Research on Social Network Inference Method Based on ConNIe Algorithm

Hailiang Chen
College of Systems Engineering,
National University of Defense Technology
Changsha, China
Email:chenhailiang_15@163.com

Bin Chen
College of Systems Engineering,
National University of Defense Technology
Changsha, China
Email:nudtcb9372@gmail.com

Jian Dong
College of Systems Engineering,
National University of Defense Technology
Changsha, China
Email:jiandong.nudt@foxmail.com

Lingnan He
The School of Communication and Design ,
Sun yat-sen University
Guangzhou, China
Email:lingnan.he@qq.com

***Abstract*—In recent years, Internet technology and online social networks have developed rapidly, enabling people to express their opinions, ideas, emotional exchanges and economic exchanges randomly. Inferences about social networks are made possible by observational data exchanged by people on the Internet. In this paper, through the analysis of ConNIe algorithm, the effects of sparse parameter, propagation time distribution model and its parameters on the inferred results of this algorithm are studied. Then, based on the research, this paper use perceptron algorithm to classify the propagation time distribution model and use particle swarm optimization algorithm to optimize the sparse parameter and the parameters of propagation time distribution model. Finally, a social network inference framework based on ConNIe algorithm is proposed to make up for ConNIe. Some of the shortcomings of the algorithm have gotten over. The research in this paper helps people to understand the social network itself, and it has a wide range of practical value in the fields of social public opinion control and marketing.**

Keywords—social network, network inference, ConNIe, Perceptron algorithm, particle swarm optimization

I. INTRODUCTION

In recent years, social informatization and networking have been accelerating, and social networks are gradually integrating into people's daily lives and exerting important influences. With the extensive use of social software such as QQ, WeChat, Weibo, Twitter, and YouTube on the Internet, a huge and complex social network has formed. Here, people can freely express their opinions, thoughts, emotional exchanges and economic exchanges by sending and receiving information such as text and voice, by which the spread of information on social networks is formed.

Social network is a complex network. Through the analysis of the structure of social networks, it can help to understand the social network itself and discover the rules of people's communication. Social networking, a bridge between the real world and the virtual world, provides a large amount of user data for studying social relationships. Network inference based on these user data has broad application value. Social networks have important practical significance for the government, who use the cluster effect, the hub nodes in social networks, in order to controlling public opinion[1].

The experiments in this article use simulated social network data. This will let you know the structure of the underlying network and compare it with the inferred results. The parts in the network have the same characteristics as the whole. Therefore, except studying the influence of nodes or edges, the rest researches use 200 nodes and 1000 edges, which is part of the actual social network, although the actual network has thousands of nodes,

This paper analyzes and studies the social network inference method based on observation data. At the begin with, the effects of sparse parameter, propagation time distribution model and its parameters on the inferred results of ConNIe algorithm are studied. Then, based on the research, this paper use perceptron algorithm to classify the propagation time distribution model and use particle swarm optimization algorithm to optimize the sparse parameter and the parameters of propagation time distribution model. In the end, a social network inference framework based on ConNIe algorithm is proposed.

II. RELATED WORK

When a message is generated on a social network, it spreads from one node to another through the edge of the potential social network, just like an infectious disease. However, real social networks are often unknown and invisible. Ghonge S et al. studied the inference of networks based on cascading[2]. Usually, you can only observe that a message has been propagated to a node at a certain time, and you don't know who passed the message to it. A large number of scholars have proposed many classic information dissemination models, including threshold models[3], cascade models, infectious disease models[4], and linear influence models[1]. Independent Cascade[5][6] (IC) is an information propagation model based on static structure graph. The model starts from a group of activated nodes. Each node in each step judges whether it is activated, according to the conditions given by the model. Once a node is activated, it remains active until no new nodes are activated and the model ends.

Today, researchers use electronic communications to build and research large social networks. They infer unobserved connections from observed communication events effectively. Psorakis I et al. proposed a methodology for extracting social network structure from spatio-temporal datasets that describe timestamped occurrences of individuals[7]. In this data, it is difficult to solve the optimal network problem. Manuel et al. proposed the scalable algorithm NETINF[8] to infer the influence and propagation of the network. This algorithm can get a near optimal network with the data of the set of k edges. The algorithm can be used for larger-scale network inference

978-1-7281-4762-8/19/$31.00

through localized renewal and lazy evaluation. With a relatively small data set, potential social networks can be accurately restored. However, the NETINF algorithm also has a problem of low accuracy. On the basis, Seth and Jure proposed ConNIe[5] to infer social networks. This method makes up for the limitation of the NETINF algorithm which can only apply to small data sets, and improves the accuracy as well. This method can almost completely restore the multi-node network structure, but the ConNIe algorithm also has shortcomings, requiring a lot of input parameters. In the case that the actual network is not known, it is impossible to know these parameters accurately, which will affect the inferred results greatly.

III. Principle and Parameter Analysis of ConNIe Algorithm

The ConNIe algorithm is a method of maximum likelihood inference based on convex optimization. First, the method establishes a maximum likelihood equation and then the equation convert to convex optimization problem to ensures that the problem can get an optimized solution. Finally, because the social network is sparse, in order to get a sparse solution and improve the computational efficiency, a penalty term is introduced. And it maintains the property of convex optimization as well. This chapter mainly studies and analyzes the ConNIe algorithm, and designs the experimental analysis of the effects of sparse parameter, network propagation time distribution models and their parameters on the inferred results of the ConNIe algorithm.

A. Principle of ConNIe Algorithm

In the ConNIe algorithm, D is used to represent an observable cascade. For each cascade, τ_i^c represents the time at which node i is propagated. if node i has not been propagated in the end, $\tau_i^c = \infty$. $\omega(t)$ is the time distribution model of propagation, representing the time it takes for a message to go from one node to another. If a node i is propagated at time τ_i , it propagates the information to node j with probability A_{ij} . Then $\tau_j = \tau_i + t$,where t obeyed $\omega(t)$. The principle of the ConNIe algorithm is as follows:

$$\min \sum_{c\in D;t_i^c<\infty} -\hat{\gamma}_c - \sum_{c\in D;t_i^c=\infty}\sum_{j\in c;t_j^c<\infty} \hat{B}_{ji} + \rho\sum_{j=1}^{N}\exp\left(-\hat{B}_{ji}\right) \quad (1)$$

subject to

$$\hat{B}_{ji} \le 0 \forall j \quad (2)$$

$$\hat{\gamma}_c \le 0 \forall c \quad (3)$$

$$\log\left[\exp\hat{\gamma}_c + \prod_{j;\tau_j\le\tau_i}\left(1-\omega_j^c+\omega_j^c\exp\hat{B}_{ji}\right)\right] \le 0 \forall c \quad (4)$$

In the above formula, $\omega_j^c \equiv \omega\left(\tau_i^c - \tau_j^c\right)$ 、 $\hat{B}_{ji}=\log(1-A_{ji})$ 、 $\hat{\gamma}_c = \log\left(1-\prod_{j\in X_c\left(\tau_i^c\right)}\left(1-\omega\left(\tau_i^c-\tau_j^c\right)A_{ji}\right)\right)$.

B. Parameters Analysis of ConNIe Algorithm

The influence parameters of ConNIe algorithm include four types: sparse parameter, network model, network scale and propagation time distribution model. In order to explore their respective influences on the inferred results of ConNIe algorithm, the following three sets of experiments are designed in this paper. And we use three evaluation indicators to evaluate the inferred result- precision (the proportion of correct sides to all sides in the inferred result), recall (the proportion of correct sides to all sides in the real network) , mse (mean square error, the ratio of the two norms of the inferred result and the true network to the number of sides of the fully coupled network corresponding to the real network).

(1) The effect of sparse parameter on ConNIe inferred results

In the experiment, the BA network was adopted. Since the features of the part of network and the whole network are similar, the study of the part is similar to the study of the whole. In this paper, a small-scale network is used, in which there are 200 nodes and 1000 edges. The propagation time distribution model is exponential distribution, and the parameter of exponential distribution is 1. The value of ρ is from 0 to 1000 with an interval of 3. The following are the experimental results:

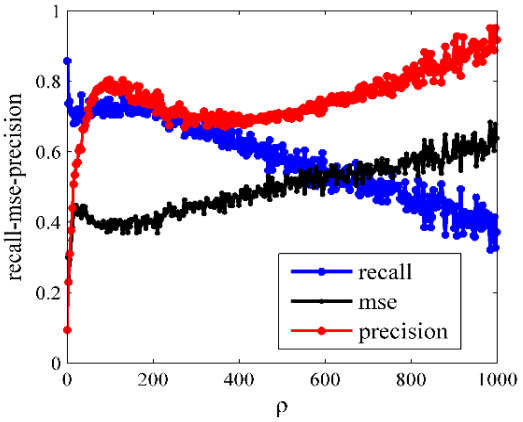

Fig. 1. The influence of sparse parameter ρ

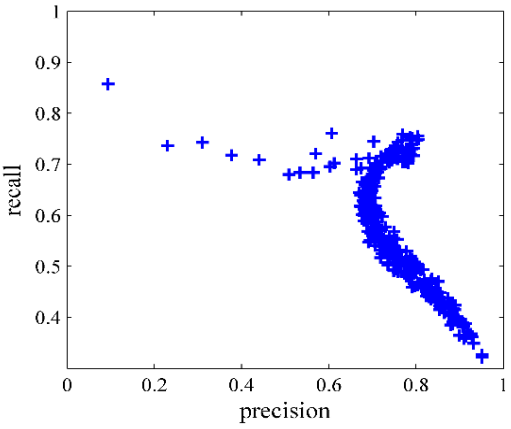

Fig. 2. PR

The better an inferred result is, the bigger its accuracy and recall rate is, and the smaller the mse is. From Fig. 1, it can be concluded that the three evaluation indicators are relatively good at around 100, and relatively good inferred results can be obtained. Fig. 2 is a graph showing the change of the recall along with the precision. The balance point of the PR map is

978-1-7281-4762-8/19/$31.00

the position corresponding to 100. The dense area on the right is recall and precision when $100 < \rho < 1000$. In a wide range, the precision is very high, and the impact of ρ on the recall is more obvious.

(2) The effect of network model and network scale on ConNIe inferred results

Since the two basic characteristics of social networks are small world[9] and scale-free[10],[11], the NW small world network model and the BA scale-free network model are selected for research. For two different networks, the number of nodes is fixed 200. By changing the number of edges, 9 different network scales are selected. And the value of ρ is 500. The propagation time distribution is also an exponential distribution with a parameter of 1.The following are the experimental results:

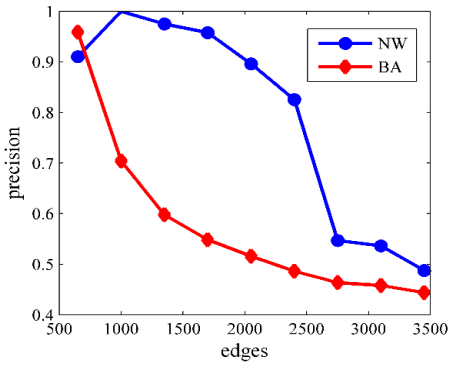

Fig. 3 Precision vs Edges

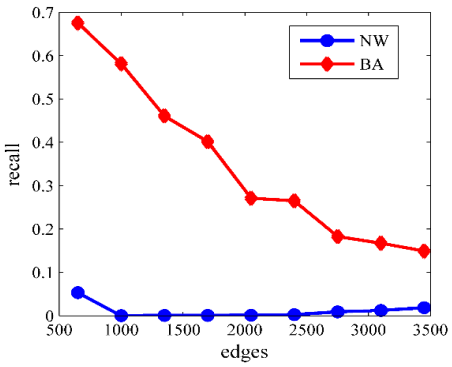

Fig. 4 Recall vs Edges

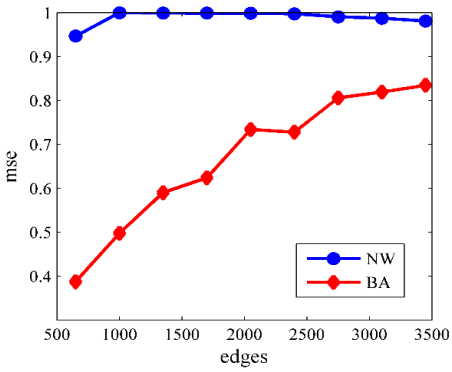

Fig. 5 Mse vs Edges

From the three pictures above, it can be found that for the NW small world network, the recall is zero almost. This indicates that the inferred result is very different from the real and there is almost no intersection. Although the precision is very big, there is no practical significance. It can be seen that the ConNIe inference algorithm is not applicable to small world networks.

For the BA scale-free network, as the network scale increases, the evaluation indicators such as the precision of the inference decrease quickly, and the mse rises rapidly, indicating that the ConNIe inference algorithm is sensitive to the network scale.

(3) The effect of propagation time distribution model on ConNIe inferred results

Because the network propagation time distribution model is needed in the ConNIe algorithm, the propagation distribution models used in this paper is divided into four types: power distribution, exponential distribution, uniform distribution and Weibull distribution. The experiment is divided into eight groups. The first four groups are experiments conducted under the four distribution conditions, when the propagation time distribution model of real data is consistent with the propagation time distribution model for the ConNIe algorithm. The last four groups are experiments when their propagation time distribution models are inconsistent. In the experiment, the BA network was adopted, ρ taking 500, and the network scale is 200 nodes and 1000 edges. The following are the experimental results:

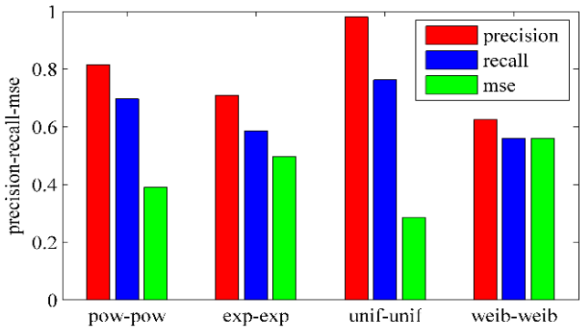

Fig. 6 Consistent in propagation time distribution model

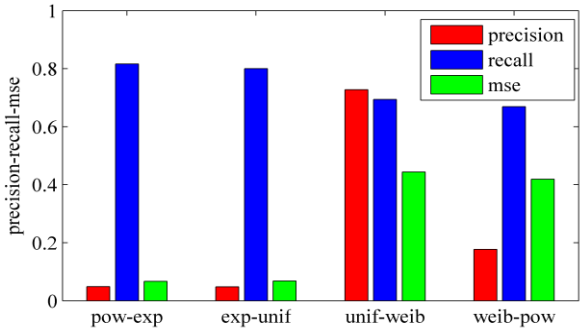

Fig. 7 Inconsistent in propagation time distribution model

Among the evaluation indicators, if any one of the precision and the recall is close to zero, and such inferred results indicate a great difference from the real network. From the comparison between Fig. 6 and Fig. 7, when the propagation time distribution model of the real network is

978-1-7281-4762-8/19/$31.00

inconsistent with the propagation time distribution model for ConNIe algorithm, the inference result is very different from the real network. Although the (uniform distribution - Weibull distribution) inferred results are relatively good, compared to the results of (uniform distribution - uniform distribution), each indicator has deteriorated. Therefore, the propagation time distribution model used in the inference is also a very important factor affecting the inferred results.

IV. Social Network Inference Framework Based on ConNIe Algorithm

In the analysis above, there are four factors that affect the ConNIe algorithm results: sparse parameter, network model, network scale, and propagation time distribution model. However, in the use of ConNIe algorithm, only two parameters need to be determined manually - sparse parameter and social network propagation time distribution model. the propagation time distribution model includes two parameters - a type parameter and a model parameter. The type parameter refers to a distribution type in which the time of information propagation obeys in a social network. This paper assumes four types (power distribution, exponential distribution, uniform distribution, and Weibull distribution). Model parameters refer to parameters in a specific distribution type, such as λ and k in a Weibull distribution. The network scale is included in the propagation data, and the ConNIe algorithm does not need to know the network model. Therefore, before the final inference result is obtained, the data needs to be processed to obtain parameter values that needs to be manually added. The ConNle-based social network inference framework proposed in this paper solves this problem well.

A. Inference Framework

In order to use the ConNIe algorithm to get a more optimized result, this paper proposes a ConNle-based social network inference framework.

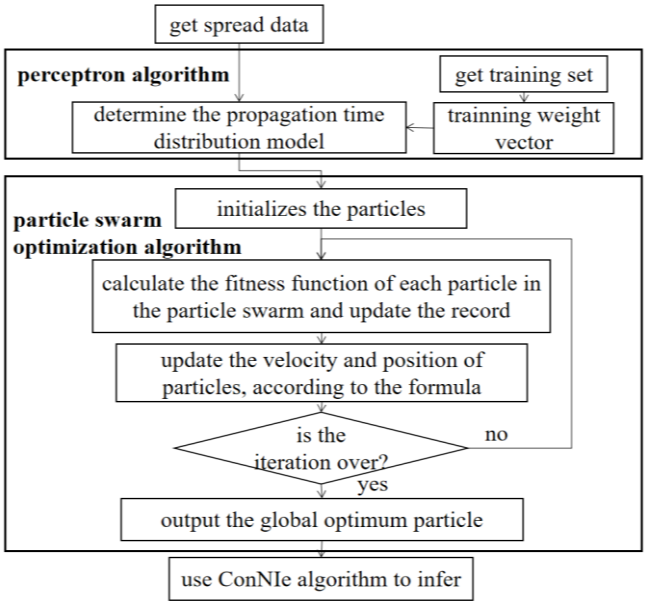

Fig. 8 Inference framework

Because different propagation time distribution models have great influence on the final inference result, the obtained observation data are classified to determine which type of propagation time distribution model belongs to. we preset four propagation time distribution models. By setting a training set of a large number of samples in advance and then using the perceptron algorithm[12], a classifier is obtained. Finally the propagation time distribution type can be identified. After identifying the propagation time distribution type, it is necessary to determine the parameters of the sparse parameter and the propagation time distribution model. Here, the particle swarm optimization algorithm[13] is used to find the optimal combination of parameters, and after the parameter combination is determined, the network inference is performed.

B. Inference of Propagation Time Distribution Model

Based on the observation data to infer, firstly, it is necessary to identify what kind model dose the propagation time distribution subject to. In practical applications, distribution model of the actual data can be determined based on the research on the social network propagation time distribution model. This paper simplifies it. It is assumed that there are four kinds of social network propagation time distribution models: power distribution, exponential distribution, uniform distribution and Weibull distribution. In this paper, the perceptron algorithm is used to classify the data according to its characteristics. Some special treatment is required before classification.

- Pretreatment

Since the location of each data in the original data does not represent the characteristics of the time distribution. In order to eliminate the influence of the position, the original data need to be sorted according to the value of the propagation time. In order to make the feature more obvious, and also to make processed data smaller, it is necessary to perform the difference processing on the sorted data to obtain the difference data. The processing diagram is as follows:

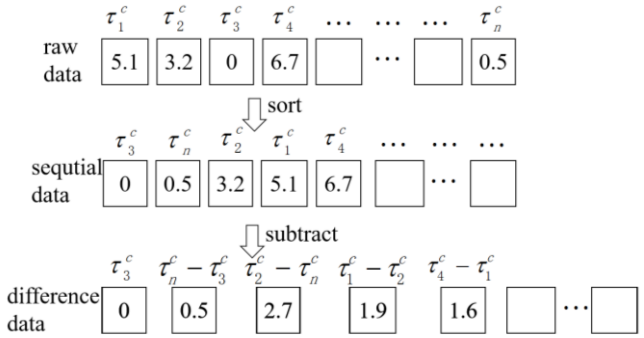

Fig. 9 Pretreatment schematic

The Fig. 9 shows the propagation data of cascade c. For the convenience of display, assume that the minimum of the five data is τ_1^c、 τ_2^c、 τ_3^c、 τ_4^c、 τ_n^c

- Training set settings

The framework proposed in this paper is faced with multiple propagation time distribution models and the training set of the perceptron is very skillful. Since the weight vector of the perceptron algorithm is generated through the training of the training sample set, the quality of the training sample set directly affects the accuracy of the final classification. In order to make the sample set more general, this paper needs to generate a training set of 60,000 samples, including 4 different

networks and 4 different propagation time distribution models and various types of training sets are staggered.

Since the training may not be completely linearly separable, a standard need to be set to allow for how many samples may not satisfy the inequalities required by the perceptron algorithm. After many trainings, it is appropriate to set 1000 tolerance errors in 60,000 training samples. The accuracy of the sample at this time is 98.3% (close to the required 100% standard), and it is also able to train the results.

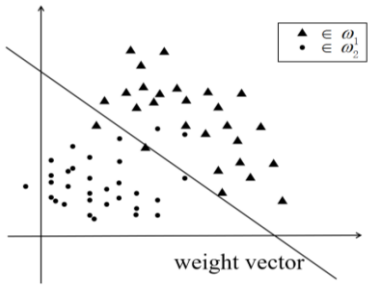

Fig. 10 Classification diagram

The dot above the diagonal line and the triangle below the diagonal line in the above figure are permissible error samples.

C. Parameter Optimization for ConNIe

After identifying the propagation time distribution model, there are two types of parameters to be determined, one is the sparse parameter and the other is the parameter of the time distribution model. Since the inference algorithm is very complicated, it is very difficult to find the optimization parameters from its analytical form. Therefore, the inference is regarded as a black box model, and only the output values corresponding to each group of parameters are known. Therefore, the particle swarm optimization method is used to find the optimal solution.

- Set fitness function

Since the actual network structure is not known in the actual application, the inference result cannot be compared with the original network structure, and the fitness function cannot be constructed according to this idea. Based on the actual propagation data, we calculate the inference network corresponding to each group parameter, and then generate the propagation data according to the inferred propagation time distribution model. Then compare the actual propagation data with the propagation data generated using the inference network. The higher the similarity between the two groups of data, the closer the inference result is to the real network.

$$f(x) = \frac{|c_{real} - c_{infer}|}{u_{real} * u_{infer}} \quad (5)$$

In the above formula, c_{real} represents the pre-processed time when the real social network node receives the information, and c_{infer} represents the pre-processed time when the inferred network node receives the information. u_{real} represents the number of cascades of the real network, and u_{infer} represents the number of cascades of the inferred network. Note that the inferred network here refers to the inferred results with each different group parameters by the ConNIe algorithm, not the final inferred network.

- Optimization Results

As for the effect of the iteration, it is still necessary to judge by the fitness function. The fitness function compares the similarity between the inferred propagation data and the real data. If the similarity is higher, the bigger the function value will be, and the closer the final inferred network is to the true network. The following graph shows the optimal fitness function value for each generation in the 20 iterations.

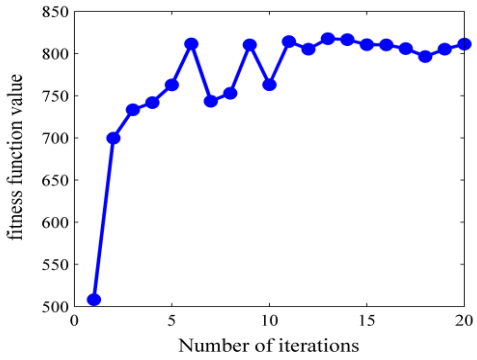

Fig. 11 Optimal for each generation

The optimal fitness function value for each iteration shown in Fig. 11 shows a great effect of particle swarm optimization. Since the first generation is randomly placed particles, the fitness function value is small. The second generation is optimized, so the effect is obvious. It is roughly a trend that rises first and then stabilizes. We can see that after the 11th iteration, the position of each particle is relatively stable and the value of the fitness function is basically unchanged.

But in the 5th to 11th iteration, the fluctuations are obvious. It is not the optimization problem, but the reason of parameters. At the beginning, one or two individual particles have reached the optimal position, but in the next generation, the optimal particle moves to the position of 0.75 due to the inertia parameter. Its fitness function value drops a lot. The other particles do not reach optimality, so the maximum fitness function value of this generation will decline on the basis of the previous generation. After several iterations, most of the particles have gradually reached the optimal position. When the fitness function value of one particle becomes smaller, other particles are in the optimal position, so the optimal value of each iteration does not fluctuate badly.

D. Experiment

In this paper, five groups of inference experiments are designed. The inference results are obtained by inputting the inferred parameters into the ConNIe algorithm. Then the standard results are obtained by using the correct parameters and the predetermined sparse parameter. Finally, compare the two results, and get a reasonable conclusion about this framework. A total of five sets of experiments were set up to compare the effects of different network models, time distribution models and their parameters. In the experiment, the network scale is 200 nodes and 1000 edges. In the inference process of standard results, the default sparse parameter is 100, because no optimization is performed.

1, 2, 3, and 4 of the model represent four propagation time distribution models, respectively, 1 for power distribution, 2 for exponential distribution, 3 for uniform distribution, and 4 for Weibull distribution. The data in the bracket is the standard value, and the data out of the bracket is the inferred value from the framework.

TABLE I Result of Parameter by Framework

	model	ρ	λ	k
BA net Weibull	4 (4)	121	11.1 (9)	2.9 (3)
BA net Weibull	4 (4)	105	8.7 (14)	4.8 (5)
BA net index	2 (2)	249	1.4 (1)	
Stochastic net Weibull	4 (4)	126	11.9 (9)	3.3 (3)
Rule net Weibull	4 (4)	53	8.2 (9)	8.2 (3)

The above data shows that the inference of the network propagation time distribution model is completely correct. This is not because there are only five groups data, but the reason is mainly due to the final result is the one that has the largest number of classification results of these samples. As long as the classification accuracy is greater than 50%, it can be judged correctly. Most of the parameters of the propagation time distribution model obtained by optimization are close to the correct value, but there are also a small part of parameters that are far away from the correct value.

The following table is a comparison of the standard inferred results and framework inferred results.

TABLE II the Comparison of Inferred Result

		precision	recall	mse
BA net Weibull	inferred	0.823	0.944	0.205
	standard	0.899	0.956	0.171
BA net Weibull	inferred	0.824	0.975	0.227
	standard	0.896	0.961	0.165
BA net index	inferred	0.396	0.896	0.415
	standard	0.522	0.893	0.403
Stochastic net Weibull	inferred	0.482	0.594	0.581
	standard	0.649	0.744	0.430
rule net Weibull	inferred	0.391	0.830	0.548
	standard	0.858	0.869	0.260

The inferring line is the result of using the inference framework proposed in this paper. The standard line is the result obtained by using the standard parameters. We can see that the inference of the Weibull distribution on the BA network is quite good and almost completely infer the real network structure.

V. Conclusion

Based on the observed data, this paper analyzes the network inference algorithm. The real network has a larger scale than the simulated network in the experiment, but the two have similar structural features, so the results of the simulation of the network in the experiment can also be applied to the real network. In order to achieve reliable and precise inference, the ConNIe algorithm with high inference precise is selected, which can be applied to the large network scale. The influence of the type of network propagation time distribution model on the inferred results of ConNIe algorithm is studied. Then, the range of sparse parameter and parameters of propagation time distribution models, and the trend of inferred results changed with these parameters are analyzed experimentally. Finally, in order to achieve the optimal inference effect, this paper proposes a social network inference framework based on ConNIe algorithm. The perceptron algorithm is used to classify the network propagation time distribution model, and the classification result can be obtained accurately. The particle swarm optimization algorithm is used to optimize the sparse parameter and the parameters of propagation time distribution models. The difference between optimized parameters value of propagation time distribution models and the real parameters are slight. The experimental results show that the precise of the inference framework is almost consistent with the inference of the known network propagation time distribution parameters, which provides a new method and idea for social network inference.

ACKNOWLEDGMENT

This study is supported by the National Key Research & Development (R & D) Plan under Grant No. 2017YFC1200300, the National Natural Science Foundation of China under Grant Nos. 71673292 and 71673294, the National Social Science Foundation of China under Grant No. 17CGL047, the Beijing National Science Foundation of China under Grant No. 91224006, and the Guangdong Key Laboratory for Big Data Analysis and Simulation of Public Opinion.

References

[1] Liang Liu. Research on simulation model of public opinion information dissemination based on data-driven social network [D]. Changsha: National University of Defense Technology, 2018.

[2] Ghonge S , Vural D C . Inferring Network Structure from Cascades[J]. Physical Review E, 2017, 96(1).

[3] Goldenberg J, Libai B. 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 9 [J]. Monthly Labor Review. 2001, 31 (3): 8–11.

[4] Goldenberg J, Libai B, Muller E. Talk of the network: A complex systems look at the underlying process of word-of-mouth [J]. Marketing letters. 2001, 12 (3): 211–223.

[5] Seth A.Myers, Jure Leskovec, On the Convexity of Latent Social Network Inference, In Advance in Neural Information Processing Systems, 2010.

[6] Granovetter M. Threshold models of collective behavior [J]. American journal of sociology. 1978: 1420−1443.

[7] Psorakis I , Roberts S J , Rezek I , et al. Inferring social network structure in ecological systems from spatio-temporal data streams[J]. Journal of The Royal Society Interface, 2012, 9(76):3055-3066.

[8] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In KDD '10, 2010.

[9] Watts D J,Strogatz S H.Collective dynamics of 'small-world'networks. Nature,19981393(6684):440~442.

[10] Barabasi A L,Albert R,Emergence of scaling in random networks. Science, 1999,286(5439):509~512.

[11] Xiaofan Wang, Xiang Li, Guanrong Chen. Complex Network Theory and Its Application [M]. Beijing: Tsinghua University Press,2006.4:18−28.

[12] Jianwei Liu, Fanglin Shen, Xionglin Luo. Research on Perceptron Learning Algorithm[J]. Computer Engineering,2010,36(07):190-192.

[13] J.Kennedy and R.C. Eberhart. Particale Swarm Optimization. In Proceedings of the IEEE international Joint Conferenec on Neural Networks, pages 1942-1948,1995.