Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Time-series prediction based on global fuzzy measure in social networks*

Li-ming YANG, Wei ZHANG, Yun-fang CHEN†‡

(*Department of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*)

†E-mail: chenyf@njupt.edu.cn

Received Jan. 18, 2015; Revision accepted July 26, 2015; Crosschecked Aug. 25, 2015

Abstract: Social network analysis (SNA) is among the hottest topics of current research. Most measurements of SNA methods are certainty oriented, while in reality, the uncertainties in relationships are widely spread to be overridden. In this paper, fuzzy concept is introduced to model the uncertainty, and a similarity metric is used to build a fuzzy relation model among individuals in the social network. The traditional social network is transformed into a fuzzy network by replacing the traditional relations with fuzzy relation and calculating the global fuzzy measure such as network density and centralization. Finally, the trend of fuzzy network evolution is analyzed and predicted with a fuzzy Markov chain. Experimental results demonstrate that the fuzzy network has more superiority than the traditional network in describing the network evolution process.

Key words: Time-series network, Fuzzy network, Fuzzy Markov chain
doi:10.1631/FITEE.1500025 **Document code:** A **CLC number:** TP393

1 Introduction

As one of the major directions of social network analysis, social network prediction has been thoroughly studied for a long time and its great potential in applications has emerged, drawing tremendous attention from the academics (de Sa and Prudencio, 2011). The recent study is not limited to link between nodes. It also involves the prediction of network global measure. Through the prediction of network global measure, corresponding measures can be taken in advance. The application scenarios have a great range. For example, one can prepare goods that may sell well in advance to get more benefits; the analysis of criminal network prediction can help us understand the social security states.

Most of the prediction methods take the certainty of relationships as a premise. In the real world, however, the social network varies with time, which makes the study on the trend of a time-series network substantial. Network density and centralization are common indicators of global network structure that can be used to analyze the trend of the network diversity over time for a sequential network. Most of the current models of social networks are built upon certain relationship. The relationship between the individuals in a social network is simplified as '1' or '0' for an unweighted graph, where '1' means the existence of a link and '0' means the opposite. More accurate models are the weighted graphs with weighted edges introduced to indicate the closeness of relationships. Traditional network centricity analysis and link prediction methods based on network structure use the deterministic network as well. Nevertheless, in fact, neither the certainty nor the probability methods can reflect the situation objectively. Because of the widespread uncertainty, there could be relation- and

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 61272422 and 61202353)

ORCID: Yun-fang CHEN, http://orcid.org/0000-0002-7897-3588

individual-absence in the real-world network, caused by lack or corruption of data, or dissemblance of privacy (Yan and Gregory, 2011). In addition, the subjectivity of cognition to relationship for different people contributes to the uncertainty. The closeness between two pairs of individuals who have the same amount of connections can be different, and the pairs who connect little on the surface could have potential connection that leads to bias. To handle these issues, a fuzzy system is introduced in this paper: the relations among nodes are fuzzified and the social network density and fuzzy centralization are predicted based on a fuzzy measure. The fuzzy network was proposed in 1965 (Zadeh, 1965) and has been applied mainly to complex and uncertain systems, such as virus propagation, Internet analysis, semantic algorithm analysis, and design of search engines. Then the fuzzy concept was introduced in computer language calculation (Khorasani *et al.*, 2011). The fuzzy network has been used to assess the effect of health warnings on the psychosocial behavior and smoking-cessation behavior of Australian smokers (Zhang *et al.*, 2011).

In our study, we innovatively combine fuzzy network with social network analysis to predict the global variation. The experimental results demonstrate that the fuzzy network has better performance than the certainty network in predicting measures of the time-series network.

2 Related work

Social network analysis is a suite of paradigms and methods that are based on systematic empirical data. The research objects are relations among individuals instead of the intrinsic property of individuals (Freeman, 2004). Relationship models are built to describe the relation between individuals and analyze its inherent structure and impact. The measurements include degree, density, shortcut, and distance, and the degree of connection, which is the most important one, posing them the pivot of social network analysis. A relatively complete theoretical system has been built for social network analysis after years of development, providing several different perspectives on social networks, including centralization analysis, aggregation sub-group analysis, core-periphery structure analysis, and structural equivalence analysis. Centralization analysis is one of the most important social network analyses, where the metric of centrality can be used as an indicator of the importance of a node in the network, by measuring the extent of an individual at the center of the network. Other than computing the centrality of individuals, it can be used to analyze the overall network metrics, such as the calculation of centralization, which reflects the tightness of various nodes connected throughout the network. Larger centralization means greater frequency of connection between the nodes in the network, and it can be calculated based on centrality (Freeman, 1978). Density is a widely applied concept in graph theory. It is also a basic measure and a focus of social network analysis. By calculating the ratio of the actual number of links among individuals to the maximum number of links that may exist among them, the measures reflect the overall closeness of the relationship among the social network (a greater density suggests more links among network members). Network density and network centralization exhibit tightness from the view of quantity and quality, respectively, making them a pair of complementary indicators.

Traditional social network analysis is based on deterministic models, by calculating the value of some measures to analyze the existing relationship in them. However, a simply determined value is not sufficient to describe these relations accurately. Since the fuzzy concept was proposed (Zadeh, 1965), several studies have shown that the introduction of the fuzzy system, using fuzzy state to replace the original value, can well solve the problem of uncertainty in social networks (Nair and Sarasamma, 2007). Brunelli and Fedrizzi (2009) used fuzzy logic to modify the original binary relations into multiple relations among individuals and achieve the same efficiency in social networks. Araujo (2008) presented a method using fuzzy logic to explain social relations. Such methods can improve the flexibility of the relationship between social networks, and thus reduce the individual's conflict. Bastani *et al.* (2013) applied a fuzzy model for link prediction based on network characteristics, and achieved better results than the traditional method. The introduction of a fuzzy model through fuzzy clustering to predict social network links also shows good performance (Ryoke *et al.*, 1995). Recently, some researchers used the ordered weighted averaging (OWA) operator to obtain the fuzzy relationship between nodes

(Brunelli *et al.*, 2014). The method needs some attributes of the network to calculate the relationship; the attributes can be, for example, the similarity index, such as common neighbors (CN), Katz, Salton, and Adamic-Adar (AA) (He *et al.*, 2015). Although the method has the disadvantage of large time complexity, it has advantages of higher prediction accuracy and higher stability. Many methods have been proposed for time-series prediction based on social network analysis. The probabilistic relational model (PRM) based prediction method has been proposed which combines the time-series network with topology (Zhu *et al.*, 2012), and the auto regressive integrated moving average (ARIMA) forecasting model was adopted in a time-series social network (Huang and Lin, 2009). The Markov chain approach has also been used in network link forecasting and analysis methods (Hasan *et al.*, 2006). Yet, these networks are not involved in prior fuzzy processing and cannot model the uncertainty well.

In our paper, through fuzzy processing in time-series social networks, we can achieve better results in predicting the global measure of social networks.

3 Preliminaries

In this section, some basic definitions, symbols, and methods about fuzzy networks are introduced for further discussion on fuzzy-based time-series social network analysis. The first subsection gives some basic concepts on social networks and their measurement, the second subsection displays how fuzzy social networks are built, and the last subsection provides some definitions of the Markov chain for future use.

3.1 Social networks and their measurement

In the traditional social network, network density and community centralization are basic metrics, and they are a couple of complementary indexes. The following definitions are combined with a fuzzy system in Section 4. A traditional undirected social network is defined as an undirected graph $G=(V,E)$, where its vertex set V represents the individuals in the network and edge set E represents the relationships between individuals in the network.

Definition 1 (Network density) The network density D is the ratio of the number of existing edges to the number of maximum possible edges. Given an undirected social network $G=(V,E)$, the ratio is represented as follows:

$$D=\frac{\sum d(v_i)}{N(N-1)/2}, \tag{1}$$

where N is the number of nodes in the social network and $d(v_i)$ is the degree of node v_i.

A larger D means a greater density of network connection. In reality, most complex networks are sparse and have a small density. Thus, through observing the change in network density, the frequency of network links can be analyzed.

Definition 2 (Centrality) The centrality $d(v_i)$ refers to the centrality of node v_i in the social network. It can be defined as

$$C_D(v_i)=d(v_i). \tag{2}$$

A node whose centrality is the largest means that it is the center of the network. If the network consists of N nodes, we define the relative centrality as follows:

$$C'_D(v_i)=\frac{d(v_i)}{N-1}. \tag{3}$$

The maximum degree of any node in the network is $N-1$. A larger centrality means that the node is connected with more direct links and is comparatively important.

Definition 3 (Network centralization) The network centralization C_D reflects the concentration of the whole network; the formula of centralization can be defined using the network centrality. It can be defined with relative centrality as follows:

$$C_D=\frac{\sum_{i=1}^{n}(C_{D\max}-C_{Di})}{N-1}, \tag{4}$$

where $C_{D\max}$ is the node that has the largest centrality in the network.

For example, the networks in Figs. 1a and 1b have the same network density ($D=0.5$), and $C_D(A)$=0 and $C_D(B)$=2.5. It can be seen that a sparse network's density is less and a balanced network has a small centralization. On the contrary, the more the links that focus on some nodes of the network center, the larger the centralization. So, an unbalanced network usually has a large centralization. Density can show the frequency of communication between individuals in the network, while centralization shows the degree of closeness in network

relation. A network that has a higher concentration usually has a larger centralization.

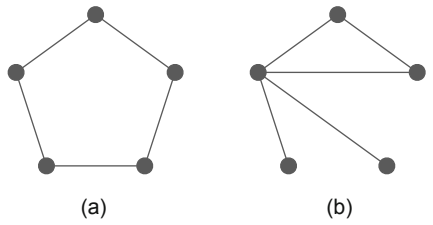

Fig. 1 Difference between a balanced network (a) and an unbalanced network (b)

3.2 Fuzzy social network

We use the similarity algorithm as the fuzzy method of the network. As shown in Fig. 2, we use the Jaccard similarity algorithm (Jaccard, 1901) to calculate the relationship between nodes A and B. There is no link between nodes A and B in the traditional network; however, the Jaccard similarity algorithm considers the influence of neighbors and we can obtain fuzzy relations between nodes A and B.

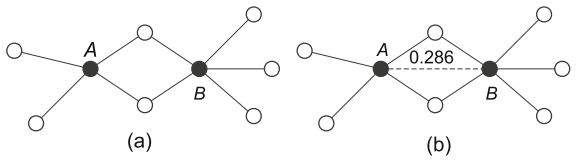

Fig. 2 Difference between a traditional network (a) and a fuzzy network (b)

Definition 4 (Relation membership function) The relation membership function is used to transform a traditional network into a fuzzy network, which is defined as follows:

$$\mu(x) = S_{v_i v_j}, \tag{5}$$

where $\mu(x) \in [0,1]$ is the membership function and $S_{v_i v_j}$ the similarity between nodes v_i and v_j.

We use the Jaccard similarity algorithm in this study. Let $\Gamma(x)$ denote the neighbor set of node x. Then the fuzzy relationship between nodes can be defined as follows:

$$S_{v_i v_j}^{\text{Jaccard}} = \frac{\Gamma(v_i) \cap \Gamma(v_j)}{\Gamma(v_i) \cup \Gamma(v_j)}. \tag{6}$$

The complexity of the fuzzification method largely influences the performance of time-series prediction in our study. To calculate the Jaccard similarity metric, it is always needed to traverse its adjacent nodes. So, its time complexity is $O(n^2)$ using the common traversal method or $O(n \log n)$ using the sort traversal method. If the number of nodes N is larger, it costs much longer time to obtain the metric. Original networks are generally sparse. However, they will become denser after being fuzzlized and the trend may affect the complete graph. As $S_{v_i v_j}^{\text{Jaccard}} \in [0,1]$, we can leave out the normalization process.

Definition 5 (Fuzzy relation) The fuzzy relation $\tilde{e}_{ij}$ can be calculated based on the traditional network relation e_{ij} through $\mu(x)$, which is defined as follows:

$$\tilde{e}_{ij} = \mu(e_{ij}). \tag{7}$$

According to Definition 4 (Eq. (5)), the fuzzy relation $\tilde{e}_{ij}$ can also be defined as follows:

$$\tilde{e}_{ij} = S(v_i v_j). \tag{8}$$

A larger $S_{v_i v_j}$ means a higher possibility of relation and more contact between nodes v_i and v_j. This kind of fuzzy algorithm is quite simple and follows people's perception of social networks. If we cannot know whether two people have a friendly relationship through the traditional network but we know they have many common friends, then we can guess that there may be a certain relationship between them (Jin *et al.*, 2001; Ebel *et al.*, 2002).

Definition 6 (Fuzzy undirected network) The fuzzy undirected network $\tilde{G} = (V, \tilde{E})$ is transformed from the traditional network $G = (V, E)$ through fuzzification. $V = \{v_1, v_2, ..., v_n\}$ is the set of individuals in the fuzzy network and $\tilde{E} = \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{E}(e_{ij})$ is the fuzzy relationship between them.

3.3 Definition of the Markov chain

Definition 7 (Markov chain) Let x be a random variable sequence, which may be in states $S_1, S_2, ..., S_r$. $X_n = i$ means that it is in state S_i at time n. It means that

$$p_{ij}(n) = \text{Prob}(X_{n+1} = j | X_n = i),\ 1 \le i, j \le r. \tag{9}$$

It shows the step transition probability in the system, if

$$\text{Prob}(X_{n+1} | X_n X_{n-1} \dots X_1) = \text{Prob}(X_{n+1} | X_n), \tag{10}$$

and if the random process of the state and the parameter of time are discrete, we can call the stochastic process a Markov chain.

4 Prediction of measure in time-series networks

4.1 Density and centralization in fuzzy networks

Consider an undirected fuzzy social network $\tilde{G}$ as defined in Definition 6. $\tilde{d}(v_i)=\sum_{j=1}^{n}\mu(e_{ij})$ is the sum of $\tilde{e}_{ij}$'s that have a relation with node v_i. In a fuzzy network, fuzzy network density reflects the degree of closeness in the fuzzy network. According to the traditional network density formula, the fuzzy network density formula can be defined as follows:

$$\tilde{D}=\frac{\sum\tilde{d}(v_i)}{N(N-1)/2}. \tag{11}$$

Fuzzy centralization can be displayed using fuzzy centrality. The fuzzy centrality of the individuals in the fuzzy network directly connects to the node with the edge of the sum $\tilde{d}(v_i)$ of membership degree $\mu(x)$:

$$\tilde{C}_D(n_i)=\frac{\tilde{d}(v_i)}{N-1}. \tag{12}$$

According to the traditional network community centralization formula, the fuzzy centralization formula can be defined as follows:

$$\tilde{C}_D=\frac{\sum_{i=1}^{n}(\tilde{C}_{D\max}-\tilde{C}_{Di})}{N-1}. \tag{13}$$

4.2 Prediction based on the fuzzy Markov chain

We use the fuzzy Markov chain to predict the measure, which is a fuzzification of the traditional Markov chain model. Many research objects have a complex inner relationship in the real world, which means the relationship is fuzzy. The unfollow-up effects between individuals are not strict. The traditional Markov chain analysis prediction methods cannot accurately reveal the characteristics of the object. We combine fuzziness with the Markov chain to ensure the accuracy of the algorithm model. In the traditional Markov chain, each transition probability of p_{ij} needs to be determined. In actual life, in many cases, these values can only be estimated or provided by the experts.

4.2.1 Establishment of the fuzzy Markov chain model

A fuzzy Markov chain model can be established by the following four steps:

1. Determining the fuzzy state of the system

Set the range of X_t as U according to the application background in the first place. Establish the fuzzy state for U as $U=\{\tilde{A}_1,\tilde{A}_2,\dots,\tilde{A}_r\}$. Let $\mu_{\tilde{A}_i}(x)$ be the membership function of state A_i and assume that $\{x_1,x_2,\dots,x_n\}$ is the observed data, where x_t is the observation at time t. We need to make sure that the amount of observed data n is greater than the number of fuzzy states r.

2. Calculating the initial probability

For each observation, the observation frequency of each fuzzy state $\{\tilde{A}_1,\tilde{A}_2,\dots,\tilde{A}_r\}$ needs to be calculated:

$$m_i=\sum_{t=1}^{n-1}\mu_{\tilde{A}_i}(x_t). \tag{14}$$

Thus, the initial probability for fuzzy state $\tilde{A}_i$ can be represented as $p_i=m_i/(n-1)$, $i=1,2,\dots,n$.

3. Calculating the first-order state transfer matrix

For the transformation of $\tilde{A}_i\to\tilde{A}_j$ from t to $t+1$, the probability can be represented as $\mu_{\tilde{A}_i}(x_t)\cdot\mu_{\tilde{A}_j}(x_{t+1})$ and the frequency of the event is $m_{ij}=\sum_{t=1}^{n-1}\mu_{\tilde{A}_i}(x_t)\cdot\mu_{\tilde{A}_j}(x_{t+1})$. We obtain the transfer probability of $\tilde{A}_i\to\tilde{A}_j$ from the whole time-series system as $p_{ij}=m_{ij}/m_i$, and the first-order state transfer matrix is $\boldsymbol{P}=[p_{ij}]_{n\times n}$.

4. Prediction

Assume that the observation value at time t is x. Then the membership degree of each state in U for x is $\boldsymbol{F}(x_t)=(\mu_{\tilde{A}_1}(x),\mu_{\tilde{A}_2}(x),\dots,\mu_{\tilde{A}_r}(x))$ and the membership degree at time $t+1$ is as follows:

$$\boldsymbol{F}(x_{t+1})=\boldsymbol{F}(x_t)\cdot\boldsymbol{P}^{\mathrm{T}}. \tag{15}$$

Finally, the transformation of the fuzzy state at $t+1$ can be judged according to the maximum membership principle. Based on the actual situation, if the maximum membership degree and the second largest membership degree have a slight deviation, the second largest state or even the third largest membership can be listed as a possible transfer target.

4.2.2 State division and membership function

State division is very important in fuzzy Markov chain prediction since the number of division states, the range of states, and the choice of the membership function will affect the accuracy of the prediction. We divide data into fuzzy states $S_1, S_2, \ldots, S_r$. The prediction result can be more accurate if fuzzy states have larger amount and shorter range. However, it requires a lot of systematic observation data $\{x_1, x_2, \ldots, x_n\}$ with the number of samples being greater than r. A small number of samples will lead to an inaccurate prediction. So, we should choose an appropriate number of states. During state division, the choice of membership function $\mu_{\tilde{A}_i}(x)$ is also very critical. One chooses the membership function based on experience. Without any experience, we need another method to choose an appropriate fuzzy distribution. Generally, we choose the following kinds of functions: trapezoidal distribution, normal distribution, Cauchy distribution, and triangular distribution. Here we choose trapezoidal distribution and Cauchy distribution. There are also some parameters in the membership function, and we can adjust these parameters in the process of an experiment to achieve the best prediction accuracy.

4.3 Relevant algorithm

At first, we divide the data into a time-series network structure $\{G_1, G_2, \ldots, G_n\}$ composed of n time slices. Then we use the Jaccard similarity metric to fuzzify the time-series network $G_t(V_t, E_t)$:

$$E_t = S^{\text{Jaccard}}_{v_i v_j},\ v_i, v_j \in G_t. \tag{16}$$

Therefore, we obtain the fuzzy network $\{\tilde{G}_1, \tilde{G}_2, \ldots, \tilde{G}_n\}$, and can calculate the fuzzy density $\tilde{D}$ and fuzzy centralization $\tilde{C}_D$ by the definition in Section 3.2. Then we need to divide the metrics into fuzzy states $S_1, S_2, \ldots, S_r$. We set the data $\{x_1, x_2, \ldots, x_n\}$ with fuzzy density and fuzzy centralization. Afterwards, we prepare to predict the data x_{n+1} ($r \ll n+1$). First, we divide the data $\{x_1, x_2, \ldots, x_n\}$ into fuzzy states. The standard of dividing data is based mainly on experience and the existing criteria or other division methods. Second, we use the sample mean value-standard deviation grade method in this paper, and the interval of the fuzzy metric can be depicted by the sample means and sample standard deviation. With data $\{x_1, x_2, \ldots, x_n\}$, we can obtain the sample average $\bar{x}$ and sample standard deviation σ, which can be divided into r states. Their state intervals are $(-\infty, x-\alpha_1\sigma), (x-\alpha_1\sigma, x-\alpha_2\sigma), \ldots, (x-\alpha_m\sigma, x+\alpha_m\sigma), \ldots, (x+\alpha_2\sigma, x+\alpha_1\sigma), (x+\alpha_1\sigma, +\infty)$, where $m = r - 1/2$ and $\alpha_1 < \alpha_2 < \ldots < \alpha_m$.

The observed data should be approximately a sequence of independent and identically distributed variables. According to the central limit theorem which shows it is a weak correlation sequence (the correlation coefficient is less than 0.3), we have

$$P\left\{\bar{X} - 1.5\sigma \le x \le \bar{X} + 1.5\sigma\right\} > 0.86.$$

Considering the amount of sample data, in this paper we use $r = 5$, which is divided into five fuzzy states. We set a step parameter θ in this study, whose states are $A_1, A_2, \ldots, A_5$. The state intervals can be $(-\infty, x-\alpha_1\sigma+\theta), (x-\alpha_1\sigma, x-\alpha_2\sigma+\theta), (x-\alpha_2\sigma, x+\alpha_2\sigma+\theta), (x+\alpha_2\sigma, x+\alpha_1\sigma+\theta)$, and $(x+\alpha_1\sigma, +\infty)$.

We set $\alpha_1 = 1.0$ and $\alpha_2 = 0.5$, and use the trapezoid membership function. The left and right sides of the function are open, i.e., semi-trapezoid, while the rest of the function is completed.

First, we calculate the average and standard deviation of the observation data to construct the membership function of different states:

$$\mu_{A_1}(x) = \begin{cases} 1, & 0 \le x \le \bar{x} - \sigma, \\ \dfrac{\bar{x} - \sigma + \theta - x}{\theta}, & \bar{x} - \sigma < x < \bar{x} - \sigma + \theta, \\ 0, & \text{otherwise}, \end{cases}$$

$$\mu_{A_2}(x) = \begin{cases} \dfrac{x - (\bar{x} - \sigma)}{\theta}, & \bar{x} - \sigma < x < \bar{x} - \sigma + \theta, \\ 1, & \bar{x} - \sigma + \theta \le x \le \bar{x} - 0.5\sigma, \\ \dfrac{(\bar{x} - 0.5\sigma) + \theta - x}{\theta}, & \bar{x} - 0.5\sigma < x < \bar{x} - 0.5\sigma + \theta, \\ 0, & \text{otherwise}, \end{cases}$$

$$\mu_{A_3}(x) = \begin{cases} \dfrac{x - (\bar{x} - 0.5\sigma)}{\theta}, & \bar{x} - 0.5\sigma < x < \bar{x} - 0.5\sigma + \theta, \\ 1, & \bar{x} - 0.5\sigma + \theta \le x \le \bar{x} + 0.5\sigma, \\ \dfrac{(\bar{x} + 0.5\sigma) + \theta - x}{\theta}, & \bar{x} + 0.5\sigma < x < \bar{x} + 0.5\sigma + \theta, \\ 0, & \text{otherwise}, \end{cases}$$

$$\mu_{A_4}(x)=\begin{cases}\dfrac{x-(\bar{x}+0.5\sigma)}{\theta}, & \bar{x}+0.5\sigma<x<\bar{x}+0.5\sigma+\theta,\\ 1, & \bar{x}+0.5\sigma+\theta\le x\le\bar{x}+\sigma,\\ \dfrac{(\bar{x}+\sigma)+\theta-x}{\theta}, & \bar{x}+\sigma<x<\bar{x}+\sigma+\theta,\\ 0, & \text{otherwise},\end{cases}$$

$$\mu_{A_5}(x)=\begin{cases}1, & \bar{x}+\sigma+\theta\le x,\\ \dfrac{x-(\bar{x}+\sigma)}{\theta}, & \bar{x}+\sigma<x<\bar{x}+\sigma+\theta,\\ 0, & \text{otherwise}.\end{cases}\tag{17}$$

We can obtain the fuzzy state classification matrix $\tilde{\boldsymbol{E}}$ with the observed data $\{x_1,x_2,\ldots,x_n\}$:

$$\tilde{\boldsymbol{E}}=\begin{bmatrix}\mu_{A_1}(x_1) & \mu_{A_1}(x_2) & \cdots & \mu_{A_1}(x_n)\\ \mu_{A_2}(x_1) & \mu_{A_2}(x_2) & \cdots & \mu_{A_2}(x_n)\\ \vdots & \vdots & & \vdots\\ \mu_{A_5}(x_1) & \mu_{A_5}(x_2) & \cdots & \mu_{A_5}(x_n)\end{bmatrix}.$$

The prediction results can be obtained according to Section 4.2.1, including the predicted membership degree and the true value. To discern from the true value x_n, we denote the result value as x_n^*. We can predict the results, which can be obtained from

$$\boldsymbol{E}_n^*=(\mu_{A_1}(x_n^*),\mu_{A_2}(x_n^*),\mu_{A_3}(x_n^*),\mu_{A_4}(x_n^*),\mu_{A_5}(x_n^*)),$$

The true membership degree can be obtained by putting the true value x_n into the membership function:

$$\boldsymbol{E}_n=(\mu_{A_1}(x_n),\mu_{A_2}(x_n),\mu_{A_3}(x_n),\mu_{A_4}(x_n),\mu_{A_5}(x_n)).$$

Finally, the value $\mu_{A_i}(x_n)$ has the largest membership degree, and for $\mu_{A_i}(x_n^*)\in\boldsymbol{E}_n^*$, we use deviation $\delta=|\mu_{A_i}(x_n)-\mu_{A_i}(x_n^*)|/\mu_{A_i}(x_n)$ $(0\le x\le 1)$ to measure the accuracy. If δ is approximately 0, the prediction is accurate. To forecast more steps, the predicted value x_n^* needs to be introduced and the same method is used to calculate the prediction value based on the original data. In the current study, the median of the maximum membership state is used as the value of x_n^*.

5 Experiments

Email data from Enron Corporation is used in our study. We choose the data of 2001 and then divide it into 12 time slices by month and fuzzify these networks for experiment. In addition, a 10-day phone call record network is used in the experiment. It is divided into 10 time slices by day.

5.1 Prediction of density and centralization in the fuzzy time-series network

To predict fuzzy centralization and density in the time-series network, first we calculate the centralization and density in the fuzzy network slices that have been known through the Jaccard similarity algorithm. The differences in density and centralization between the traditional network and fuzzy networks using the email data from Enron Corporation and a 10-day phone call record network are illustrated in Figs. 3–6.

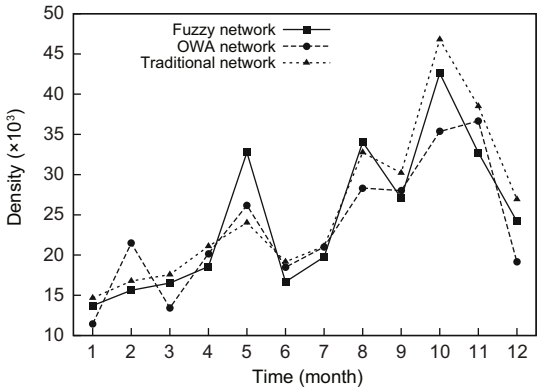

Fig. 3 The difference in density among the traditional network, fuzzy network, and OWA network (Enron)

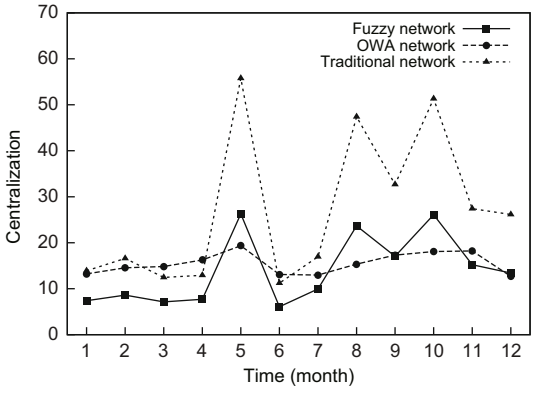

Fig. 4 The difference in centralization among the traditional network, fuzzy network, and OWA network (Enron)

We use the data from Jan. to Oct. of a complete traditional network and a fuzzy network as the observed values to predict density. We obtain

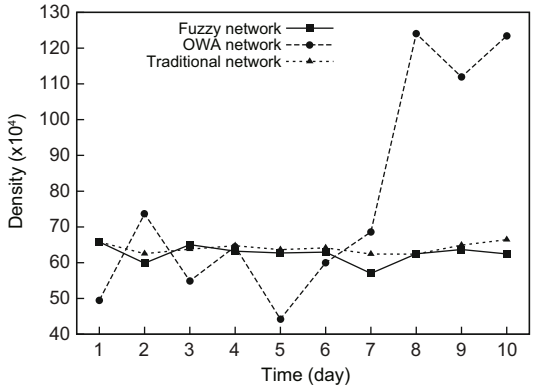

Fig. 5 The difference in density among the traditional network, fuzzy network, and OWA network (10-day call)

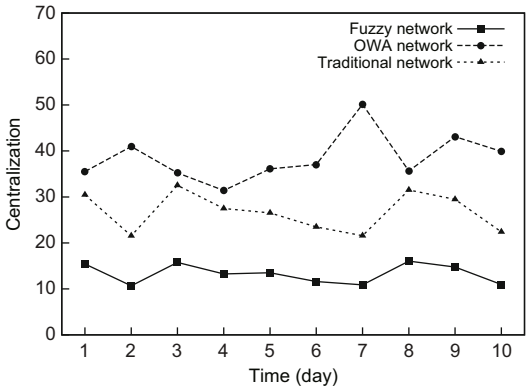

Fig. 6 The difference in centralization among the traditional network, fuzzy network, and OWA network (10-day call)

$\bar{x} = 25.79$, $\sigma = 9.70$, and the real value $x_{11} = 38.50$ in the traditional network, whereas $\bar{x} = 24.52$, $\sigma = 9.23$, and the real value $x_{11} = 32.73$ in the fuzzy network. We use the trapezoidal membership function and set step length $\theta = \alpha\bar{x}$. Let $\alpha = 0.05$, 0.10, 0.15, 0.20, and 0.25 respectively, and then calculate the deviation value using the real value. For example, when $\alpha = 0.05$ is set and the data of the fuzzy network for experiment is used, we can obtain the degree of each time slice network membership for different states by establishing the fuzzy Markov chain model. The membership states are shown in Table 1.

After that, calculate the transfer matrix $\boldsymbol{P}^{\mathrm{T}}$:

$$\boldsymbol{P}^{\mathrm{T}} = \begin{bmatrix} 0.420 & 0.579 & 0.500 & 0 & 0 \\ 0 & 0.532 & 0 & 0.412 & 0.054 \\ 0 & 0 & 0 & 0 & 1.000 \\ 0 & 0.361 & 0.638 & 0 & 0 \\ 0 & 0 & 0.190 & 0.810 & 0 \end{bmatrix}.$$

Table 1 The membership status table

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.
μ_{A_1}	1	0.72	0	0	0	0	0	0	0	0
μ_{A_2}	0	0.28	1	1	0	1	1	0	0	0
μ_{A_3}	0	0	0	0	0	0	0	0	1	0
μ_{A_4}	0	0	0	0	1	0	0	0.77	0	0
μ_{A_5}	0	0	0	0	0	0	0	0.23	0	1

Then we can calculate the degree of membership of each state:

$$\tilde{\boldsymbol{E}}_n^* = (0 \quad 0 \quad 0.190 \quad 0.810 \quad 0).$$

The ranges of membership states $A_1, A_2, ..., A_5$ are $(-\infty, 16.5)$, $(15.3, 21.1)$, $(19.9, 30.3)$, $(29.1, 34.9)$, and $(33.7, +\infty)$, respectively.

The largest membership value $\mu_{A_4}(x_{11}^*) = 0.810$ indicates that x_{11}^* is most likely in state A_4 and the range of A_4 is $(29.1, 34.9)$. We can obtain the real value of Nov. as $x_{11} = 32.73$ from the original data. The value is within the scope of A_4, and this means that the prediction is accurate. The real degree of membership of each state is as follows:

$$\tilde{\boldsymbol{E}}_n = (0 \quad 0 \quad 0 \quad 1 \quad 0).$$

Then we can calculate the deviation value $\tilde{\delta} = 0.190$ and the degree of the traditional network with the same algorithm:

$$\tilde{\boldsymbol{E}}_n^* = (0 \quad 0 \quad 0.500 \quad 0 \quad 0.500).$$

The ranges of membership states $A_1, A_2, ..., A_5$ are $(-\infty, 17.3)$, $(16.1, 22.2)$, $(20.9, 31.9)$, $(30.6, 36.7)$, and $(35.5, +\infty)$, respectively.

The largest membership value $\mu_{A_5}(x_{11}^*) = 0.500$ indicates that x_{11}^* is most likely in state A_5 and the range of A_5 is $(35.5, +\infty)$. We can obtain the real value of Nov. as $x_{11} = 38.50$ from the original data. The value is within the scope of A_5, and this means that the prediction is accurate. The real degree of membership of each state is as follows:

$$\tilde{\boldsymbol{E}}_n = (0 \quad 0 \quad 0 \quad 0 \quad 1).$$

Then we can calculate the deviation value $\tilde{\delta} = 0.5$. Analogously, we calculate the deviation values when $\alpha = 0.10$, 0.15, 0.20, and 0.25, respectively (Tables 2 and 3).

Apparently $\tilde{\delta} < \delta$, and the range is 5.8 in the fuzzy network but 6.1 in the traditional network under the same conditions. The statistics show that

Table 2 The prediction of density in the fuzzy network

Step length α	Maximum status	Range	Deviation δ
0.05	A_4	29.1–39.8	0.190
0.10	A_4	29.1–38.6	0.105
0.15	A_4	29.1–37.4	0.072
0.20	A_4	29.1–36.2	0.055
0.25	A_4	29.1–34.9	0.044

Table 3 The prediction of density in the traditional network

Step length α	Maximum status	Range	Deviation δ
0.05	A_5	<35.5	0.500
0.10	A_5	<35.5	0.500
0.15	A_5	<35.5	0.436
0.20	A_5	<35.5	0.367
0.25	A_4	30.6–41.9	0.317

using the fuzzy network for prediction is more accurate than using the traditional network.

The prediction of centralization is the same. We can obtain $\bar{x} = 26.69$, $\sigma = 4.23$, and the real value $x_{11} = 29.48$ in the traditional network. Moreover, we can obtain $\bar{x} = 14.04$, $\sigma = 7.62$, and the real value $x_{11} = 15.21$ in the fuzzy network. We use the trapezoidal membership function and set step length $\theta = \alpha\bar{x}$. Let $\alpha = 0.05$, 0.10, 0.15, 0.20, and 0.25, respectively. Then we calculate the deviation value δ from the real value.

Through Fig. 7, we know that with the same parameter, the deviations of the prediction obtained through metrics of the fuzzified network are smaller than those of the traditional network in terms of both centralization and network density. This indicates that utilization of the fuzzy Markov chain would result in a more accurate result in the fuzzy network than in the traditional network in predicting centralization and density. In addition, we use a 10-day phone call network for this experiment. As Fig. 8 shows, most values of the metric predicted in the fuzzy network have a smaller deviation than in the traditional network. This means that the fuzzy method can improve the accuracy of metric prediction in a different sense.

Similarly, we use the fuzzy network based on the OWA method to do the experiment. As Figs. 3 and 4 show, we select some attributes such as CN, Katz, Salton, RA, and PA. We can calculate the deviation values using the same method (Fig. 9).

We know that the fuzzy network based on the OWA method can also have a good performance in prediction. However, it costs much more time.

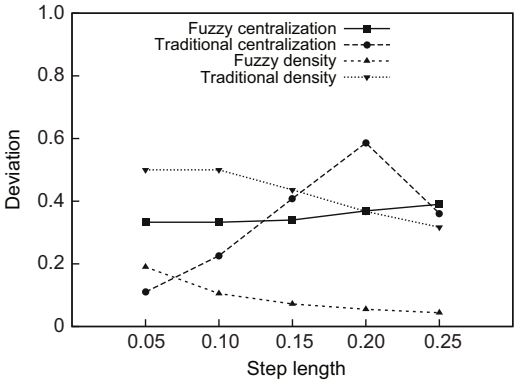

Fig. 7 The difference in deviation between the traditional network and the fuzzy network (Enron)

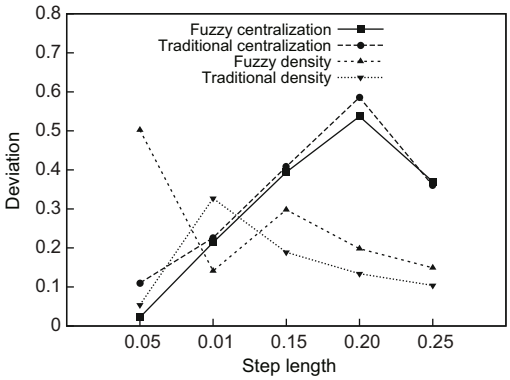

Fig. 8 The difference in deviation between the traditional network and the fuzzy network (10-day call)

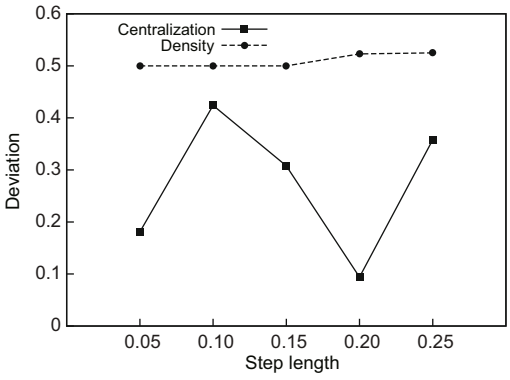

Fig. 9 The difference in deviation in the OWA network

5.2 Impact on prediction results with varied fuzzy parameters

The membership function $\mu_A(x)$, which is used to divide the state, is the pivot of the prediction method. We examine the impact of the membership function by adjusting the parameters. In addition to

trapezoidal function, the Cauchy function is introduced here as a membership function:

$$\mu_{A_i}(x) = \frac{1}{1+\alpha(x-a)^{\beta}},$$

where a can be $\pm 0.5\sigma \sim \pm 1.5\sigma$ according to the sample-standard deviation grading method, and β is an even number. Take the fuzzy network data from Jan. to Oct. as observations, and let $\alpha = 0.04$ and $\beta = 2$. By predicting the network density, we obtain the predicted membership degree as follows:

$$\tilde{\boldsymbol{E}}_n^* = (0.001 \quad 0.003 \quad 0.004 \quad 0.008 \quad 0.038).$$

The true degree of membership for each membership state is as follows:

$$\tilde{\boldsymbol{E}}_n = (0.011 \quad 0.017 \quad 0.022 \quad 0.030 \quad 0.065),$$

where the predicted maximum membership degree is 0.038 and the true maximum membership degree is 0.065, both at state A_5. The bias value is $\delta = 0.544$.

Let $\alpha = 0.004$ and $\beta = 2$. The predicted membership degree by predicting the network density is as follows:

$$\tilde{\boldsymbol{E}}_n^* = (0.109 \quad 0.177 \quad 0.231 \quad 0.306 \quad 0.489).$$

The true degree of membership for each membership state is as follows:

$$\tilde{\boldsymbol{E}}_n = (0.103 \quad 0.152 \quad 0.189 \quad 0.240 \quad 0.412).$$

In the above results, we have the predicted maximum membership degree 0.489 and the true maximum membership degree 0.412, with both at state A_5, which is also accurate. The bias value is $\delta = 0.009$.

The experimental results indicate that under the same conditions, trapezoidal distribution outperforms Cauchy distribution in terms of accuracy. Besides, the uncertainty of the state range is a flaw when used here. In addition to using data from Jan. to Nov. to forecast the Dec. data, we use data from Jan. to Oct. to predict the Nov. data. Then the data is combined with the predicted data from Nov. 1 to Oct. to predict the Dec. data. Then deviation is calculated. When $\alpha = 0.05$, the predicted maximum degree of membership is in a limited range. We can obtain its state interval. Its membership is obtained as follows:

$$\tilde{\boldsymbol{E}}_n^* = (0 \quad 0 \quad 0.190 \quad 0.810 \quad 0).$$

x_{11}^* is most likely in state A_4. The range of membership state A_4 is $(29.1, 34.9)$, taking the number of intermediate states as 32, along with data from Jan. to Sept. to forecast Nov. data. Also, let $\alpha = 0.05$, 0.10, 0.15, 0.20, and 0.25, respectively. The deviation value δ with respect to the true value is shown in Fig. 10.

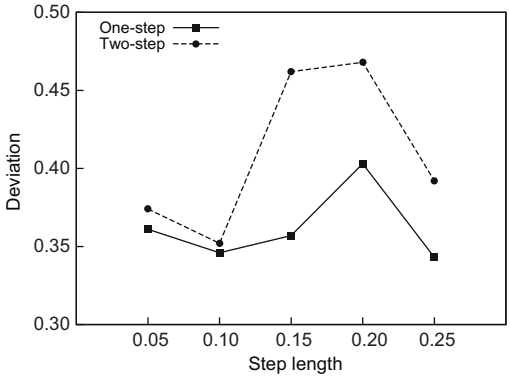

Fig. 10 The difference in deviation with different steps

Fig. 10 shows that under the same conditions, the deviation to use the two-step fuzzy Markov chain prediction is slightly larger than that of the one-step method, but the real value falls into the predicted maximum membership degree range, which means that the prediction is accurate. The results show that although the time range of samples is limited, the use of multi-step prediction in the fuzzy network can still maintain a certain degree of accuracy and choosing a reasonable state interval can increase accuracy.

5.3 Network size and time span effect on prediction accuracy

To verify the impact of accuracy with different sizes, we sample fuzzy network individuals of 80%, 50%, 25%, respectively, 10 times each, and calculate the average of their fuzzy density and fuzzy centralization. The results are shown in Figs. 11–13.

The prediction bias δ is computed as shown in Fig. 13. It can be found from the experimental results that the reduction in data scale does not cause the deviation δ to become larger, and the prediction is accurate. Furthermore, to verify the influence of the selection of different time frames on the final prediction, we observe the raw data. It shows that fuzzy density and centralization both have a significant increase in May and June when they change back to

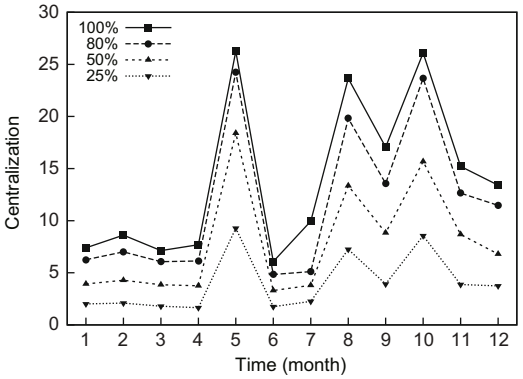

Fig. 11 The centralization at different network sizes

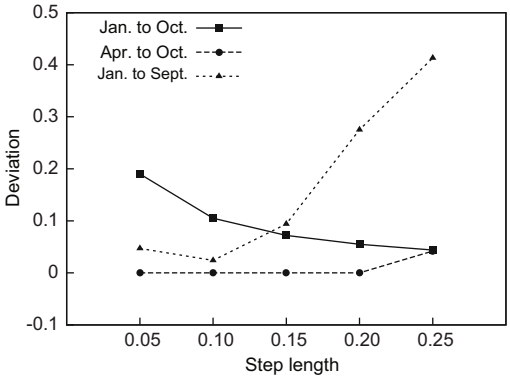

Fig. 14 The deviation at different time spans

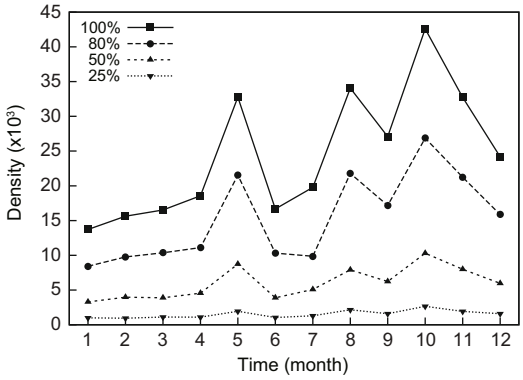

Fig. 12 The density at different network sizes

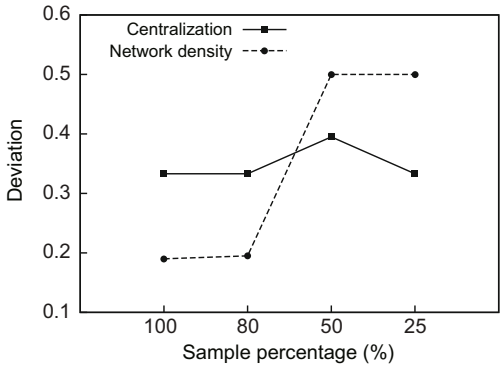

Fig. 13 The deviation at different sample percentages

the original scope. If we select Apr. to Oct. data to predict Nov. data, the time sheet data that is no longer than that from Jan., the data fluctuates less within its scope. The average density from Apr. to Oct. is $\bar{x} = 27.35$ and the standard deviation is $\sigma = 9.62$. Using the trapezoidal membership function and taking $\theta = \alpha\bar{x}$ and $\alpha = 0.05$, 0.10, 0.15, 0.20, and 0.25 respectively, the results are as shown in Fig. 14.

The true value $x_{11} = 32.73$ is within the largest range of membership states of the predicted value x^*_{11}, and the deviation δ between them is smaller in a small state range, for example, when $\alpha = 0.05$ or $\alpha = 0.10$. This shows that although the number of samples is reduced, the accuracy does not decrease. We can exclude some of the data and choose a suitable state range to increase the prediction accuracy. In addition, we select Jan. to Sept. data to predict the Oct. data, with the average centralization $\bar{x} = 24.55$ and standard deviation $\sigma = 9.69$. Taking $\alpha = 0.05$, 0.1, 0.15, 0.2, and 0.25, we obtain the results as shown in Fig. 10. The true value $x_{10} = 42.60$ is in the interval of the largest membership states of the predicted value x^*_{10}. Using Jan. to Sept. data to predict Oct. data can also have a very good effect. Experiments show that the fuzzy Markov chain has catholicity in fuzzifying network metric prediction.

6 Conclusions and prospects

Enron Corporation announced that it was in financial crisis in the autumn of 2001, and then the American Ministry of Economic Affairs inquired this. The Securities and Exchange Commission (SEC) investigation began to step in incident investigation. Through network analysis, it can be learned that although centralization and density have a considerable rise and fall in May and June, they are growing slowly until Aug. and Sept., which means that the internal network links have more frequency and more closeness. Centralization and density in Nov. decreased as the experiment predicted, which means that the contacts within the network began to be weakened. In fact, Enron's stock price dropped to four dollars in Nov. 2001 and finally fell to only a few cents as expected. In Dec. 2001, Enron finally

quitted the stock market and ended in bankruptcy. This indicates the importance of global measures in social network analysis. Through this experiment, we know that the fuzzy time-series network is more advantageous than the fuzzy Markov chain in global measures. In comparison to the traditional time-series network, it has a smaller standard deviation under the global measures, thus a smaller fluctuation, which makes forecasting more accurate. Besides, fuzzy network forecasting has shown its accuracy in prediction using two-step and node-screened methods, indicating that it has a universality in time slice selection and data selection. In addition, it can be applied to different datasets and obtain a favorable result. It is found that there is facility and advantage in the prediction of a fuzzy time-series network measure with the fuzzy Markov chain model. However, some problems are raised in the experiment. For example, the border of the membership function or the bad data that should be repaired will affect the final prediction accuracy and the method of expanding the application range of the fuzzy network. So, we will apply the fuzzy network to other prediction models in a better manner and will choose appropriate data. Optimization of the observed data is exactly our future work.

References

Araujo, E., 2008. Social relationship explained by fuzzy logic. Proc. IEEE Int. Conf. on Fuzzy Systems, p.2129-2134. [doi:10.1109/FUZZY.2008.4630664]

Bastani, S., Jafarabad, A.K., Zarandi, M.H.F., 2013. Fuzzy models for link prediction in social networks. *Int. J. Intell. Syst.*, **28**(8):768-786. [doi:10.1002/int.21601]

Brunelli, M., Fedrizzi, M., 2009. A fuzzy approach to social network analysis. Proc. Int. Conf. on Advances in Social Network Analysis and Mining, p.225-230. [doi:10.1109/ASONAM.2009.72]

Brunelli, M., Fedrizzi, M., Fedrizzi, M., 2014. Fuzzy m-ary adjacency relations in social network analysis: optimization and consensus evaluation. *Inform. Fusion*, **17**:36-45. [doi:10.1016/j.inffus.2011.11.001]

de Sa, H.R., Prudencio, R.B.C., 2011. Supervised link prediction in weighted networks. Proc. Int. Conf. on Neural Networks, p.2281-2288. [doi:10.1109/IJCNN.2011.6033513]

Ebel, H., Davidsen, J., Bornholdt, S., 2002. Dynamics of social networks. *Complexity*, **8**(2):24-27. [doi:10.1002/cplx.10066]

Freeman, L.C., 1978. Centrality in social networks conceptual clarification. *Soc. Netw.*, **1**(3):215-239. [doi:10.1016/0378-8733(78)90021-7]

Freeman, L.C., 2004. The Development of Social Network Analysis: a Study in the Sociology of Science. Empirical Press, Vancouver.

Hasan, M.A., Chaoji, V., Salem, S., *et al.*, 2006. Link prediction using supervised learning. Proc. SDM Workshop on Link Analysis, Counter-Terrorism and Security, p.1-10.

He, Y.L., Liu, J.N.K., Hu, Y.X., *et al.*, 2015. OWA operator based link prediction ensemble for social network. *Expert Syst. Appl.*, **42**(1):21-50. [doi:10.1016/j.eswa.2014.07.018]

Huang, Z., Lin, D.K.J., 2009. The time-series link prediction problem with applications in communication surveillance. *INFORMS J. Comput.*, **21**(2):286-303. [doi:10.1287/ijoc.1080.0292]

Jaccard, P., 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull. Soc. Vaud. Sci. Nat.*, **37**:547-579 (in French).

Jin, E.M., Girvan, M., Newman, M.E.J., 2001. The structure of growing social networks. Available from http://ideas.repec.org/p/wop/safiwp/01-06-032.html [Accessed on June 30, 2015].

Khorasani, E.S., Rahimi, S., Patel, P., *et al.*, 2011. CWJess: an expert system shell for computing with words. Proc. IEEE Int. Conf. on Information Reuse and Integration, p.396-399. [doi:10.1109/IRI.2011.6009580]

Nair, P.S., Sarasamma, S.T., 2007. Data mining through fuzzy social network analysis. Proc. 26th Annual Meeting of the North American Fuzzy Information Processing Society, p.251-255. [doi:10.1109/NAFIPS.2007.383846]

Ryoke, M., Nakamori, Y., Suzuki, K., 1995. Adaptive fuzzy clustering and fuzzy prediction models. Proc. Int. Joint Conf. of 4th IEEE Int. Conf. on Fuzzy Systems and 2nd Int. Fuzzy Engineering Symp., p.2215-2220. [doi:10.1109/FUZZY.1995.409987]

Yan, B., Gregory, S., 2011. Finding missing edges and communities in incomplete networks. *J. Phys. A*, **44**:495102.1-495102.15.

Zadeh, L.A., 1965. Fuzzy sets. *Inform. Contr.*, **8**(3):338-353.

Zhang, J.Y., Borland, R., Coghill, K., 2011. Evaluating the effect of health warnings in influencing Australian smokers' psychosocial and quitting behaviours using fuzzy causal network. *Expert Syst. Appl.*, **38**(6):6430-6438. [doi:10.1016/j.eswa.2010.11.042]

Zhu, J., Xie, Q., Chin, E.J., 2012. A hybrid time-series link prediction framework for large social network. Proc. 23rd Int. Conf. on Database and Expert Systems Applications, p.345-359. [doi:10.1007/978-3-642-32597-7_30]