MONITORING NETWORK DATA BY CONSIDERING THE CORRELATION OF NETWORK FEATURE STATISTICS

Panpan Zhou[1*], Dennis Lin[2], Xiaoyue Niu[2], and Zhen He[1†]
[1]College of Management and Economics,
Tianjin University, Tianjin, China
zhoupanpan@tju.edu.cn
zhhe@tju.edu.cn

[2]Department of Statistics,
The Penn State University, USA

ABSTRACT

Network monitoring has wide applications in computer and social network surveillance, pathological diagnosis in neuroscience and bioscience among others. Motivated by a real example of brain networks, we focus our interests on monitoring networks by considering correlations of feature statistics. Structural statistics of numbers of edges, stars and triangles, are adopted to summarize the main features of a network - density, degree variability, and transitivity. A multivariate chart is proposed to monitor the multiple statistics simultaneously, which has not been paid much attention to in previous studies. A simulation study is conducted to compare the performances of the multivariate chart and individual charts for the structural statistics as well as a model-based approach as a benchmark. The results show that the multivariate chart for the structural statistics perform well in most scenarios. In particular, it is more advantageous in timely detecting large shifts of connection propensity and degree variability locally and globally. A real case of monitoring Enron email networks is analyzed as an illustration.

Keywords: network monitoring, correlation, statistical process control, performance comparison, structural statistics, multivariate chart

* Corresponding Author

† Corresponding Author

1 INTRODUCTION

Networks are a type of data describing a set of connected entities. Many complex systems are modeled as networks. Typical examples include energy flows through food chains, synthesis and decomposition among cells, communications among friends, and data transmission through the Internet to name a few. Interest is rapidly growing in network monitoring for its wide applications in fraud detection, disease spreading control, detection of cells pathological changes, and computer network surveillance among others.

Network data are typically represented as adjacency matrices. Given a random network G with n nodes and its adjacency matrix Y, the element Y_{ij} is a binary variable indicating whether a link exists, or a variable quantifying the frequency or weights of the link between node i and node j. In graph theory, a link is also called an edge. The procedure of monitoring network data can be summarized into three steps based on Woodall et al. [1] and Savage et al. [2]. As shown in Figure 1, the first step is to aggregate raw data into network data by time or space intervals. While it is sometimes overlooked, the preprocessing step of data aggregation may significantly affect the monitoring performance for not only networks but also any type of data (Zwetsloot and Woodall [3] and Zhao et al.[4]). Especially, networks and matrices are usually not directly available. When measuring a network process or collecting observational network data, the raw data are often the links among different pairs of nodes, which might happen sequentially at different time points (e.g. email communication networks). With network data obtained through suitable aggregation, the next step is to decide the unit of interest and determine the network features. With respect to specific problems and targets, anomalies of node level, subgraph level and graph level can be of interest. Then the features of networks in corresponding levels should be determined and quantified. Two popular ways of representing the network features are (1) to directly subtract summary statistics of networks, and (2) to fit a model to networks and estimate the model parameters to summarize network features. The third step is to develop statistical methods for monitoring the network features. Various approaches including control chart and hypothesis testing methods, Bayesian methods, scan methods, time series model methods and others can be applied to retrospective analysis of historical data and monitoring online data (see Woodall et al. [1] for a thorough review).

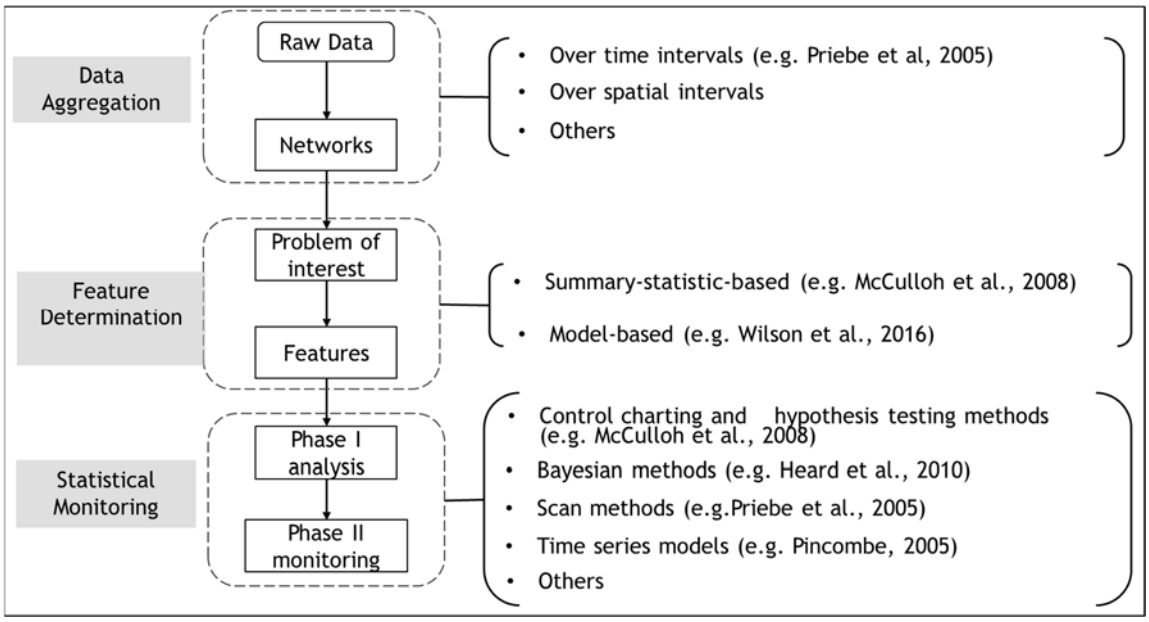

Figure 1: Procedures of network monitoring and methods summarized based on Woodall et al. [1] and Savage et al. [2]

From the perspective of the ways of representing network features, existing methods can be roughly categorized into summary-statistic-based approaches and model-based approaches.

Examples of summary-statistic-based approaches include the work of McCulloh and Carley [5, 6, 7], McCulloh et al. [8]. CUSUM or EWMA charts for centrality statistics including average betweenness and closeness are employed to detect global changes of networks. In Priebe et al. [9], scan method is used to detect outliers of k^{th} order neighborhood statistics by a two-step standardization moving window approach. Model-based approaches are to monitor the parameters of network models. An example is the work of Wilson et al. [10]. Shewhart charts are employed to monitor the estimates of parameters of degree corrected stochastic block models fitted to the network data (denoting this approach as the DCSBM-based approach). It is usually difficult to represent network features by only one summary statistic or one model parameter. As such, multiple features were adopted in previous studies as aforementioned. In existing literature, the network features were monitored by univariate charts separately. However, it is possible that these statistics are correlated. Strong correlations between the centrality statistics has already been proven empirically by Valente et al. [11] and mathematically by Li et al. [12]. As shown in Montgomery [13], monitoring correlated quality characteristics individually by univariate charts can be very misleading. The correlation between the network features has been much neglected and is addressed in this paper.

Various statistical methods can be adapted for monitoring network feature statistics. Here we mainly focus on statistical process monitoring methods. To fill up the research gap, we study statistical process monitoring for network data by considering the correlation among the statistics representing network features. The contributions of this paper are (1) emphasizing the correlation test before applying any statistical methods for network monitoring, (2) proposing multivariate charts for monitoring both undirected and directed networks through the summary statistics, and (3) providing a comparison between the multivariate chart method and the DCSBM-based approach and giving some advice for practical use of the charts.

The organization of this paper is as follows. In section 2, a motivating example is analysed. In section 3, a multivariate chart for the network statistics is proposed by considering the correlations of the structural statistics. The proposed approach is compared with individual Shewhart charts method and the DCSBM-based approach by a simulation experiment. The results show that overall, the Hotelling T2 chart performs better than the individual Shewhart charts and the DCSBM-based charts; the DCSBM-based charts perform better in detecting change of variance in particular. In section 5, a real example is provided to illustrate the proposed method.

2 A MOTIVATING EXAMPLE: BRAIN NETWORKS

Pathology detection in bioscience plays an important role in facilitating early treatment to prevent pathological area growing. It has been found that many brain disorders are associated with the abnormal topological structures of brain networks (Liu et al. [14]). For example, high degree nodes in functional MRI graphs are shown to have greater local deposition of amyloid protein than less topologically central brain regions for patients with Alzheimers disease (Buckner et al. [15]); node degree and other measures of topological centrality in functional connectivity networks are positively correlated with local grey matter atrophy across a range of neurodegenerative disorders (Zhou et al. [16]). Thus, monitoring brain networks through its feature statistics is of great use in pathology diagnosis. In this section, correlation of various feature measures and correlation of model parameters are explored based on a real dataset of brain networks of healthy subjects. The results show that strong correlation exists among different categories of summary statistics, motivating us to further study monitoring network features considering their correlations.

We use the data of brain connectivity structures in the dataset KKI-42 (Landman et al.[17]), which is further processed and analysed by Durante et al. [18]. Data are collected for 21 healthy subjects with no history of neurological disease under a scan-rescan imaging session. Each subject has been observed twice. Brain regions are constructed according to the Desikan et al. [19] atlas, for a total of 68 nodes equally divided in left and right hemispheres. The

matrices of the total number of white matter fibers connecting two brain regions are processed as undirected binary network data.

Compared with many social networks which tend to be time-varying, brain networks of different subjects are independent of each other. For such independent networks, we are interested in (1) what features of networks to monitor, (2) whether the features are correlated, and (3) how to monitor the correlated or uncorrelated features.

Selecting feature from a variety of measures of networks is not standardized and may vary from case to case. As commented by Savage et al. [2],

> *"…the lack of papers clearly describing the reasons for examining a particular set of features suggests to us that selection of a suitable feature space may be extremely difficult in practice".*

Here, both summary-statistic-based and model-based features are considered for a more complete understanding. We studied the popular centrality measures, the size of 1st and 2nd neighborhood as well as the transitivity measures from the Exponential Random Graph Model (ERGM) family. Those measures are either used for network monitoring in previous studies or well interpretable for network structures (e.g. McCulloh et al.[8]; Priebe et al.[9]; Snijders et al. [20]; and Fornito et al. [21]). Since global changes are more generalized and representative, we study the network measures from a graph level perspective and omit the node-level and subgraph-level characteristics. The total number of edges describes the overall density of a network and thus is included here. We studied the typical centrality measures such as betweenness and closeness. These two measures together with the sizes of 1st and 2nd order neighborhood are node-wise metrics. Therefore, their averages were taken over all nodes within networks as graph-level measures. The numbers of 2-stars and triangles, which will be further explained in a later section, are global summary statistics and were calculated. The adjacency matrices of the subjects show a pattern of nodes clustering into 2 blocks, corresponding to brain regions in the left and right hemispheres. Examples of the adjacency matrix plots are shown in Figure 2. As such, a block model with two communities is suitable for fitting the brain networks. To explore correlations among the model parameters for the brain networks, we adopted the degree-corrected stochastic block model, which is proposed by Karrer and Newman [22] and adapted to a dynamic version by Wilson et al. [10].

Figure 2: Adjacency matrices of brain networks of subject 16 for the first scan (left) and subject 3 for the second scan (right) with black representing an edge (adapted from Durante et al. [18])

We calculated the Pearson correlations to study the simplest linear relationship among the network features (Pearson [23]). Figure 3 is the plot of pairwise correlation among the network summary statistics. The correlation coefficients are shown in upper triangular part and the scatter plots are shown in the lower triangular part. The high correlation coefficient values (close to 1) and the linear shapes indicate strong pairwise linear correlations among the statistics, and the correlations are statistically significant with all p-values below 0.01. For the degree-corrected stochastic block models, index the two blocks of left and right hemispheres with L and R. We estimated four parameters P_{LL}, P_{LR}, P_{RR}, and δ for each network

because they are considered as the charting statistics for network monitoring in Wilson et al. [10]. The four parameters express the propensity of connection between nodes in left hemisphere, across left and right hemisphere, in right hemisphere, and the variability of the propensity of connection of the nodes, respectively. We calculated the average of the estimated parameters for the replicate networks from one subject. The correlations and scatter plots among P_{LL}, P_{LR}, P_{RR}, and δ are shown in Figure 4. The upper and lower triangular parts are the pairwise Pearson correlation coefficients, and the scatter plot, respectively. The diagonal part are the histograms of the parameter estimates. The *p*-values for the significance test of the correlations are shown in Table 1. At a 95% confidence level, only P_{RR}, and δ are negatively correlated. There is no evidence to reject the hypothesis that all other pairs of the parameters are not linear correlated.

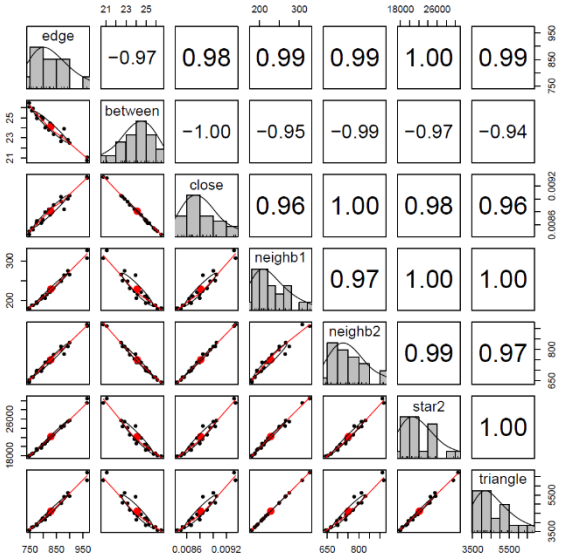

Figure 3: Correlation plot for the network summary statistics (from top to bottom: total number of edges, average betweenness, average closeness, average size of 1st order neighborhoods, average size of 2nd order neighborhoods, total number of 2-stars, and total number of triangles)

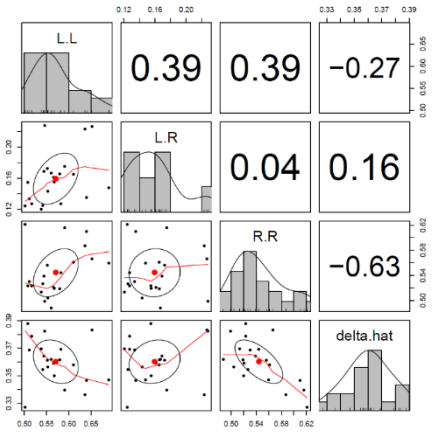

Figure 4: Correlation plot for the parameter estimates of fitted block models (from top to bottom: P_{LL}, P_{LR}, P_{RR}, and δ)

In this specific brain network case, we found strong correlations among the commonly used summary statistics representing network features. The correlations within centrality measures coincide with the conclusion drawn from empirical and theoretical studies by Valente et al. [11] and Li et al. [12]. Moreover, strong correlations exist within neighborhood size measures, within transitivity measures, and between the three classes of measures. It implies that monitoring multiple summary statistics without considering their correlations could be

misleading in this case. Although a general conclusion of network summary statistics being correlated cannot be drawn from one individual dataset, we conjecture correlations among the summary statistics exist in many real-world networks. More empirical studies and mathematical proofs can be explored on the relationships among the network summary statistics. We found significant correlation only between P_{RR}, and δ parameters of degree-corrected stochastic block models fitted to the brain networks. Individual charts are reasonable for monitoring the model parameters for this case.

Table 1: *P*-values for Pearson correlation tests for parameter estimates of block models for the brain network data

***p*-value**	P_{LL}	P_{LR}	P_{RR}	δ
P_{LL}		0.08	0.08	0.24
P_{LR}	0.08		0.86	0.50
P_{RR}	0.08	0.86		0.00
δ	0.24	0.50	0.00	

Motivated by this real example, we suggest a correlation test should be done for network feature statistics prior to network monitoring. Multiple network quality characteristics should be monitored simultaneously when they are not independent.

3 MONITORING NETWORKS CONSIDERING CORRELATION BETWEEN FEATURE STATISTICS

Monitoring feature statistics by individual charts could be misleading when they are shown correlated in statistical tests. In this section, we follow the procedures of network monitoring shown in Figure 1, discuss the selection of network features and propose a method of monitoring them considering their correlations.

3.1 Selection of network features

We focus on global changes of networks due to their generality. Counts of triadic structures are commonly used statistics for well characterizing global properties of networks (Frank and Strauss [24]; Holland and Leinhardt [25]). The numbers of edges, stars, and triangles are basic triadic structures. More complicated network statistics, which were extended from those star and triangle counts, were proposed for ERGM modeling (Snijders et al. [20]), which became a popular tool afterwards. The number of edges reflects the overall density of a network. The number of k-stars reflects the propensities for individual node to have connections with multiple network partners. Since introducing too many variables might decrease the power of the multivariate charts, the number of 2-stars is considered here and higher order star counts will not be included. The number of triangles reflects the transitive relationship.

Here we propose to monitor the statistics in a multivariate chart if their correlation is tested to be significant. The adjacency matrix for a random binary network G is Y. Its elements Y_{ij} equals 1 when a tie exists between node i and node j; and it equals 0 otherwise. According to Frank and Strauss [24], the statistics are

$$\begin{aligned} S_1(Y) &= \textstyle\sum_{1\le i<j\le n} Y_{ij}, \textit{ number of edges} \\ S_2(Y) &= \textstyle\sum_{1\le i<j\le n} \binom{Y_{i+}}{2}, \textit{ number of 2-stars} \\ T(Y) &= \textstyle\sum_{1\le i<j<h\le n} Y_{ij}Y_{jh}Y_{ih}, \textit{ number of triangles} \end{aligned} \quad \textbf{(1)}$$

where the + sign denotes summation over the index, and Y_{i+} is the degree of node i. For directed networks, Y_{ij} equals 1 when there is a tie from node i to node j. The number of edges equals the number of ingoing edges and equals the number of outgoing edges.

The counterparts of the 2-star statistics are the numbers of 2-in-stars and 2-out-stars. The number of the triangles contains the number of transitive triples and cyclic triples, corresponding to the sets of three edges ($i \rightarrow j$), and ($j \rightarrow k$) , and either ($i \rightarrow k$) or ($k \rightarrow i$). Thus, the statistics for directed networks are (Frank and Strauss [24]; Holland and Leinhardt [25])

$$\begin{aligned} S_1(Y) &= \textstyle\sum_{1\le i<j\le n}(Y_{ij} - Y_{ij}Y_{ji}), \textit{ number of edges} \\ S_2^{in}(Y) &= \tfrac{1}{2}\textstyle\sum_{i,j,h;j\neq h} Y_{ij}Y_{hi}, \textit{ number of 2-in-stars} \\ S_2^{out}(Y) &= \tfrac{1}{2}\textstyle\sum_{i,j,h;j\neq h} Y_{ij}Y_{ih}, \textit{ number of 2-out-stars} \\ T(Y) &= \textstyle\sum_{1\le i<j<h\le n} Y_{ij}Y_{jh}\,(Y_{ih} + \tfrac{1}{3}Y_{hj}). \textit{ number of triangles} \end{aligned} \quad \textbf{(2)}$$

3.2 Multivariate chart for global statistics

When the statistics are correlated to each other, two ways of monitoring them are (1) extracting principal components and applying individual charts to monitor the orthogonal principal components and the residuals, and (2) monitoring the statistics simultaneously by a multivariate chart. The number of edges, stars and triangles are very interpretable regarding the network structures, while principal components might be difficult to interpret in practice. Thus, we apply multivariate charts for network monitoring.

Write the statistics into a vector $\boldsymbol{S}(Y)$. Although $\boldsymbol{S}(Y)$ are counts, we can approximately assume $\boldsymbol{S}(Y)$ is normally distributed. Conventionally, statistical monitoring is classified into Phase I and Phase II monitoring. In Phase II, online monitoring is implemented for each individual network. Given m observations of networks $\{y^{(g)}\}_{g=1:m}$, $\boldsymbol{S}(y^{(g)})$ can be obtained by Equation (1) or (2) depending on the type of networks under study. Write $\boldsymbol{S}(y^{(g)})$ as S^g for simplicity. In Phase I, historical data are available to identify an in-control process and estimate process parameters. Given m observations of the networks, we can estimate μ and Σ as $\hat{\mu} = \frac{1}{m}\sum_{g=1}^{m} S^g$, and $\hat{\Sigma} = \frac{1}{m-1}\sum_{g=1}^{m}(S^g - \hat{\mu})(S^g - \hat{\mu})'$.The Hotelling T^2 statistic for S^g is

$$T_g^2 = (S^g - \hat{\mu})'\hat{\Sigma}^{-1}(S^g - \hat{\mu}). \quad \textbf{(3)}$$

When the process is in-control, T_g^2 follows a beta distribution and the control limits are

$$\begin{aligned} UCL &= \frac{(m-1)^2}{m}\beta_{\alpha,\ d/2,(m-d-1)/2'} \\ LCL &= 0, \end{aligned} \quad \textbf{(4)}$$

Where d is the degree of freedom, equal to 3 for undirected networks and 4 for directed; α is the false alarm probability; $\beta_{\alpha,\ d/2,(m-d-1)/2}$ is the upper α percentage point of the central beta distribution with parameters $d/2$ and $(m-d-1)/2$ (Tracy et al.[25]). Remove network g from the samples if $T_g^2 >$ UCL. Repeat the estimation and outlier detection procedures until only $m^*\alpha$ networks are shown out-of-control.

4 PERFORMANCE COMPARISON

We simulate network data based on degree corrected stochastic block models following the same settings as in Wilson et al. [10]. Since they directly monitor parameters of the stochastic block models, DCSBM-based charts serve well as a benchmark. The aim here is to evaluate and compare the performances of Hotelling T^2 charts and Shewhart charts of the statistics, as well as Wilson et al. [10]'s DCSBM-based charts.

We generate m = 1000 undirected networks as Phase I samples. Each network has n = 100 nodes, and k = 2 equally sized communities. We set the connectivity matrix $P = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$, and the standard deviation of the degree parameters δ_j = 0.5 for j= 1,2. For Phase II data, changes are implemented at time t^* = 25 and thereafter networks are generated as many as it took to observe the first signal on each control chart. A total of 7 scenarios are implemented as shown in Table 2. For each control chart in each scenario, we run simulations 1000 times and obtain 1000 run length values.

Table 2: Simulation settings by the dynamic DCSBMs method in Wilson et al. [10]

Scenario	Change	Description
0	No change	no structural change has occurred
1	$P^*_{1,1} = P_{1,1} + \epsilon$, ϵ = 0.01, 0.05, 0.10	local outbreak in community 1
2	$P^* = P + \epsilon$, ϵ = 0.01, 0.05, 0.10	global outbreak
3	$\delta_1^* = \delta_1 + \lambda$, λ = 0.05, 0.10, 0.25	increase of local variability in community 1
4	$\delta^* = \delta + \lambda$, λ = 0.05, 0.10, 0.25	increase of global variability
5	merge communities	merge communities
6	split communities	split community 1 into 2 communities

The average run lengths (ARLs) are obtained for individual charts for the four parameters δ, $P_{1,1}$, $P_{1,2}$, and $P_{2,2}$, respectively in Wilson et al. [10] as shown in Table 3. In statistical process monitoring practice, any one of the four individual charts signalling indicates an anomaly. As such, we calculate the ARLs for the combined use of the four charts based on the equation $1 - \frac{1}{ARL_{combined}} = (1 - \frac{1}{ARL_{P_{1,1}}})(1 - \frac{1}{ARL_{P_{1,2}}})(1 - \frac{1}{ARL_{P_{2,2}}})(1 - \frac{1}{ARL_{\delta}})$. The result is shown in the last column in Table 3. In this simulation study, we obtain ARL values for individual charts for the number of edges $S_1(Y)$, the number of 2-stars $S_2(Y)$ and the number of triangles $T(Y)$ separately, corresponding control limits were set as $\hat{\mu} \pm 3\hat{\sigma}$. For the T^2 chart, we obtain the *UCL* based on Phase I samples with type I error α = 0.9973. The ARL results are shown in Table 3 named as T^2. The best performance among the individual charts are marked in bold. The best performance among all charts in each scenario is italicized and underlined.

From Table 3, we see the ARLs of the charts of number of edges, T^2 and $P_{1,2}$ have lower false alarm probability when the process has no change. While each of the DCSBM-based charts show a reasonable ARL value for the in-control case when evaluated separately, the combined use of the four charts has an ARL of 94.49, substantially increasing the false alarm probability. Regardless of its over-sensitivity in the in-control case, the combined use of the four parameter charts largely improves the performance for anomaly detection when the process has a small shift. By contrast, it doesn't make much contribution when the shifts are large. Among the individual charts of numbers of edges, 2-stars and triangles, the edge count chart performs better when the process is in-control while the triangle count chart over-alarms. It partially accounts for the overall best detection power of the triangle count chart for process shifts. From the underlined values of the chart of number of triangles, we found that overall, the Shewhart charts of the count statistics can more timely detect the change of connection propensity locally and globally. All four types of charts perform almost equally well when the

shift of propensity parameter is large. The combined use of DCSBM parameter charts provides a better performance in detecting small shifts. For changes of degree variability, T^2 chart and the combined DCSBM parameter charts show a better performance. Especially, T^2 chart performs better in detecting large shifts, and the combined DCSBM parameter charts performs better in detecting small shifts. For the merging of communities, the DCSBM-based charts and their combined use is obviously more advantageous. For the splitting of communities, the Shewhart chart of edge count, the T^2 chart, the $P_{1,1}$ chart and the combined use of DCSBM parameter charts show comparable performances.

In summary, the Shewhart charts performs well in detecting changes of propensity of connections in local and global communities. The $P_{1,1}$, $P_{1,2}$, and δ charts performs well in detecting local outbreak, global outbreak, and the changes of degree variability, respectively. The combined use of the DCSBM parameter charts can boost the detection power when the process has a small shift with a cost of much higher false alarm probability. The T^2 chart of numbers of edges, 2-stars and triangles has an overall reasonably good performance. In particular, it performs very well in detecting anomalies when connection propensity or degree variability has large shifts.

Table 3: ARL results for Shewhart charts and T^2 chart of numbers of edges, 2-stars and triangles, DCSBM-based parameter charts as well as the combined use of the DCSBM-based parameter charts

Scenario	edges	2-stars	triangles	T^2	δ	$P_{1,1}$	$P_{1,2}$	$P_{2,2}$	Combined
none	**375.43**	340.04	283.64	445.32	317.18	439.25	***446.50***	338.25	94.49
$P_{1,1}$ + 0.01	251.15	177.54	**126.94**	226.05	294.80	**134.00**	413.70	332.40	*61.77*
$P_{1,1}$ + 0.05	27.01	19.07	**11.57**	17.46	284.90	**9.87**	257.27	207.70	*8.91*
$P_{1,1}$ + 0.10	4.71	3.20	***2.12***	2.42	524.40	**2.23**	289.9	325.90	2.21
P + 0.01	34.85	26.22	***21.41***	40.51	498.80	140.9	**64.65**	142.30	31.98
P + 0.05	1.10	1.06	***1.04***	1.08	211.10	9.48	**1.71**	12.17	1.51
P + 0.10	1.10	1.07	**1.05**	*1.00*	93.30	2.01	**1.01**	2.28	1.00
δ_1 + 0.05	277.33	233.67	**175.33**	156.39	**106.51**	221.40	260.10	202.70	*44.44*
δ_1 + 0.10	201.86	176.38	**111.31**	69.06	**115.70**	152.33	305.29	544.60	*49.56*
δ_1 + 0.25	89.94	67.67	**30.77**	*6.51*	**18.81**	63.35	107.20	431.00	12.67
δ + 0.05	216.17	178.49	**109.22**	72.38	**93.58**	232.30	246.10	216.10	*42.58*
δ + 0.10	120.37	99.97	**55.22**	*16.35*	**36.33**	142.00	185.94	218.50	22.75
δ + 0.25	38.03	28.77	**11.30**	*1.35*	**4.94**	52.88	92.23	53.87	4.16
Merge	249.19	**181.37**	205.06	121.55	247.00	39.79	**1.66**	27.61	*1.59*
Split	**36.31**	99.04	582.35	36.35	127.50	**33.90**	313.39	426.20	*23.57*

5 AN ILLUSTRATIVE EXAMPLE

Enron email communication network data have been widely studied. The version adopted here is from Priebe et al. [9]. The dataset contains 184 unique email addresses of about 150 users (mostly executives and some assistants and traders). With the email addresses being nodes, a directed edge exists if there is no less than one email from the sender to the receiver in a week. A total of 189 directed binary networks are obtained by aggregating the data by weeks from November, 1998 to June, 2002.

The numbers of edges, 2-in-stars, 2-out-stars, and triangles are used to characterize the network structures. They represent the overall density of communications, the degree variability of senders, the degree variability of receivers and the amount of transitive communications among a local triad. A Hotelling T^2 chart for the numbers of edges, 2-in-stars, 2-out-stars, and triangles is shown in Figure 5. Outliers are detected in the weeks around dates of 2001-04-30, 2001-05-21, 2001-08-20, 2001-10-01, 2001-10-08, 2001-10-22, 2001-10-29, 2001-11-12, 2002-01-28, and 2002-02-04. The first outlier signals before the critical point of its stock price on May 5, 2001. The second alarm corresponds to the reaction to the selling of 1.1 million stock shares by the chief executive of Enron Xcelerator Lou Pai on May 18. Many outliers appear from October to November, 2001, consistent with the period when large amount of Enron shares was sold at the end of September and Enron reported a $618 million loss on October 16, 2001, until it was under a formal investigation from SEC on October 31, 2001. The suicide of the former Enron Executive J. Clifford Baxter on January 25, 2002 results in a suddenly increased connections among the executives after a period of stable communications.

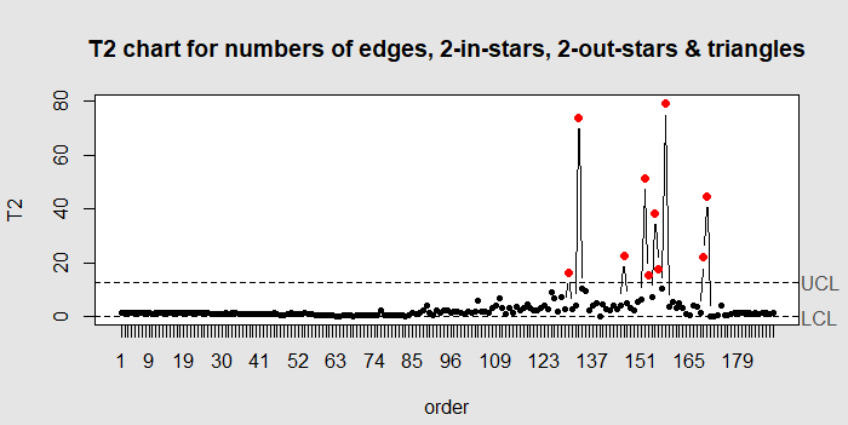

Figure 5: T^2 chart of numbers of edges, 2-stars and triangles

6 CONCLUSION

Monitoring network feature statistics by individual charts without considering their correlations might be misleading. It should be checked whether correlations exist among the feature statistics before applying network control charts. Characterizing network global properties by the number of edges, 2-stars and triangles, a T^2 chart of these feature statistics show a good balance of in-control performance and out-of-control performance. Compared with individual charts of the structural statistics and the DCSBM-based parameter charts, the T^2 chart shows an overall competitive performance and significant advantages in detecting large shifts of connection propensity and degree variability globally.

7 REFERENCES

[1] **Woodall, W. H., Zhao, M. J., Paynabar, K., Sparks, R., and Wilson, J. D.** 2017. An overview and perspective on social network monitoring. *IISE Transactions*, 49, pp 354-365.

[2] **Savage, D., Zhang, X., Yu, X., Chou, P., and Wang, Q.** 2014. Anomaly detection in online social networks. *Social Networks*, 39, pp 62-70.

[3] **Zwetsloot, I. M. and Woodall, W. H.** 2018. A review of some sampling and aggregation strategies for basic statistical process monitoring. *Submitted for publication*.

[4] **Zhao, M. J., Driscoll, A. R., Sengupta, S., Stevens, N. T., Fricker, R.D., J., and Woodall,W. H.** 2018. The effect of data aggregation level in social network monitoring. *Submitted for publication*.

[5] **McCulloh, I. and Carley, K. M.** 2008. Dynamic network change detection. *Technical report,* Military Academy West Point NY.

[6] **McCulloh, I. and Carley, K. M.** 2008. Social network change detection. *Technical report*.

[7] **McCulloh, I. and Carley, K. M.** 2011. Detecting change in longitudinal social networks. *Technical report*, Military Academy West Point NY Network Science Center (NSC).

[8] **McCulloh, I., Webb, M., Graham, J., Carley, K., and Horn, D. B.** 2008. Change detection in social networks. *Technical report*, Military Academy West Point NY Dept of Mathematical Sciences.

[9] **Priebe, C. E., Conroy, J. M., Marchette, D. J., and Park, Y.** 2005. Scan statistics on Enron graphs. *Computational & Mathematical Organization Theory*, 11, pp 229-247.

[10] **Wilson, J. D., Stevens, N. T., and Woodall, W. H.** 2016. Modeling and estimating change in temporal networks via a dynamic degree corrected stochastic block model. arXiv preprint arXiv:1605.04049.

[11] **Valente, T. W., Coronges, K., Lakon, C., and Costenbader, E.** 2008. How correlated are network centrality measures? *Connect*, 28, pp 16-26.

[12] **Li, C., Li, Q., Van Mieghem, P., Stanley, H. E., and Wang, H.** 2015. Correlation between centrality metrics and their application to the opinion model. *The European Physical Journal B*, 88, pp 65-78.

[13] **Montgomery, D. C.** 2009. *Introduction to statistical quality control*. 7th Edition, John Wiley & Sons.

[14] **Liu, J., Li, M., Pan, Y., Lan, W., Zheng, R., Wu, F. X., and Wang, J.** 2017. Complex brain network analysis and its applications to brain disorders: A survey. *Complexity*, 2017, pp 1-27.

[15] **Buckner, R. L., Sepulcre, J., Talukdar, T., Krienen, F. M., Liu, H., Hedden, T., Andrews-Hanna, J. R., Sperling, R. A., and Johnson, K. A.** 2009. Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to alzheimer's disease. *The Journal of neuroscience*, 29, pp 1860-1873.

[16] **Zhou, J., Gennatas, E. D., Kramer, J. H., Miller, B. L., and Seeley, W. W.** 2012. Predicting regional neurodegeneration from the healthy brain functional connectome. *Neuron*, 73, pp 1216-1227.

[17] **Landman, B. A., Huang, A. J., Giord, A., Vikram, D. S., Lim, I. A. L., Farrell, J. A., Bogovic, J. A., Hua, J., Chen, M., Jarso, S., et al.** 2011. Multi-parametric neuroimaging reproducibility: a 3-t resource study. *Neuroimage*, 54,pp 2854-2866.

[18] **Durante, D., Dunson, D. B., and Vogelstein, J. T.** (2017). Nonparametric bayes modeling of populations of networks. Journal of the American Statistical Association, 112, pp 1516-1530.

[19] **Desikan, R. S., Sgonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J.** 2006. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, 31, pp 968-980.

[20] **Snijders, T. A., Pattison, P. E., Robins, G. L., and Handcock, M. S.** 2006. New specications for exponential random graph models. *Sociological methodology*, 36, pp 99-153.

[21] **Fornito, A., Zalesky, A., and Bullmore, E. T.** 2016. *Fundamentals of Brain Network Analysis*. Academic Press, San Diego.

[22] **Karrer, B. and Newman, M. E. J.** 2011. Stochastic blockmodels and community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 83, 016107.

[23] **Pearson, K.** (1895). Vii. note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58, pp 240-242.

[24] **Frank, O. and Strauss, D.** 1986. Markov graphs. *Journal of the American Statistical Association*, 81, 832-842.

[25] **Holland, P.W., & Leinhardt, S.** 1981. An exponential family of probability-distributions for directed-graphs. *Journal of the American Statistical Association*, 76(373), 33–50.

[26] **Tracy, N., Young, J., and Mason, R.** 1992. Multivariate control charts for individual observations. *Journal of Quality Technology*, 24, pp 88-95.