Chen, *Member, IEEE*, and Jie Tang, *Senior Member, IEEE*

**Abstract**—When a “following” link is formed in a social network, will the link trigger the formation of other neighboring links? We study the diffusion phenomenon of the formation of “following” links by proposing a model to describe this link diffusion process. To estimate the diffusion strength between different links, we first conduct an analysis on the diffusion effect in 24 triadic structures and find evident patterns that facilitate the effect. We then learn the diffusion strength in different triadic structures by maximizing an objective function based on the proposed model. The learned diffusion strength is evaluated through the task of link prediction and utilized to improve the applications of follower maximization and followee recommendation, which are specific instances of influence maximization. Our experimental results reveal that incorporating diffusion patterns can indeed lead to statistically significant improvements over the performance of several alternative methods, which demonstrates the effect of the discovered patterns and diffusion model.

**Index Terms**—Link diffusion, triad formation, social network

## 1 INTRODUCTION

$$0 \leq t - t' \leq \delta, \quad \delta = \min_{\substack{A, B, C \\ A', B', C'}} \left( \inf_{t \in [0, T]} |x_t^A - x_t^{A'}|, \inf_{t \in [0, T]} |x_t^B - x_t^{B'}|, \inf_{t \in [0, T]} |x_t^C - x_t^{C'}| \right).$$

- J. Zhang and Z. Fang are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.  
E-mail: {zhangjing12, fzp13}@mails.tsinghua.edu.cn.
- W. Chen is with Theory Group, Microsoft Research, Beijing 100080, China. E-mail: weic@microsoft.com.
- J. Tang is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, and Tsinghua National Laboratory for Information Science and Technology (TNList).  
E-mail: jietang@tsinghua.edu.cn.

*Manuscript received 11 Jan. 2014; revised 5 Jan. 2015; accepted 9 Feb. 2015.  
Date of publication 25 Feb. 2015; date of current version 2 July 2015.*

Recommended for acceptance by G. Das.

For information on obtaining reprints of this article, please send e-mail to: [reprints@ieee.org](mailto:reprints@ieee.org), and reference the Digital Object Identifier below.

Digital Object Identifier no. 10.1109/TKDE.2015.2407351

## 2 “FOLLOWING” LINK CASCADE MODEL

neighbor-  
ing links

( )

$G = (V, E, t)$ ,  $v \in V$ ,  $e_{uv} \in E$ ,  $u \sim v$ ,  $e \in A \cup B$ ,  $t : E \rightarrow \mathbb{N} \cup \{\perp\}$ .  
 $t(e_{uv}) = n \in \mathbb{N}$ ,  $t(e_{uv}) = \perp$ ,  $t(e_{uv}) = t_e$ ,  $e' \in A \cup B$ ,  $C$ .

**A. 1. Diffusion effect between links decays over time.**  
 discovery probability  $g_{e'e}$ , diffusion probability  $h_{e'e}$ ,  $A$ ,  $B$ ,  $C$ .

$t' + \delta$ ,  $t'$ ,  $t'$ ,  $e'$ ,  $g_{e'e}$ ,  $e'$ ,  $h_{e'e}$ ,  $e$ ,  $\lambda$ ,  $t' + \lambda$ ,  $e'$ ,  $e$ ,  $\delta$ .

Organization.

Fig. 2. Diffusion

$$(r, \delta = \frac{1}{2})$$

3.1 Data Collection

0,000

0/ / 0 0 / / 0 0.

0/ / 0 0 / / 0 0.

(

3.2 Observations

1 12

13 24

( )

( )

$t'$   $t'$

( )  $t$

$0 \leq t - t' \leq \delta$  ( $\delta$ ).

$C_{\Delta}$

$\Delta,$   $C_{\Delta}^+$

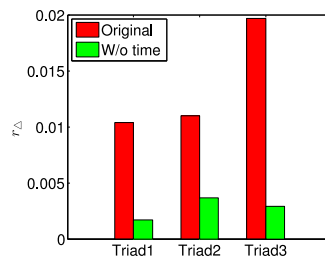
$B$   $C$   $[t', t' + \delta]. |C_{\Delta}|$

$\Delta.$   $r_{\Delta}$

$B$   $C$

$$r_{\Delta} = \frac{|C_{\Delta}^+|}{|C_{\Delta}|}. \tag{1}$$

- -
- Pattern significance.

 $\leq 0.05,$ 
$$r_{\Delta} \dots \delta \dots (\dots, t = \perp), \dots r_{\Delta} \dots$$

0,000

0.0,

$$A \quad B \quad B \quad C \quad C'$$

$$- \quad - \quad - \quad - \quad -$$

$$C \quad W^- \quad , \quad , \quad , \quad 0,$$
$$e_{AC} = \frac{C}{C - C} = 1$$
$$C \quad \quad \quad e_{AC}$$
 $e_{AC} \dots$ 

*Diffusion decay.*

$$\delta \quad 1, 2, 3, 5, 7 \quad 10$$
$$\delta, \dots, r_{\Delta} \dots$$
 $r_{\Delta}$ 
$$\delta \dots \dots \dots , r_{\Delta}$$
 $r_{\Delta}$ 
$$B \quad A \quad C, \quad A'$$
$$( \quad , \quad )_C, \quad B \quad A \quad C \quad , B$$
 $\delta$ 

*Other observations.*

 $A \quad B.$ 

0- ).

$$e_{BA}, \dots$$
$$e_{AB}, \dots, e_{AB}, \dots, e_{AB}, \dots, e_{BA}.$$
$$B \dots \begin{matrix} A & B \\ & C \\ A & B \end{matrix} \left( \dots + \dots \right).$$

*A. C.*

$$\left( \frac{1}{2} - \frac{1}{2} \right) \left( \frac{1}{2} - \frac{1}{2} \right) = \frac{1}{4}$$
$$\begin{array}{ccc} & A & C \\ B & & C \\ & C & A(+, -, 0, \dots). \end{array}$$
$$\begin{array}{ccccc} B & & A & & B \\ A' & & C & & A' \\ & C' & & A' & C \end{array}$$

*Summary.*

$$\begin{aligned}
 & \dots (+, \dots) \dots - \dots - \\
 & \dots A \dots C \dots - \\
 & \dots A \dots C \dots B \dots C (+, -, 0, \dots).
 \end{aligned}$$

## 4 MODEL LEARNING

$$\begin{aligned}
 & \dots / \dots \\
 & \dots \bullet \dots \\
 & \dots \bullet \dots \\
 & \text{Likelihood function.} \dots / \dots \\
 & \dots \bullet \dots \\
 & \dots \theta = \{h_{e'e}, g_{e'e}\} \dots - \\
 & \dots \bullet \dots / \dots \\
 & \dots h_{e'e} \dots \\
 & \dots g_{e'e} \dots (e', e) \dots \\
 & \dots (e \dots
 \end{aligned}$$

$$e \begin{matrix} & y_{e'e} & & e' \\ [t_{e'}, t_e], & & & t_{e'} \\ t_e, e' & e & & y_{e'e} \\ & e' & e & t_{e'} \end{matrix}$$

$$\begin{aligned} y_{e'e} &= 1 - h_{\Delta} g_{\Delta} \sum_{t=t_{e'}}^{t_e} (1 - g_{\Delta})^{t-t_{e'}} \\ &= h_{\Delta} (1 - g_{\Delta})^{t_e - t_{e'} + 1} + (1 - h_{\Delta}). \end{aligned} \tag{6}$$

$$\begin{aligned} & e \in \mathcal{E} \\ & \delta \\ & e' \quad y_{ee'} \quad e' \in R_e, \quad R_e \\ & e \quad t_e + \delta. \\ & y_{ee'} \quad ( ), \quad t_{e'} \\ & t_e \quad t_e \quad t_e + \delta \end{aligned}$$

$$\log \mathcal{L} = \sum_{e \in \mathcal{E}} \left\{ \log \sum_{\vec{\alpha}_{S_e}} \prod_{e' \in S_e} x_{e'e}^{\alpha_{e'}} y_{e'e}^{1-\alpha_{e'}} + \sum_{e' \in R_e} \log y_{ee'} \right\}.$$

EM algorithm.

$$q(e|\vec{\alpha}_{S_e}) = \frac{p(e|\vec{\alpha}_{S_e})}{p(e|S_e)}$$

$$\begin{aligned} \log \mathcal{L} &= \sum_{e \in \mathcal{E}} \left\{ \log \sum_{\vec{\alpha}_{S_e}} \hat{q}(e|\vec{\alpha}_{S_e}) \frac{p(e|\vec{\alpha}_{S_e})}{\hat{q}(e|\vec{\alpha}_{S_e})} + \sum_{e' \in R_e} \log y_{ee'} \right\} \\ &\geq \sum_{e \in \mathcal{E}} \left\{ \sum_{\vec{\alpha}_{S_e}} \hat{q}(e|\vec{\alpha}_{S_e}) \log \frac{p(e|\vec{\alpha}_{S_e})}{\hat{q}(e|\vec{\alpha}_{S_e})} + \sum_{e' \in R_e} \log y_{ee'} \right\}, \end{aligned}$$

$$\begin{aligned} & \hat{q}(e|\vec{\alpha}_{S_e}) \log \hat{q}(e|\vec{\alpha}_{S_e}) \\ & Q(\theta, \hat{\theta}) \end{aligned}$$

$$Q(\theta, \hat{\theta}) = \sum_{e \in \mathcal{E}} \left\{ \sum_{\vec{\alpha}_S}$$





where  $u$  and  $v$  are the two nodes connected by the edge  $e$ ,  $S_e$  is the set of nodes that are connected to both  $u$  and  $v$ , and  $p(e|S_e)$  is the probability of observing the edge  $e$  given the set of nodes  $S_e$ .

### Evaluation metrics.

We use two metrics to evaluate the performance of the proposed method. The first metric is the  $F_1$  score, which is the harmonic mean of precision and recall. The second metric is the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve.

The  $F_1$  score is calculated as follows:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

The AUC of the ROC curve is a measure of the model's ability to distinguish between the positive and negative classes. It is calculated as the area under the curve of the ROC curve.

The AUC of the ROC curve is calculated as follows:

$$\text{AUC} = \frac{\sum_{i=1}^n \text{rank}(x_i)}{n(n+1)},$$

where  $x_i$  is the score assigned to the  $i$ -th instance, and  $n$  is the total number of instances.

### Comparison methods.

#### Basic.

The Basic method is a simple baseline method that uses the degree of the nodes to predict the following links.

#### SVM.

The SVM method is a machine learning method that uses a Support Vector Machine to predict the following links.

#### LRC.

The LRC method is a machine learning method that uses a Logistic Regression Classifier to predict the following links.

#### Collaborative filtering (CF):

The CF method is a machine learning method that uses Collaborative Filtering to predict the following links. It is based on the assumption that users who have followed similar users in the past will follow similar users in the future.

$$CF\_score(u, v) = \sum_w I(w, v) sim(w, u),$$

where  $sim(w, u)$  is the similarity between user  $w$  and user  $u$ , and  $I(w, v)$  is the indicator function that is 1 if user  $w$  has followed user  $v$ , and 0 otherwise.

The similarity between two users  $u$  and  $v$  is calculated as follows:

$$sim(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|},$$

where  $N(u)$  is the set of users that user  $u$  has followed, and  $N(v)$  is the set of users that user  $v$  has followed.

The Random-random model (RR) is a baseline method that randomly selects users to follow.

The Random-random model (RR) is a baseline method that randomly selects users to follow. It is based on the assumption that the probability of a user following another user is proportional to the degree of the user being followed.

The Random-random model (RR) is a baseline method that randomly selects users to follow. It is based on the assumption that the probability of a user following another user is proportional to the degree of the user being followed.

#### Random-random model (RR).

The Random-random model (RR) is a baseline method that randomly selects users to follow. It is based on the assumption that the probability of a user following another user is proportional to the degree of the user being followed. The score for the Random-random model (RR) is calculated as follows:

$$RR\_score(u, v) = \frac{1}{|F(u)|} \sum_w I(u, w) I(w, v) \frac{1}{|F(w)|},$$

where  $|F(u)|$  is the number of users that user  $u$  has followed, and  $|F(w)|$  is the number of users that user  $w$  has followed.

The Random-random model (RR) is a baseline method that randomly selects users to follow.

#### Preferential attachment with communities (PAC).

The PAC method is a machine learning method that uses Preferential Attachment with Communities to predict the following links. It is based on the assumption that the probability of a user following another user is proportional to the degree of the user being followed, and the probability of a user following another user is also proportional to the probability of the user being followed being in the same community as the user following.

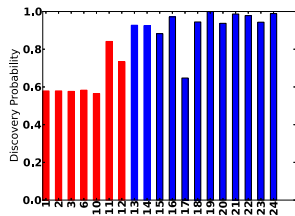
The PAC method is a machine learning method that uses Preferential Attachment with Communities to predict the following links. It is based on the assumption that the probability of a user following another user is proportional to the degree of the user being followed, and the probability of a user following another user is also proportional to the probability of the user being followed being in the same community as the user following.

The PAC method is a machine learning method that uses Preferential Attachment with Communities to predict the following links. It is based on the assumption that the probability of a user following another user is proportional to the degree of the user being followed, and the probability of a user following another user is also proportional to the probability of the user being followed being in the same community as the user following.

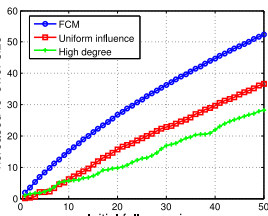
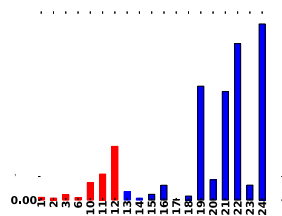
$$PAC\_score(u, v) = \beta \left( \alpha \frac{|N(v)|}{\sum_{v \in C(u)} |N(v)|} + (1 - \alpha) \frac{1}{|C(u)|} \right) + (1 - \beta) \left( \alpha \frac{|N(v)|}{\sum_{v \in V} |N(v)|} + (1 - \alpha) \frac{1}{|V|} \right),$$

where  $|N(v)|$  is the degree of user  $v$ ,  $|C(u)|$  is the number of users in the same community as user  $u$ , and  $|V|$  is the total number of users in the network.

$$\begin{array}{rcl}
 u, & & \\
 0. & V & \\
 u, & & \\
 u. & & - \\
 \alpha & \beta & 0 \quad 0. \quad . \\
 & & -
 \end{array}$$



(a) Discovery Probabilities



(a) Follower maximization

(b) Followee maximization

Fig. 11. Results for “following” influence maximization on Twitter. X-axis: the number of initial users. Y-axis: the number of newly activated users.

### Per-triad analysis.

### Delay analysis.

### Convergence analysis.

### Model parameter analysis.

## 6.3 Application Improvement

..... / ..... • .....  
..... / .....  
..... / ..... • .....  
.....  
..... / .....  
..... / ..... •

## 7 RELATED WORK

*Diffusion model and influence maximization.* .....  
.....  
.....  
.....  
..... / .....  
.....

## REFERENCES

- Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 00
- Proc. 4th ACM Int. Conf. Web Search Data Mining, 0
- Science, 0
- Proc. IEEE 12th Int. Conf. Data Mining, 0
- Nature, 0
- Proc. 11th SIAM Int. Conf. Data Mining, 0
- Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 0
- Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 0
- Phys. Rev. E, 0
- Proc. 16th Int. Conf. World Wide Web, 00
- Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 00
- Networks, Crowds, and Markets: Reasoning about a Highly Connected World, 00
- Tsinghua Sci. Technol., 0
- Permutation, Parametric and Bootstrap Tests of Hypotheses, 00
- Proc. 3rd ACM Int. Conf. Web Search Data Mining, 00
- Amer. J. Sociol., 0
- Proc. 13th Int. Conf. World Wide Web, 00
- Proc. 11th SIAM Int. Conf. Data Mining Workshop, 00
- Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 00
- Psychometrika, 0
- Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 00
- Proc. Roy. Soc. A, 00
- Proc. ACM Conf. Inf. Knowl. Manage., 00
- Intell. Data Anal., 0
- Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 0
- Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 00
- Proc. 21st Int. Conf. World Wide Web, 00
- Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 00
- Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 00
- Math. Biosci., 0
- Proc. ACM Int. Conf. Web Search Data Mining, 0
- J. Amer. Soc. Inf. Sci. Technol., 00
- Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 00
- Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 00
- Data Mining Knowl. Discovery, 0
- Sci. Technol., 0
- ACM Trans. Knowl. Discovery Data, 0
- Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 0
- Phys. Rev. E, 00
- Computer Intensive Methods for Testing Hypotheses, 0
- Proc. 4th Int. AAAI Conf. Weblogs Social Media, 00
- Int. J. Semantic Web Inf. Syst., 00
- Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 00
- Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 00
- Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 00
- Proc. 6th Int. Conf. Data Mining, 00
- Proc. 7th IEEE Int. Conf. Data Mining, 00

