Analyzing Dynamic Change in Social Network Based on Distribution-Free Multivariate Process Control Method

Yan Liu[1, *], Lian Liu[1], Yu Yan[2], Hao Feng[1] and Shichang Ding[3]

Abstract: Social organizations can be represented by social network because it can mathematically quantify and represent complex interrelated organizational behavior. Exploring the change in dynamic social network is essential for the situation awareness of the corresponding social organization. Social network usually evolves gradually and slightly, which is hard to be noticed. The statistical process control techniques in industry field have been used to distinguish the statistically significant change of social network. But the original method is narrowed due to some limitation on measures. This paper presents a generic framework to address the change detection problem in dynamic social network and introduces a distribution-free multivariate control charts to supervise the changing of social network. Three groups of network parameters are integrated together in order to achieve a comprehensive view of the dynamic tendency. The proposed approaches handle the non-Gaussian data based on categorizing and ranking. Experiments indicate that nonparametric multivariate procedure is promising to be applied to social network analysis.

Keywords: Dynamic social network analysis, statistical process control, email network, categorization, rank.

1 Introduction

Social organizations can be represented with different networks, such as communication network, resources sharing network, and so on. Social Network Analysis (SNA) is a usual approach for studying and analyzing groups of actors and their ties.

Organizations are not static. Their structure, composition, and patterns of communication may change over time. These changes may occur quickly, such as when a corporation restructures, but they often happen gradually, as individual roles expand or contract. These tendentious changes often reflect the gradual evolution of an organization. However, the trend change is easily confused with the normal fluctuation. In the normal operation of an organization, there will be a certain degree of fluctuation, which reflects the normal daily changes.

However, most techniques in social network analysis focus on static relationships between actors in social organizations. As the organizations are developing, their corresponding social networks are fluctuating overtime. Researches on dynamics of social network are

[1] State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, 450001, China.

[2] The First Affiliated Hospital of Chongqing Medical University, Chongqing, 400016, China.

[3] Goettingen University, Goettingen, 37077, Germany.

[*] Corresponding Author: Yan Liu. Email: ms_liuyan@aliyun.com.

CMC. doi:10.32604/cmc.2019.05619

significant for understanding the behaviors of organization. Drastic changes are obvious, but the network may experience gradually changes which are too subtle to notice at most of the time. At present, network change detection technology has been widely used in social network analysis, industrial control network [Wan, Yao, Jing et al. (2018)] and other fields.

McCulloh et al. [McCulloh and Carley (2008)] originated a new area of researching named social network change detection (SNCD), which combined the theories of Social Network Analysis with the techniques of Statistical Process Control (SPC) to rapidly detect statistically significant changes in dynamic social network. However, this approach has some limitation. On one hand, the effectiveness of the approach is sensitive to missing information. On the other hand, this approach assumes that network measures are normally distributed. As a matter of fact, complete information is hard to achieve because of the limitation of privacy, encryption, and inaccurate operations and so on. The strict assumption limits the selection of measures and the using of approach.

Our research attaches importance to the issues above, we put forward three groups of parameters for measuring communication network, and adopt another SPC procedure proposed by Qiu [Qiu (2018)] to explore the trend of dynamic network. The approach is tested on Enron Email dataset and proves that it can be applied in analyzing dynamic tendency in social networks well with careful adjustments.

2 Related work

The methods based on graph is the general technique for analyzing the dynamitic of network. Related work is classified as bellow.

2.1 Distance measure applied to dynamic network anomaly detection

One common dynamic network anomaly detection method is to use network distance metrics to find time slices that are extremely different from the network in the previous time step. The metrics measured in graphs are typically structural features, Once the summary metrics are found for each graph, the difference or similarity, which are inversely related, can be calculated. The variation in the algorithms lies in the metrics chosen to extract and compare, and the methods they use to determine the anomalous values and corresponding graphs. Hamming distance is often used in binary networks to measure the distance between two networks. Euclidean distance is similarly used for weighted networks. Jaccard coefficient and graph edit distance are used to measure the number of same actors and edges between two graphs. Another examples of network distance metrics include [Akoglu, Tong and Koutra (2015); Koutra, Shah, Vogelstein et al. (2016); Berlingerio, Koutra and Faloutsos (2013)].

While these methods may be effective at quantifying difference among static networks, they lack an underlying statistical distribution. This is a constraint on identifying a statistically significant change, as opposed to normal and spurious fluctuations in the network.

2.2 Probabilistic models applied to dynamic network anomaly detection

With a foundation in probability theory, distributions, and scan statistics, these methods typically construct a model of what is considered 'normal' or expected, and flag deviations from this model as anomalous.

Some of the most popular models for hypothesis testing are the exponential random graph models (ERGM) family. An example is dynamic ERGM variant for hypothesis testing in a network stream [Desmarais and Cranmer (2012); Snijders, van de Bunt and Steglich (2010)]. A dynamic ERGM model can be used with likelihood ratio testing to perform anomaly detection. But the model is limited because it is not suitable for a network of nodes larger than ten thousand.

Scan statistics are often called ‘moving window analysis’ [Ranshous, Shen, Koutra et al. (2015)], where the local maximum or minimum of a measured statistic is found in specific regions of the data. In a graph, a scan statistic can be considered as the maximum of a graph invariant feature. Wan et al. [Wan, Milios, Kalyaniwalla et al. (2009)] proposed a link-based event detection method that clusters vertices with similar communication patterns together and then, considers deviations from each vertex's individual profile, as well as its cluster profile.

The methods of scan statistics have advantages at detailed network statistics. As a result, it could be used to locate the change area of the large network. However, the high cost of computation for each window should be considered when these algorithms are applied to really application.

2.3 Statistical process control applied to dynamic network anomaly detection

SNCD was discussed by McCulloh et al. using social network analysis techniques and statistical process control to identify small changes via monitoring network measures independently [McCulloh and Carley (2008)]. By taking measures of a network over time, a control chart can be used to signal when significant changes occur in the network. CUSUM chart was recommended for longitudinal social network analysis. Azarnoush et al. [Azarnoush, Paynabar, Bekki et al. (2016)] also proposed a statistical method to monitor Enron Email formation mechanism via network attributes. Ebrahim Mazrae Farahani et al. [Ebrahim, Reza, Rassoul et al. (2017)] used multivariate exponentially weighted moving average (MEWMA) and multivariate cumulative sum (MCUSUM) control charts to monitor the network formation process.

However, these approaches assume that network measures are normally distributed. Gaussian distribution plays an important role in SPC, the observed measurements are required to follow the normal distribution when applying traditional CUSUM control chart. However, the condition is very strict in applications, especially when we consider the measures of dynamic network. Usually, people transform the original data into Gaussian data in order to use the control chart, but it is extremely difficult for the transformation of multivariate data, because of the requirement that all its lower-dimensional marginal distribution must be normally distributed.

3 Dynamic network change analysis based on distribution-free multivariate process control

Equations and mathematical expressions must be inserted into the main text. Two different types of styles can be used for equations and mathematical expressions. They are: in-line style, and display style.

3.1 Problem definitions

In this subsection, social network modeling and notation are discussed. Firstly, the notation and definitions used to formulate the model and method are presented. A social network can be provided in the form of network relationships matrices. The notation to characterize a social network is presented by the following equations:

$$g(t) = \big(V(t), E(t)\big); t = 1, \dots, n \tag{1}$$

$$V(t) = \{v_1, v_2, \dots v_i, \dots v_m\} \tag{2}$$

$$E(t) = \{e_1, e_2, \dots e_j, \dots e_l\} \tag{3}$$

where $V(t)$ and $E(t)$ represent nodes and edges in time period t, respectively. A relationship may be defined as any possible communications such as Email, phone calls, SMS, Telegram messages, etc.

A social dynamic network G is presented by the following equation:

$$G = \{G_t, t = 1, \dots, n\} \tag{4}$$

where G_t represents a slice network, which is a snapshot of the social network during the given period. Sometimes $G_t = g(t)$ if the network is split without any operation.

Considering the actual situation, it is impossible for social communication relationship to completely coincide with social relationship. Many factors can lead to interference, which makes it impossible to accurately reflect the corresponding actual social relationship in the dynamic network through simple split operation. The time slice network only inspects the social communication in this period, which has great uncertainty. By referring to the historical data, increasing the period of investigation and using the historical data to correct the current data, the revised connection relationship can be more stable and accurate. In this paper, G_t is superimposed by the first t time slices network.

$$G_t = g_1 \cup g_2 \cup \dots \cup g_t \tag{5}$$

The superposition of time slice networks can be specific by the following equation:

$$G_t = \theta_1 g_1 \underset{\oplus}{\cup} \theta_2 g_2 \underset{\oplus}{\cup} \dots \underset{\oplus}{\cup} \theta_t g_t \tag{6}$$

where $\underset{\oplus}{\cup}$ represents network superposition operation. The proportion of each time slice network is satisfied $\theta_i \geq 0$ and $\sum \theta_i = 1$. Normally, the longer the distance, the smaller the impact of time slice network on the current situation, so the proportion of θ_i should increase with i. The exponential smoothing method can be used to select the proportion of each time slice network, and then the slice network G_t can be defined recursively:

$$G_t = \theta G_{t-1} \underset{\oplus}{\cup} (1-\theta) g_t, 1 > \theta \geq 0 \tag{7}$$

If θ =0, the impact of history is not considered, the current network snapshot is used directly as a slice network for this period, then $G_t = g_t$. If θ >0, as time series t increases, the influence of time slice network g_t on G_t increases.

In order to avoid the excessive number of edges, the minimum edge weight of the network is stipulated. When the edge weight is less than ε, its weight is regarded as 0 and is deleted. Deleting edges with weights less than ε may result in some nodes having no connections in the network and becoming isolated nodes. The isolated node is not valid for the linked relationship, so it is deleted.

The task of dynamic network change detection is to analyze the trend of network evolution from these overlapped networks and find abnormal changes.

3.2 Dynamic network change analysis framework

In this section, the dynamic network analysis framework is introduced systematically. The relationship among the organization is viewed as dynamic in our work. If we use graph to express the communication network, some edges appear during one period and some disappear during another. The analysis framework is shown in Fig. 1.

First, the time span of the dynamic network is split into disjoint time intervals (e.g., one week). Within each time interval t, a static graph $g(t)$ is built to summarize the dynamic network. In other words, all the edges that ever appeared during this time interval are kept in the static graph. For each $g(t)$, three groups of network parameters are measured. Stable and continuous time interval is selected and trained [Bush, Chongfuangprinya, Chen et al. (2010).]. Then the parameters of the subsequent time interval are compared with the trained result. If the parameters show one-way growth trend, then some clues may be found out for early warning detection.

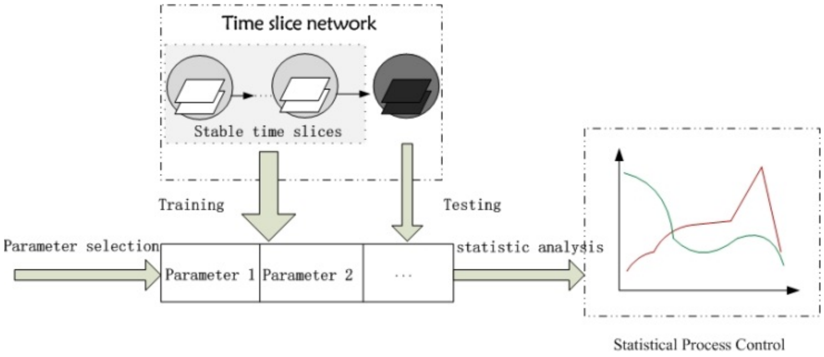


Figure 1: Dynamic network change analysis framework

As for social network, which is not a strict industrial process, parameters have two characteristics. First, many parameters' distribution is unknown in social network. Second, network properties are reflected by various parameters, among which the correlation is unknown, and as a result, independently testing on some parameters may cause one-sided mistakes.

According to the two characteristics above, our improvement lies in 1) analyzing various parameters synthetically and 2) overcoming the restrictions of the normal distribution on parameters.

For 1), we select three groups of network parameters in order to study the network change behavior from different perspective as introduced in section 5. The parameters could be analyzed by multivariate statistical analysis.

For 2), most existing multivariate SPC procedures assume that the in-control distribution of the multivariate process measurement is known and it is a Gaussian distribution. In applications, however, the measurement distribution is usually unknown and it needs to be estimated from data. Furthermore, multivariate measurements often do not follow a Gaussian distribution (e.g., cases when some measurement components are discrete). Existing statistical tools for describing multivariate non-Gaussian data or, transforming the multivariate non-Gaussian data to multivariate Gaussian data are limited, making appropriate multivariate SPC difficult especially in high dimensional data. Qiu2 suggested a methodology for estimating the in-control multivariate measurement distribution when a set of in-control data is available, which is based on log-linear modeling and which takes into account the association structure among the measurement components. As described by his method, data were categorized, and the in-control distribution is estimated by log-linear model. We supervise dynamic network measures according to this procedure as follows.

3.3 Three groups of network parameters

In order to fully understand the network changes behavior, it is needed to study the characteristics of various dedicated network changes. The network parameters can quantify the characteristics of the network from different perspective, which are divided into three categories in this paper.

3.3.1 Scale parameters

Being faced with the target network which is corresponding to an organization, our interesting focus on what change will be triggered when some events happen. The occurrence of event will affect the number of records collected and results in the change of the network scale. In this section, we select three parameters to measure the network scale.

Edge Count. Edge count denotes how many times the communication is happened during the given period. We use $EC(G)$ to marker edge count of the network, which is defined as

$$EC(G) = |E| \tag{8}$$

Node Count. Node count denotes how many actors who are take part in the communication. We use $NC(G)$ to marker node count of the network, which is defined as

$$NC(G) = |V| \tag{9}$$

Weight Sum. The relationship between the two actors is viewed as different. More frequent communication means higher score of weight. Weight sum is the sum of all the relationship, namely

$$WS(G) = \sum_{e \in E} w(e) \tag{10}$$

3.3.2 Connectivity parameters

Connectivity refers to the connected features between the actors in the network. Stronger connectivity implies easier communication among the social entities and more fluency of information transfer. Three parameters are used to quantify the connectivity of a network.

Average Distance. Average Distance reflects the average number of sending and receiving messages between any social members. Given a directed network, let the average distance is the average distance between any two actors, namely

$$AD(G) = \frac{\sum_{v_i, v_j \in V, i \neq j} d_{ij}}{|V|(|V|-1)} \tag{11}$$

where v_i and v_j denote two actors in the network and d_{ij} denotes the number of edges of the shortest path from v_i to v_j. If there is no shortest path, then d_{ij} equals to the number of nodes in the network.

Reciprocal Link Counts. With symmetric dyadic data, two actors are connected to each other. Reciprocal link counts of a network reflect how many proportions the number of the bidirectional relationship accounts for, namely

$$RL(G) = \frac{|\{e_{ij} \in E | e_{ji} \in E\}|}{|E|} \tag{12}$$

where e_{ij} denotes an edge from v_i to v_j.

Transitivity. We are interested in the proportion of triads that are "transitive" (that is, display a type of balance where, if A directs a tie to B, and B directs a tie to C, then A also directs a tie to C). Transitivity of the communication network is defined as how many proportions the number of transitive triads accounts for, namely

$$T(G) = \frac{|\{(v_i, v_j, v_k) \in V^3\} | e_{ij}, e_{jk}, e_{ik} \in E|}{|\{(v_i, v_j, v_k) \in V^3\} | e_{ij}, e_{jk} \in E|} \tag{13}$$

3.3.3 Compactness parameters

If we take into account a network that describes the social relationship, compactness reflects to what extent the network has a closely structure as a whole. More compact the network means more frequent communication and as a result, more closely relationship. Density and Centralization are two kinds of parameters which are usually used to weight the compactness of a network.

Density. Density of the communication network implies the cohesion level. It is defined as the actual number of network edges versus the maximum possible edges for a network N, namely

$$D(G) = \frac{|E|}{|V|(|V|-1)} \tag{14}$$

Network Closeness Centralization. Closeness centrality approaches emphasize the distance of an actor to all others in the network by focusing on the distance from each actor to all others. Loosely, closeness is the inverse of the average distance in the network between the node and all other nodes.

Let $dist = \sum_{j \in V} d_G(v_i, j)$, if every node is reachable from v_i, then closeness centrality of node v_i is defined as

$$CC_i = (|V| - 1)/dist = \frac{|V|-1}{\sum_{j\in V} d_G(v_i,j)} \tag{15}$$

If some node is not reachable from v then the closeness centrality of v is $|V|$.

Network closeness centralization based on the closeness centrality of each node in a square network. Let $CC_{max} = \max\{CC_i|1 \leq i \leq |V|\}$, the network closeness centralization is defined as

$$CC(G) = \frac{\sum_{v_i\in V}(CC_{max} - CC_i)}{\frac{(|V|-2)(|V|-1)}{(2|V|-3)}} \tag{16}$$

Network betweenness centralization. With binary data, betweenness centrality views an actor as being in a favored position to the extent that the actor falls on the geodesic paths between other pairs of actors in the network. That is, the more people depend on the actor to make connections with other people, the more power it has. The betweenness centrality of node v in a network is defined as: across all node pairs that have a shortest path containing v, the percentage that passes through v. Let g_{jk} represents the number of shortest path from v_j to v_k, and $g_{jk}(v_i)$ represents the number of shortest path from v_j to v_k and v_i exists in the nodes set in the path, then the betweenness centrality of node v_i is defined as

$$BC_i = \frac{\sum_{j<k}\frac{g_{jk}(v_i)}{g_{jk}}}{(|V|-1)(|V|-2)} \tag{17}$$

Network centralization based on the betweenness score for each node in a square network. Let BC_{max} represents the greatest degree of mediation of all nodes, the network betweenness centralization is defined as

$$BC(G) = \frac{\sum_{v_i\in V}(BC_{max} - BC_i)}{n-1} \tag{18}$$

3.4 Nonparametric multivariate procedures

In this section, we adopt nonparametric multivariate process control procedures to supervise the measures of dynamic network as developments of SNCD. The goal of the improvements is to explore the trend of social network more comprehensively and more precisely.

It is addressed that we focused on the trend of the dynamic network in this research, not the particular change, so the control limit (a constant value denoted h in a common control chart) will be ignored. Thus, there are some modifications in our methods differentiate from original SPC procedures. The purpose is to be convenient to understand the trend of statistics.

3.4.1 Categorization method

Phrase I SPC.

First, the time span of the whole observed dynamic network is split into disjoint time intervals. Each time interval is called a slice. Take a steady time-period of the dynamic network as the in-control period when the network seems stable. Let n_0 represents the length of in-control network snapshots. Three groups of network parameters are selected for investigation. Let r parameters are measured, and the in-control data can be denoted by $\{X(i) = (X_1(i), X_2(i), \dots, X_r(i))', i = 1,2,\dots, n_o\}$.

Then calculate the median m_j of$X_j(i)$ from the in-control data. Let $Y_j(i) = I(X_j(i) > m_j), j = 1,2,\dots, r$, where $I(x)$ is an indicator function which equals 1 if x is "true", and 0 otherwise. Thus, the r components are converted to a r-dimensional vector: $Y(i) = (Y_1(i), Y_2(i), \dots, Y_r(i))'$. It is rational to use the categorized data for detection, because the changes of network measures make the median vector altered that will also change the distribution of Y(i) . Therefore, we are able to detect shifts in such a location parameter vector (e.g., the median vector $(\mathrm{m}_1, \mathrm{m}_2, \dots, \mathrm{m}_\mathrm{r})'$).

In statistics, a contingency table (also referred to as cross tabulation or cross tab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. It is often used to record and analyze the relation between two or more categorical variables. A r-dimensional contingency table of the in-control data can be easily formed from these binary variables $Y(i)$. Let $\mathrm{O}_{\mathrm{j}_1,\mathrm{j}_2,\dots,\mathrm{j}_\mathrm{r}}$ be the observed cell count of the $(\mathrm{j}_1, \mathrm{j}_2, \dots, \mathrm{j}_\mathrm{r})$-th cell in the table. Estimate expected count $\mathrm{E}_{\mathrm{j}_1,\mathrm{j}_2,\dots,\mathrm{j}_\mathrm{r}}$ of each cell in the contingency table using log-linear model, so we have an estimator $\{\frac{\mathrm{E}_{\mathrm{j}_1,\mathrm{j}_2,\dots,\mathrm{j}_\mathrm{r}}}{\mathrm{n}_0}, \mathrm{j}_1, \mathrm{j}_2, \dots, \mathrm{j}_\mathrm{r} = 0,1\}$ for the joint distribution of $\mathbf{Y}$(i) , denoted as $\{\mathrm{d}_{\mathrm{j}_1,\mathrm{j}_2,\dots,\mathrm{j}_\mathrm{r}}, \mathrm{j}_1, \mathrm{j}_2, \dots, \mathrm{j}_\mathrm{r} = 0,1\}$, in which $\mathrm{d}_{\mathrm{j}_1,\mathrm{j}_2,\dots,\mathrm{j}_\mathrm{r}} = \frac{\mathrm{E}_{\mathrm{j}_1,\mathrm{j}_2,\dots,\mathrm{j}_\mathrm{r}}}{\mathrm{n}_0}, \mathrm{j}_1, \mathrm{j}_2, \dots, \mathrm{j}_\mathrm{r} = 0,1$. So far, we calculated the distribution of in-control period.

Phrase II SPC.

In this section, distribution-free, multivariate CUSUM is used for detecting shifts in the joint distribution of $\mathbf{Y}$(i). Network measures of successive slices are calculated as the same way in Phrase I. The binary vector $\mathbf{Y}$(i) is re-calculated and transformed on the new data set involved the new network of the successive slice one by one. And the contingency table is recount by adding the new vector on it at each step.

Assume that the in-control joint distribution of $\mathbf{Y}$(i) is $\{\mathrm{d}^0_{\mathrm{j}_1,\mathrm{j}_2,\dots,\mathrm{j}_\mathrm{r}}, \mathrm{j}_1, \mathrm{j}_2, \dots, \mathrm{j}_\mathrm{r} = 0,1\}$, which can be estimated by the log-linear modeling procedure discussed in Phrase I. In the statistical literature, the Pearson's χ^2 test is well-known for testing whether or not the distribution of a random vector equals a given distribution. Let $g_{j_1,\dots,j_r}(i) = I(Y_1(i) = j_1, \dots, Y_r(i) = j_r)$, where $j_1, \dots, j_r = 0$ or 1 . Then $g^n_{j_1,j_2,\dots,j_r} = \sum_i^n g_{j_1,\dots,j_r}(i)$ is the observed count of the $(j_1, \dots, j_r) - th$ cell as of time point n. The conventional Pearson's χ^2 statistic is defined by

$$\sum_{j_1,j_2,\dots,j_r=0,1} \frac{(\frac{g^n_{j_1,j_2,\dots,j_r}}{n} - d_{j_1,j_2,\dots,j_r})^2}{d_{j_1,j_2,\dots,j_r}} \qquad (19)$$

In order to measure the deviation between the observed value and the expected value and accumulate the statistics. An existed CUSUM procedure [Liu, Liu and Luo (2011)] is used for detecting possible shifts in a location parameter vector, of which following steps are:

For each observation, define

$$Z_n = \ [\left(\mathrm{S}_{n-1}^{obs} - \mathrm{S}_{n-1}^{exp}\right) + \left(\frac{\mathrm{g}(n)}{n} - d\right)]'\left[diag\left(\mathrm{S}_{n-1}^{exp} + d\right)\right]^{-1} \ [(\left(\mathrm{S}_{n-1}^{obs} - \mathrm{S}_{n-1}^{exp}\right) + \left(\frac{\mathrm{g}(n)}{n} - d\right)] \qquad (20)$$

where $d$is a vector of all $(d_{j_1,j_2,\dots,j_r})$ values, $g(n)$ is a vector of all $(g_{j_1,j_2,\dots,j_r}^{n})$ values, and $\frac{g(n)}{n}$ is the observed frequency vector, $diag(x)$ denotes a diagonal matrix of vector x, $S_0^{obs} = S_0^{exp}$, and if $Z_n \leq k, \begin{cases} S_n^{obs} = 0 \\ S_n^{exp} = 0 \end{cases}$, if $Z_n > k, \begin{cases} S_n^{obs} = \frac{(S_{n-1}^{obs} + \frac{g(n)}{n})(Z_n - k)}{Z_n} \\ S_n^{exp} = \frac{(S_{n-1}^{exp} + d)(Z_n - k)}{Z_n} \end{cases}$, in which, constant k is the reference value used in CUSUM procedures depending on the magnitude of a target shift. And we can calculate the cumulative statistic finally:

$$C_n = \left[\left(\mathrm{S}_n^{obs} - \mathrm{S}_n^{exp}\right)\right]'[diag\left(\mathrm{S}_n^{\mathrm{exp}}\right)]^{-1}[\left(\mathrm{S}_n^{obs} - \mathrm{S}_n^{exp}\right)] \qquad (21)$$

In a typical SPC procedure, the constant k and $h$are commonly decided by the Average Run Length (ARL) which is a criterion of the control chart. And the original search of k is to simulate a detect process by adjusting the reference value to achieve an ideal ARL value. However, the value of k cannot be inferred from ARL here because of the absence of h. In another way, we consider that the fluctuation of in-control distribution is under the range of target shift, so we simulate an in-control data set to evaluate the average discrepancy from the expect value. A series of random vectors is generated from the multinomial distribution with probability parameters$\{d_{j_1,j_2,\dots,j_r}, j_1, j_2, \dots, j_r = 0,1\}$, and calculate the average conventional Pearson's χ^2 statistic between the expect frequency and random vectors as the value of k:

$$k = \sum_{j_1,j_2,\dots,j_r=0,1} \frac{\left(\frac{g_{j_1,j_2,\dots,j_r}^{L}}{L} - d_{j_1,j_2,\dots,j_r}\right)^2}{d_{j_1,j_2,\dots,j_r}} \qquad (22)$$

where $g_{j_1,j_2,\dots,j_r}^{L}$is each random vector, assuming there are L vectors total.

3.4.2 Ranking method

Based on the sorting multi-variable non-parametric control chart [Bush, Chongfuangprinya, Chen et al. (2010)], the data of the test set are added to the training set for sorting, and the degree of the test data deviating from the training set is examined according to the sorting value.

It is assumed that each inspection data contains r network parameters. The training set consists of n_0 samples. $\{X(i) = (X_1(i), X_2(i), \dots, X_r(i))', i = 1,2, \dots, n_0\}$. The test set is represented by a data pool. Each time a vector $X(i)$ is added to the training set, the data pool capacity is $n_0 + 1$. The vectors in the pool are sorted to a sequence. First, we compute the median vector of the data pool as the first vector of the sequence. Then the distance is calculated between the data in the pool and the data in the sequence. The distance between two vectors is measured by Mahalanobis distance [Maesschalck, Jouan-Rimbaud and Massart (2000)]:

$$d_{ij} = (\mathrm{X}(i) - \mathrm{X}(j))'\mathrm{V}^{-1}(\mathrm{X}(i) - \mathrm{X}(j)) \qquad (23)$$

where V represents the covariance matrix of training set samples. The distance D_i between each data and sequence is computed by K nearest neighbor method. That is, the distance between the nearest k data in the sequence is used as the distance between the current data and the sequence.

$$D_i = \frac{1}{k}\sum_{h=1}^{k} d_i(h) \tag{24}$$

where $d_i(h)$ represents the small h value of the distance between the vector and all the vectors in the sequence. Finally, all $n_0 + 1$ data are added to the sequence, the sort R_n of $X(n)$ in the data pool is got, which is converted to sort statistics $u_n = R_n/(n_0 + 1)$. The range of u_n is $[1/(n_0+1),1]$. After each sorting gets u_n, $X(n)$ is deleted from the data pool. Then the next vector $X(n+1)$ in the test set is added to reorder the data in the data pool until statistics of all data in the test set are obtained sequentially.

The larger the u_n, the farther the distance between $X(n)$ and the training set is, that is, the larger the N time slice network changes. The network changes relatively sharply when the ranking statistics value is relatively large. However, the maximum value of u_n can only be obtained by 1. If the ranking statistics of a time slice network have reached 1 but continue to increase, it is impossible to know the specific details from the results. This is the defect of this sort-based method. Therefore, we use both methods to make up for each other's defect and give full play to its advantages.

4 Experiments

This section evaluates the effectiveness of the proposed methods based on Enron dataset.

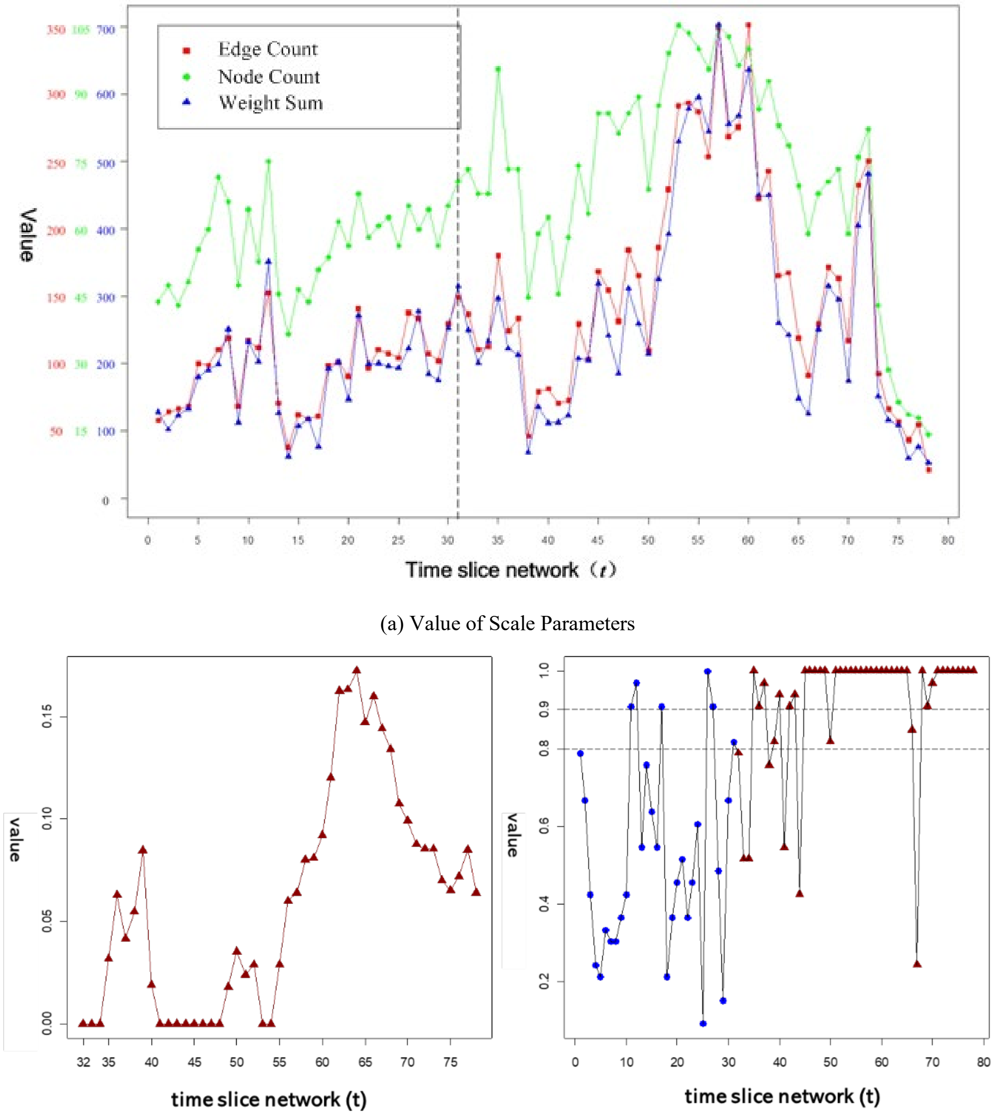
4.1 Dataset

Our methods are applied on the Enron Email dataset. The whole time-period is partitioned weekly, and a direct network in which node represents Email address and arc represents Email communication is exacted from each week. The scales of the weekly networks are very small before 2000, so we delete them. According to the previous result [Crosier (1988)], we select the weeks from the 40-th week in 2000 to 17-th in 2001 as the stable period. The 31 weeks is regarded as the in-control data, named 2000.w40~2001.w17. Then the next 47 weeks are selected as testing data for dynamic tendency analysis.

4.2 Plot of values vs. cumulative statistic of values

As described in Section 3, the behavior of network is measured by three groups of parameters, which are scale parameters, connectivity parameters and compactness parameters. Then, we select three parameters from each group to test their distribution. Fig. 2~Fig. 4 shows that our method is superior to the simple plotting value in terms of the sensitive on highlighting network dynamic tendency.

The (a) of the three figures plot the values of the given 9 parameters from week 1 to week 78. Values are fluctuating, but the fluctuation could not reveal some dynamic character of the network. If we make a careful study on the right part (b) and (c) of the four figures, we can draw a conclusion that there are some changes happened during the period from week 48 (2001.08.26~2001.09.01) to week 60 (2001.11.18~2001.11.24).



(a) Value of Scale Parameters

(b) Cumulative Statistic

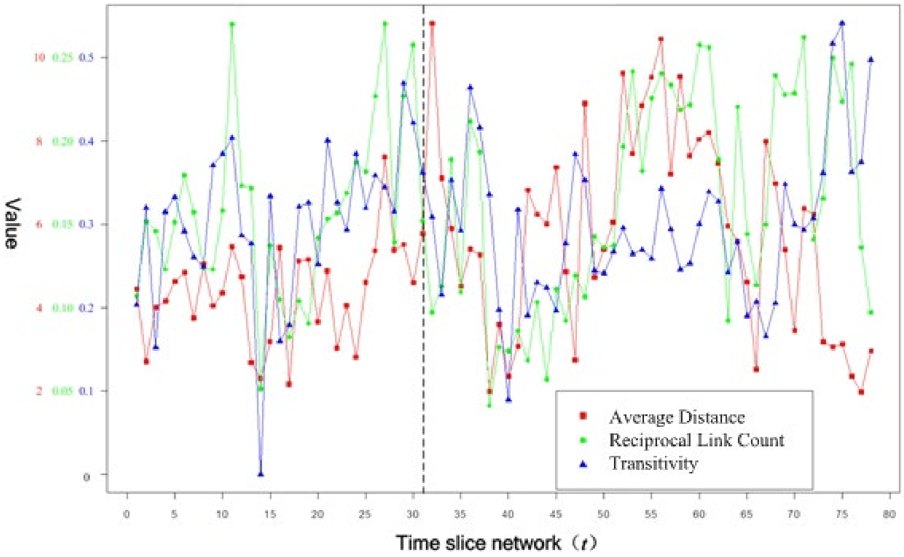
(c) Ranking Statistic

Figure 2: Plot of values vs. cumulative statistic of values on scale parameters

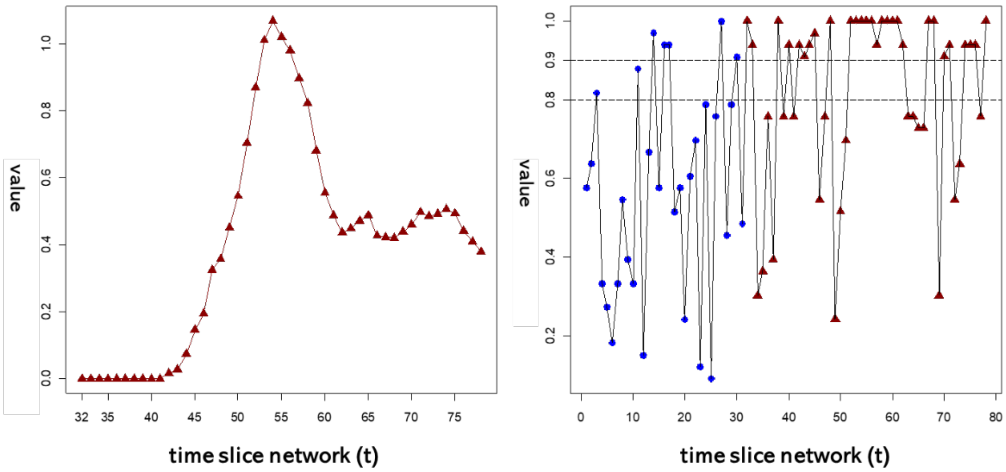
According to Fig. 2(b), we can see the change trend of network scale characteristics. From 35 weeks (2001.5.27-2001.6.2) to 40 weeks (2001.7.1-2001.7.7) the network scale has a small change, but from 41 weeks (2001.7.8-2001.7.14) to 48 weeks (2001.8.26-2001.9.1) it restores stability, and from 49 weeks (2001.9.2-2001.9.8) it occurs again. Significant changes, and gradually intensified, reached the maximum in 64 weeks (2001.12.16-2001.12.22), and then gradually slows down. According to Fig. 2(c), it can be found that the scale characteristics of time slice network after training set has obvious changes, especially after 45 weeks (2001.8.5-2001.8.11), almost all of them get the maximum of ranking statistics.

Comparing the results of the two methods, the cumulative statistic changes of the end-time slice network (g70-g78) are slowed down, while the ranking statistics get the maximum, indicating that the network changes dramatically. This is because the extreme value of ranking statistics is limited, and the data that is seriously inconsistent with the distribution of training set data is added to the training set to rank at the end of the queue at best, so that the maximum statistics can only be taken as 1. Therefore, the understanding of the conflict between the two results is: in fact, the network changes are still dramatic at this time, but the more drastic changes have been mitigated compared with the previous ones, so the cumulative statistical scales show a gentle decline.

According to Fig. 3(b), the characteristics of network connectivity begins from 42 weeks (2001.7.15-2001.7.21), and the changes gradually accumulat and rapidly increas, reaching the maximum value in 53 weeks (2001.9.30-2001.10.6). After that, the accumulation of changes begins to decline gradually, but do not drop to 0. The training set of comparison still remains different until 62 weeks (2001.12.2-2001.10.6). 1.12.8) maintaining relatively stable differences. According to Fig. 3(c), most of the time slice network connectivity characteristics after training set have obvious changes, especially from 52 weeks (2001.9.23-2001.9.29) to 62 weeks (2001.12.2-2001.12.8), and the ranking statistics reach the maximum.



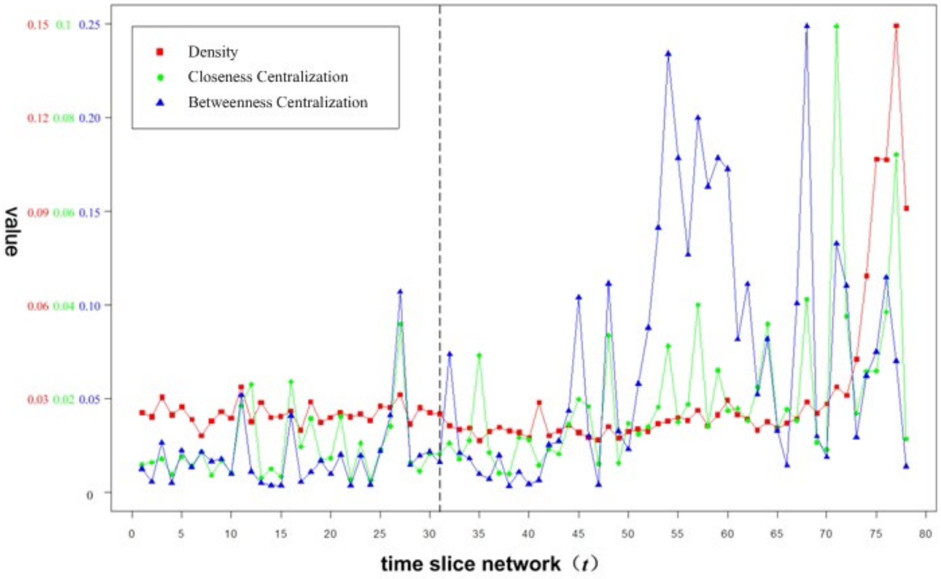
(a) Value of Connectivity Parameters



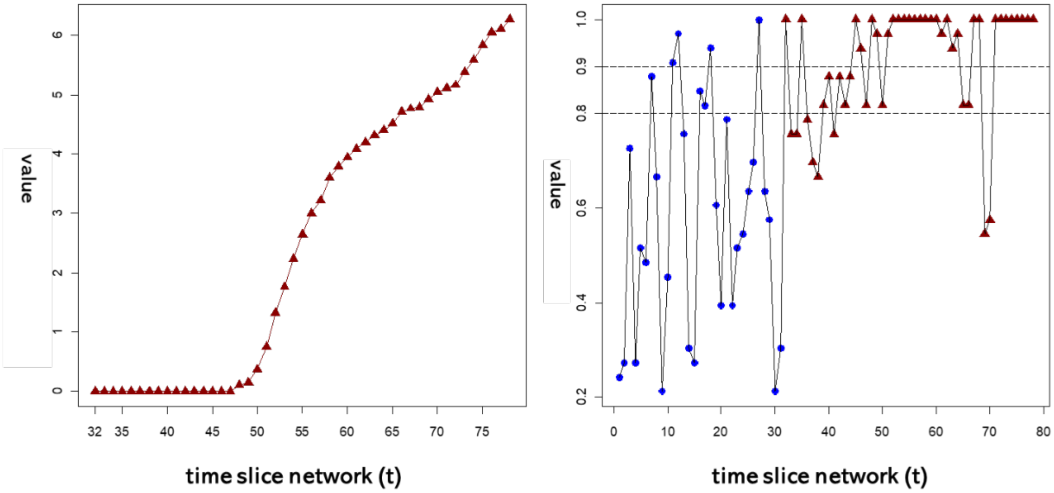
(b) Cumulative Statistic

(c) Ranking Statistic

Figure 3: Plot of values vs. cumulative statistic of values on connectivity parameters



(a) Value of Compactness Parameters



(b) Cumulative Statistic

(c) Ranking Statistic

Figure 4: Plot of values vs. cumulative statistic of values on compactness parameters

Fig. 4(b) shows that the compactness characteristics of time slice network accumulate with the change of time after stationary period. The larger the value of each point corresponding to the cumulative statistics of the time slice network data relative to the training set, the greater the change, and the continuous increase of the statistics indicates that the change is accumulating and increasing. From the graph, we can see that from 48 weeks (2001.8.26-2001.9.1), the compactness of the network begins to change, and then the change continues to accumulate and increase.

Fig. 4(c) shows how the compactness of the whole network varies over time (each difference is measured by a Mahalanobis distance). The values of each point in the graph indicate the ranking statistics of the network parameters relative to the training data. The larger the values, the greater the deviation from the training set. 0.8 and 0.9 are taken as reference lines. The blue dots represent the training set data, and the red triangles represent the subsequent network data. From the graph, we can see that the time slice network ranking statistics after the training set are significantly larger than the training set, especially after the fiftieth week (2001.9.9-2001.9.15), the sixtieth week (2001.11.18-2001.11.24), and the seventy-two weeks (2002.2.3-2002.2.9). The statistics get the maximum value, which indicates the change of compactness parameters of the network in these periods is very intense.

In order to validate our results, our detection result about network changes behavior is compared with the reality of organization behavior. Enron was rated the most innovative large company in America in Fortune magazine’s survey of Most Admired Companies. Yet within a year, Enron’s image was in tatters and its stock price had plummeted nearly to zero. [Healy and Palepu (2003)] lists some of the critical events for Enron between August and December 2001. The contrast above validates the effective of our method.

5 Conclusion

It is an innovation of using SPC for social network analysis. But there are some inevitable shortcomings with this initiated approach. For some of these restricts, we introduce multivariate distribution-free procedures from industry field to evaluate the trend of social network. The developments lie in two main aspects, one is the integration of more parameters for measurement, and the other is overcome the limit of the Gaussian distribution assumption on the network parameters which can be hardly satisfied in many circumstances. And the results of our methods are in accordance with the real case of the evolvement in organization in Enron Email dataset.

The distribution-free multivariate CUSUM procedure based on categorizing provides us very intuitive result, and after the original measures are transformed to binary data, the computational complexity is decreased greatly. And being good at discovering subtle shifts especially the gradual changes is the intrinsic advantage of CUSUM method. However, in-control distribution is estimated by the log-linear model in this approach, which would become challenging when the number of in-control observations is relatively smaller or the number of measurement components is larger, that will lead the contingency table to become sparse.

There are a lot of issues need to be addressed in our future work. The choice of observed measures and constants in control charts need more rigorous proof. And for the result, details about what the control charts indicate need a deep investigation.

Acknowledgement: This work was supported by the National Key R&D Program of China (Nos. 2016YFB0801303, 2016QY01W0105), the National Natural Science Foundation of China (Nos. 61309007, U1636219, 61602508, 61772549, U1736214, 61572052) and Plan for Scientific Innovation Talent of Henan Province (No. 2018JR0018).

References

Akoglu, L.; Tong, H.; Koutra, D. (2015): Graph based anomaly detection and description: A survey. *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 626-688.

Azarnoush, B.; Paynabar, K.; Bekki, J.; Runger, G. (2016): Monitoring temporal homogeneity in attributed network streams. *Journal of Quality Technology*, vol. 48, no. 1, pp. 28-43.

Berlingerio, M.; D. Koutra, T.; Faloutsos, C. (2013): Network similarity via multiple social theories. *Proceedings of the 5th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 1439-1440.

Bush, H.; Chongfuangprinya, P.; Chen, V.; Sukchotrat, T.; Kim, S. B. (2010): Nonparametric multivariate control charts based on a linkage ranking algorithm. *Quality and Reliability Engineering International*, vol. 26, no. 7, pp. 663-675.

Crosier, R. (1988): Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, vol. 30, no. 3, pp. 291-303.

Desmarais, B.; Cranmer, S. (2012): Micro-level interpretation of exponential random graph models with application to estuary networks. *Policy Studies Journal*, vol. 40, no. 3, pp. 402-434.

Ebrahim, M.; Reza, B.; Rassoul, N.; Ghazaleh, R. (2017): A statistical approach to social network monitoring, *Communications in Statistics-Theory and Methods,* vol. 46, no. 22, pp. 11272-11288.

Healy, P.; Palepu, K. (2003): The fall of Enron. *Journal of Economics Perspectives*, vol. 17, no. 2, pp. 3.

Koutra, D.; Shah, N.; Vogelstein, J.; Gallagher B.; Faloutsos, C. (2016): DELTACON: A principled massive-graph similarity function and applications. *ACM Transactions on Knowledge Discovery from Data*, vol. 10, no. 3.

Liu, Y.; Liu, L.; Luo, J. (2011): Fast identifying steady phase of communication network based on network signature. *Journal of Computer Research and Development*, vol. 48 (Suppl.). pp. 67-72.

Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D. (2000): The mahalanobis distance, *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 1, pp. 1-18.

McCulloh, I.; Carley, K. (2008): Social network change detection. Carnegie Mellon University, School of Computer Science, Institute for Software Research, *Technical Report CMU-ISR-08-116*.

Qiu, P. (2018): Some perspectives on nonparametric statistical process control. *Journal of Quality* Technology, vol. 50, no. 1, pp. 49-65.

Ranshous, S.; Shen, S.; Koutra, D.; Faloutsos, C.; Samatova, N. F. (2015): Anomaly detection in dynamic networks: a survey. *WIREs Computational Statistics*, vol. 7, no. 3, pp. 223-247.

Snijders, T.; van de Bunt, G.; Steglich, C. (2010): Introduction to stochastic actor-based models for network dynamics. *Social Networks*, vol. 32, no. 1, pp. 44-60.

Wan, M.; Yao, J.; Jing, Y.; Jin, X. (2018): Event-based anomaly detection for non-public industrial communication protocols in SDN-based control systems. *Computers, Materials & Continua*, vol. 55, no. 3, pp. 447-463.

Wan, X.; Milios, E.; Kalyaniwalla, N.; Janssen, J. (2009): Link-based event detection in Email communication networks. *Proceeding of the 2009 ACM symposium on Applied Computing*, pp. 1506-1510.