Reconstruction of missing data in social network Based on Affinity Propagation

Rongxin LIU[1, a], Qun LIU[2, b]

[1] Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications University, Chongqing 400065, China;

[2]School of Computer Technology, Chongqing University of Posts and Telecommunications University, Chongqing 400065, China.

[a]liurongxin2012@live.com, [b]liuqun@cqupt.edu.cn

Keywords: Network Topology, social networks, affinity propagation, missing data recovery.

Abstract. With the appearance of data explosion, important data in incomplete network could be missed caused by some factors. To address the problem, we present a reconstruction framework based on Hawkes process with self-exciting and TAP (Topical affinity propagation) to effectively and efficiently reconstruct data. The existing methods mainly focus on how to replace missing values with some plausible estimate, but do not consider reconstruction efficiency and dynamic network. In our paper, we analyze temporal patterns and affinity propagation in the series of interactive events between nodes.

Introduction

In recent years, the prediction of missing information is one of the most important parts in the data analysis for social science [1-3]. Assuming that all of events in the network are known, only some end nodes are missing among these events. The problem is how to reconstruct these incomplete events based on some behavior models. In this paper, we pay attention to networks that changed over time and the strength of the nodes impact are quantified. We focus on the order of weight for each incomplete node to get the "most likely" nodes because the weights between nodes are proportional to the likelihood function.

Reconstruction Model

In this section, we make a brief description of the notations used in our paper firstly. Then we introduce the reconstruction method based on Hawkes Process and describe how to quantify the strength of the nodes impact in the process.

Table 1 Notations

SYMBOL	DESCRIPTION
N	total number of events
M	number of edges in the network
V	the set of nodes in the network
K	total number of pairs=k(k-1)/2
n	number of incomplete events
k	number of nodes
v_i	The *i*-th node

A Hawkes Process with self-exciting point process is proposed in [3, 4]. It has been commonly used in seismology to model the rate of earthquakes occurring and the rate jumps up following an event as one expects aftershocks. It is defined by intensity function as following.

$$\lambda(t) = \mu + \theta \sum_{t_i<t} g(t - t_i) \quad (1)$$

© 2015. The authors - Published by Atlantis Press

where μ represents Poisson background rate and $\theta \sum_{t_i<t} g(t-t_i)$ is a self-exciting component. The elevated rate spreads in time according to the kernel g, and θ reveals the scaling factor of the effect. Based on Eq. (1), Eq. (2) represents the intensity function among two nodes (such as α and β), which has their own interaction parameters.

$$\lambda_{\alpha\beta}(t) = \mu_{\alpha\beta} + \theta_{\alpha\beta} \sum_{t_i^{\alpha\beta}<t} \omega_{\alpha\beta} e^{-\omega_{\alpha\beta}(t-t_i^{\alpha\beta})} \quad (2)$$

where parameters $\mu_{\alpha\beta}$ represents Poisson background between the nodes α and β, $\theta_{\alpha\beta}$ reveals the scaling factor of the effect and $\omega_{\alpha\beta}$ represents the time scale.

In order to solve these problems, we introduce TAP (Topical affinity propagation) model [5] in incomplete networks to improve Hawkes process. The model can be used to quantify the strength of the nodes impact [5]. The influence function can be given as follow:

$$\varphi_{st}^{z} = \frac{1}{1+e^{-(\varphi_{ts}^{z}+\varphi_{st}^{z})}} \quad (3)$$

where φ_{st}^{z} represents the strength of the nodes impact for node v_s to node v_t in incomplete network z , φ_{ts}^{z} corresponds to the influence for node v_t to node v_s. We introduce influence function result φ_{st}^{z} (Eq.3) as a parameter to rewrite intensity function. It will be shown in Eq.5.

Now, we bring the strength of nodes impact to the energy function by substituting Eq. (2) and Eq. (3) into Eq. (4), Eq. (4) is rewritten as following:

$$\Lambda = \sum_{\alpha\beta} \sum_{i,j} \delta_{ij} \varphi_{\alpha\beta} + \frac{1}{2}(1-\delta_{ij})\theta_{\alpha\beta}\omega_{\alpha\beta} e^{-\omega_{\alpha\beta}|t_i^{\alpha\beta}-t_j^{\alpha\beta}|} \quad (4)$$

Thus, Λ is decoupled into the sum of the energies of the missing events. The background rates $\sum_{\alpha\beta} \sum_{i,j} \delta_{ij} \varphi_{\alpha\beta}$ and the sum of both nodes' energies $\frac{1}{2}(1-\delta_{ij})\theta_{\alpha\beta}\omega_{\alpha\beta} e^{-\omega_{\alpha\beta}|t_i^{\alpha\beta}-t_j^{\alpha\beta}|}$ are the determinant factors. The m_i is the weights for each incomplete event on the various time .The weights are proportional to the energy function [5]. So we only need to focus on the order of the various $m^{\alpha\beta}$ for each incomplete event and we can get the most likely nodes by Eq. (4). Eq. (5) aims to calculate the likelihood $m_i^{\alpha\beta}(f)$ by which the incomplete event i belongs to pair of nodes under metric f= $\mathcal{L}$[2] and the Eq. (5) is given as following:

$$m_i^{\alpha\beta}(f) = \sum_{\mathcal{A}_i^{\alpha\beta}} f \quad (\mathcal{A}) \quad (5)$$

where $\mathcal{A}$ represents the sum of possibilities for incomplete events and $\mathcal{A}_i^{\alpha\beta}$ represents possibility in which incomplete event i belongs to pair of nodes. As shown in Algorithm 1, we use pseudocode to represent process of the reconstruction with Hawkes process.

Experimental Study

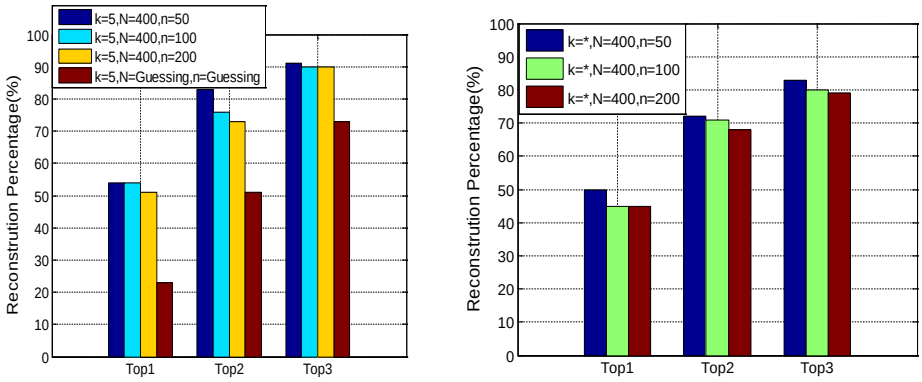
In order to evaluate the performance and accuracy of our approach, we adopt two real datasets: call record data (9,835 calling events and 75 cell towers) and DNS data from 360 Security Company (16051 users' actions for DNS address).We randomly select 30% events from datasets as missing data in events and make reconstruction for them by using our methods.

Performance on call record data

In the following experiments, one of the participants is unknown for each incomplete event. The parameters are as same as the above. In Figure 1(a) —(b), we compare our results of reconstruction percentage to Hawkes without influence. The $\star$ value of *k* corresponds to the real call records. The "Guess" rows show the parameter that obtained from random guessing. We can find that our method has better reconstruction percentage than those parameters obtained by guessing.

Table 2 Dimensions of the network and Possibility of events.

k	N	n	Top1	Top2	Top3	k	N	n	Top1	Top2	Top3
5	400	50	54%	83%	91%	7	400	50	46%	68%	80%
5	400	100	54%	76%	90%	7	400	100	45%	68%	79%
5	400	200	51%	73%	90%	7	400	200	45%	65%	77%
5	Guess	Guess	23%	51%	73%	7	Guess	Guess	16%	31%	47%
k	**N**	**n**	**Top1**	**Top2**	**Top3**	**k**	**N**	**n**	**Top1**	**Top2**	**Top3**
9	400	50	40%	64%	72%	*	400	50	50%	72%	83%
9	400	100	40%	61%	71%	*	400	100	45%	71%	80%
9	400	200	39%	55%	66%	*	400	200	45%	68%	79%
9	Guess	Guess	13%	25%	38%						



(*a*) *k*=5 (*b*) *k*=*

Fig1 Reconstruction percentage

Performance on DNS data

In the following experiments, we test our methods on the DNS network datasets received from Qihoo 360.We also use same parameters with call record data for each pair of nodes: $\mu = 10^{-2} days^{-1}$, $\omega = 10^{-1} days^{-1}$ and $\theta = 0.5$. We randomly select 40% events and assume the participants missing among these event. Then we make reconstruction using our methods. Table 3 demonstrates our method performance in DNS network. We can find that our method has better percentage than those parameter obtained by random guessing.

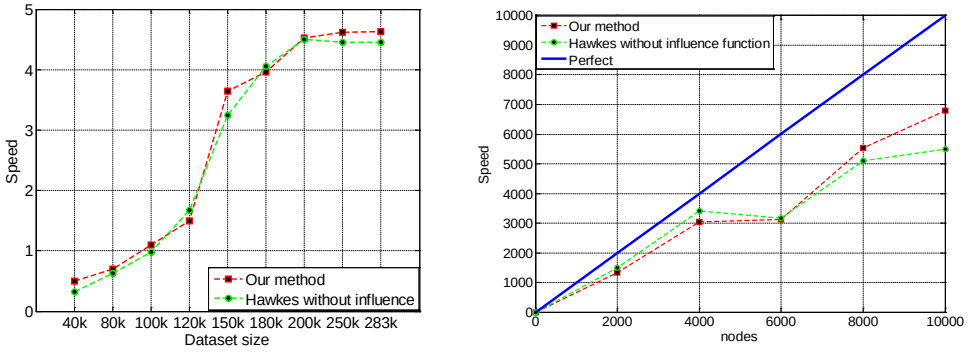


Fig2 Call record dataset size vs. speed and nodes vs. Speed

Table 3 Dimensions of the network and Possibility of events.

k	N	n	Top1	Top2	Top3	k	N	n	Top1	Top2	Top3
10	860	150	32%	51%	72%	15	860	150	40%	57%	70%
10	860	200	40%	51%	70%	15	860	200	41%	52%	66%
10	860	250	39%	55%	63%	15	860	250	33%	49%	71%
10	Guess	Guess	19%	23%	32%	15	Guess	Guess	13%	29%	43%
k	**N**	**n**	**Top1**	**Top2**	**Top3**	**k**	**N**	**n**	**Top1**	**Top2**	**Top3**
25	860	150	40%	63%	72%	*	860	150	53%	66%	73%
25	860	200	40%	59%	71%	*	860	200	48%	62%	66%
25	860	250	39%	55%	69%	*	860	250	44%	59%	71%
25	Guess	Guess	15%	25%	44%						

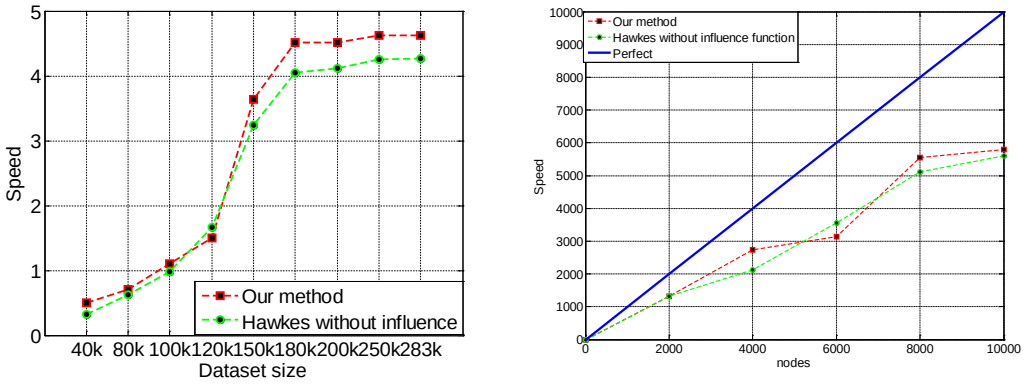


Fig3 DNS dataset size vs. speed and nodes vs. speed

As shown in the Table 2 and Table 3, we can see that our method captures the qualitative features of interaction process better than those obtained by random guessing. From Figure 1, it can be reported that our method has higher percentage than parameter guessing. Comparing with the Hawkes without influence, the result of our reconstruction model show the higher reconstruction speed and efficiency in Figure 2 and Figure 3.

Summary

Hawkes process is one of the most popular method to reconstruct missing data in incomplete network. However, studies about reconstruction missing data ignore the network change over time and computation expensive. In this paper, a new reconstruction method based on affinity propagation and temporal patterns of interactions events is proposed. Meanwhile, the dynamic network time parameters are considered. According to the results of experiments, our method outperforms other existed algorithm in terms of reconstruction speed and percentage.

References

[1] J. van der Geer, J.A.J. Hanraads, R.A. Lupton, Reconstruction of missing data in social networks based on temporal patterns of interactions, Inverse Problems. 27 (2011) 13-15.

[2] Blundell, Charles, Jeff Beck, and Katherine A. Heller, Modelling reciprocating relationships with Hawkes processes, Advances in Neural Information Processing Systems. 60 (2012) 79-89.

[3] Guimerà, Roger, and Marta Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, Proceedings of the National Academy of Sciences. 52 (2009) 22073-22078.

[4] Wang G N, Gao H, Chen L, Predicting Positive and Negative Relationships in Large Social Networks, PloS one. 15 (2015)154-163.

[5] Huisman M, Imputation of missing network data: some simple procedures, Journal of Social Structure. 18 (2009) 72-94.