Received: 31 August 2016 | Revised: 21 October 2017 | Accepted: 2 January 2018

DOI: 10.1002/qre.2273

RESEARCH ARTICLE

WILEY

Phase I monitoring of social networks based on Poisson regression profiles

Hatef Fotuhi | Amirhossein Amiri | Mohammad Reza Maleki

Department of Industrial Engineering, Faculty of Engineering, Shahed University, Tehran, Iran

Correspondence
Amirhossein Amiri, Department of Industrial Engineering, Faculty of Engineering, Shahed University, Tehran, Iran.
Email: amiri@shahed.ac.ir

Abstract

Nowadays, due to the increasing role of social networks in our daily life, monitoring and forecasting social trends have attracted the attention of many researchers. To the best of the authors' knowledge, the literature includes few studies of monitoring social networks. Existing researches have focused on analyzing only the existence of communications between people and have neglected to monitor the number of such communications. In this paper, first counts of communications between people are modeled using Poisson regression profiles. Then, 3 Phase I monitoring methods, extended T^2, F, and a standardized likelihood ratio test method is suggested to detect step changes, drift, and outliers in the parameters of Poisson regression profiles. The proposed methods are evaluated via simulation studies in terms of signal probability criterion. The results show that in most out-of-control situations the standardized likelihood ratio test method outperforms the T^2 and F methods. Then, a numerical example and a case study based on Enron email data are presented to illustrate the application of the extended methods.

KEYWORDS

Phase I, Poisson regression profile, signal probability, social network, statistical process monitoring

1 | INTRODUCTION

In recent years, social networks have attracted the attention of academic and industrial researchers, especially in the area of statistical process monitoring. Monitoring communications within networks is essential for analysis of social issues such as terrorism and crime. Governments are interested in methods that can provide them with information about terrorist attacks and increases in communications in particular social groups. Social network monitoring can also be used to manage crises, such as fires in populated areas or early diagnosis of diseases. Analyzing and monitoring social network communications are highlighted in security issues for early detection of threats of terrorist attacks or damage to sites and facilities. This helps security agencies to focus their resources. Large numbers of communications and increases in the flow of information between different groups make these networks complex. One goal of researchers in this area is to find simple and accurate methods for monitoring communications in social networks.

In recent years, several studies have been conducted in the area of monitoring communications in social networks. They can generally be classified into 2 categories: monitoring fixed (static) networks and monitoring dynamic networks. In static networks, the number of nodes is constant and does not change over time. In dynamic networks, the number of nodes may change over time. Static networks have been studied by Erdos and Renyi,[1] Barabasi and Albert,[2] Leskovec et al,[3] Chakrabarti et al,[4] and Pennock et al.[5] Various methods have been used to analyze static networks. Exponential random

graph models (ERGM) were used by Frank and Strauss[6]; latent space models by Hoff et al[7]; attributed networks by Kim and Leskovec[8]; and random effect models by Hoff.[9] In dynamic networks, nodes and edges can be added or deleted over time. Dynamic networks have been studied by Snijders,[10] Hanneke et al,[11] Sarkar and Moore,[12] Ho et al,[13] Xu and Hero,[14] McCulloh and Carley,[15] and Park et al.[16] Scan statistics were used by Priebe et al,[17] Marchette[18], Neil,[19] and Sparks.[20] In some applications, the quality of the process can be described by the relationship between a response variable and 1 or more independent variables, called a profile. Practical applications of profiles have been discussed by researchers such as Kang and Albin,[21] Mahmoud and Woodall,[22] and Amiri et al.[23] Monitoring of different profiles has also been addressed in many studies, such as those by Kim et al[24] and Mahmoud et al.[25]

One of the models used to monitor social networks is generalized linear model-based profiles, which is proposed by Azarnoush et al[26] based on logistic regression profiles for Phase II monitoring of the existence of communications between nodes. Their proposed method has no ability to detect anomalies in the average counts of communications between people who are currently in contact with each other. However, in some practical situations, it is important to monitor the counts of communications between the nodes of a given social network. Therefore, to overcome this problem, Poisson regression profiles are proposed to model counts of communications rather than just the existence of communications. To accomplish this objective, for any attributes of 2 given nodes, a new index is presented and considered as an explanatory variable. Then, 3 methods, T^2, F, and standardized likelihood ratio test (SLRT), are proposed to monitor the parameters of Poisson regression profiles in Phase I. To illustrate the advantages of the proposed method compared with that provided by Azarnoush et al,[26] an example with 5 nodes is presented and illustrated in Figure 1. In the first network (Figure 1A), only connections are considered, while in the second (Figure 1B), along with the existence of communications, counts of communications are also provided and reported on the edges. Obviously, if the number of communications increases between 2 nodes, the network in Figure 1A still shows an in-control state; however, based on the proposed methods, the social network in Figure 1B would be out of control. This case involves increasing probability of Type II errors when the method proposed by Azarnoush et al[26] is used.

The remainder of this paper is organized as follows. In Section 2, social networks are discussed, and the proposed modeling procedure is described. In Section 3, Poisson regression profiles are briefly described. In Section 4, 3 methods, T^2, F, and SLRT, are applied to monitor the counts of communications in social networks. Section 4 is devoted to simulation studies and comparison between the methods. In Section 6, a numerical example is given to illustrate the proposed methods. Finally, concluding remarks and a recommendation for future research are presented in Section 7.

2 | PROBLEM DEFINITION AND MODEL ASSUMPTIONS

In this section, social networks modeling and notations are discussed. The notations and definitions used to formulate the problem are presented in Table 1.

A social network can be described in the form of a network relationship matrix. The notations to characterize a social network are presented by the following equations:

$$\begin{aligned} G(t) &= (V(t), Y(t)); t = 1, ..., T \\ V(t) &= \{v_1, v_2, ..., v_i, ..., v_n\} \\ Y(t) &= \left\{y_{12t}, y_{13t}, ..., y_{ijt}, ..., y_{n-1,n,t}\right\}, \end{aligned} \tag{1}$$

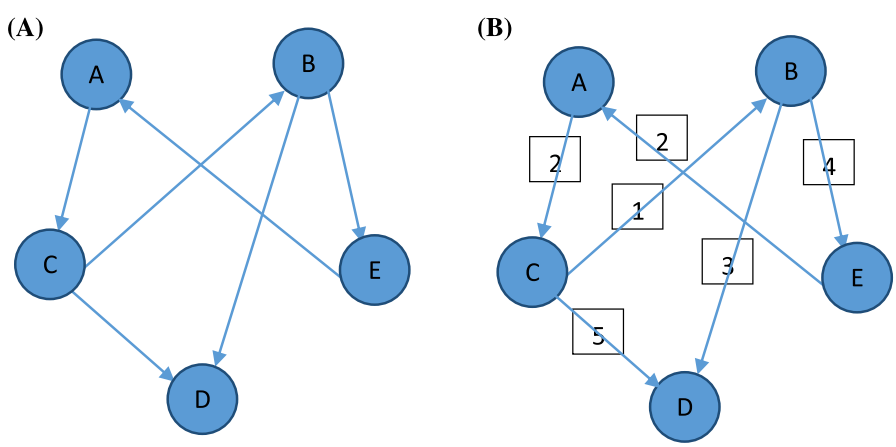

FIGURE 1 Comparing 2 social networks: A, considering just the existence of a communication; and B, considering counts of communications [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 The notations used in modeling of the problem

Notation	Description
i	Set of nodes
j	Set of nodes
k	Set of attributes
n	Number of nodes
l	Number of communications between any 2 nodes
T	Number of time periods
p	Number of attributes
a_{ik}	Value of kth attribute at i^{th} node
x_{ijk}	Value of explanatory variable between i and j for kth attribute
$\boldsymbol{\beta}_t = (\beta_{0t}, \beta_{1t}, ..., \beta_{pt})$	Vector of Poisson regression parameters at time t when average count of communications is in control
$\boldsymbol{\beta}_t'$	Vector of Poisson regression parameters at time t when average count of communications is out of control
δ	Parameter of step change
ζ	Parameter of linear trend
θ	Parameter of outlier
y_{ijt}	Number of contacts between nodes i and j at t^{th} time period
λ_{ijt}	Expected value for count of communications between nodes i and j at time t
$\widehat{\boldsymbol{\beta}}_t$	Estimated value of model parameters at t^{th} time period
$\widehat{\lambda}_{ijt}$	Estimated value of λ_{ij} at t^{th} time period

where $V(t)$ and $Y(t)$ represent nodes and edges in time period t, respectively. To define the edges, it is necessary to first recognize who a given person (node) communicates with in period $t = 1, 2, ..., T$. Note that, in this paper, time periods are defined as 1 week. Relationships are defined as any possible communications, such as emails, phone calls, and SMS. A matrix is defined whose diagonal elements are considered equal to zero, because people cannot communicate with themselves. These elements may be a nonzero number if relationships between individuals and their immediate family members are considered. The number of communications between the nodes of the network are presented by matrix $\mathbf{Y} = [y_{ijt}]_{n \times n}$, where y_{ijt}; $i \neq j$ is the number of communications between nodes i and j during the period t, while y_{iit} denotes the number of contacts of node i with closed family. $\boldsymbol{\lambda} = [\lambda_{ijt}]_{n \times n}$ is also defined as the matrix containing the expected values for the counts of communications between the nodes. It is assumed that the number of communications between nodes in different time periods follows a Poisson distribution as in Equation 2:

$$y_{ijt} \sim poisson(\lambda_{ijt}); i = 1, ..., n; j = 1, ..., n; t = 1, ..., T. \quad (2)$$

To model the social network as a profile, the counts of communications between nodes at time period $t = 1, .., T$ are considered the response variables. Consequently, the Poisson regression profiles can be used to model the relationship between the response and explanatory variables at any time period. Suppose that the node attributes are characterized by matrix $\mathbf{A} = [a_{ik}]_{n \times p}$, where a_{ik} denotes the value of the kth attribute at the ith node. Obviously, at t^{th}, $t = 1, .., T$ time period, there is a total of $l = \binom{n}{2}$ possible communications between any 2 nodes i and j; $i \neq j$, where n is the number of nodes. The matrix of explanatory variables $\mathbf{X}$ (attributes index) is defined as Equation 3:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{121} & x_{122} & \cdots & x_{12p} \\ 1 & x_{131} & x_{132} & \cdots & x_{13p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n(n-1)1} & x_{n(n-1)2} & \cdots & x_{n(n-1)p} \end{pmatrix}_{n \times (p+1)}, \quad (3)$$

where

$$x_{ijk} = \frac{\min\{a_{ik}, a_{jk}\}}{\max\{a_{ik}, a_{jk}\}}; i = 1, ..., n; j = 1, ..., n; k = 1, ..., p. \quad (4)$$

Equation 4 is used when the attributes of each node are numerical. This equation is used to compute the explanatory variable of an edge based on the attributes of the corresponding nodes. If the attributes are nominal or ordinal, a new explanatory variable is defined based on the attributes of the corresponding nodes and the values of 1,2,... are assigned to this new variable. Assume the attribute "gender" for 2 nodes. Because gender includes male and female, the new variable categories and the corresponding values would be as shown in Table 2. Then, a Poisson regression profile is used to model the average counts of communications between people and the new variable. As an alternative method, one can use a Poisson regression model with indicator explanatory variables to model this relationship.[27]

TABLE 2 New variable categories and corresponding values

New variable categories	M-M	M-F	F-F
Values	1	2	3

3 | POISSON REGRESSION MODEL

This section describes the Poisson regression model for characterizing a social network. For nodes i and j, the vector of explanatory variables considering p attributes is given as follows:

$$\mathbf{x_{ij}} = \left(x_{ij1}, x_{ij2}, .., x_{ijp}\right)^T; i = 1, ..., n, j = 1, ..., n. \quad (5)$$

Recall that y_{ijt} is the count of communications between the ith and j^{th} nodes at the tth profile (time period). This is assumed to follow a Poisson distribution as y_{ijt}~$poisson(\lambda_{ijt})$; $i = 1, ..., n; j = 1, ..., n; t = 1, ..., T$ with the following probability distribution function:

$$f_{y_{ijt}}\left(y_{ijt}\right) = e^{-\lambda_{ijt}}\frac{\lambda_{ijt}{}^{y_{ijt}}}{y_{ijt}!}; y_{ijt} = 0, 1, ...; \lambda_{ijt} > 0, \quad t = 1, ..., T. \quad (6)$$

The value of parameter λ_{ijt} depends on the vectors of $\boldsymbol{\beta}_t = (\beta_{0t}, \beta_{1t}, ..., \beta_{\text{pt}})$ and $\mathbf{x}_{ij} = (1, x_{ij1}, x_{ij2}, .., x_{ijp})$ as $\ln(\lambda_{ijt}) = \beta_{0t} + \beta_{1t}x_{ij1} + ... + \beta_{pt}x_{ijp}$ where β_{0t} denotes the model intercept. Obviously, the mean of the count between nodes i and j at a given network (profile) is $\lambda_{ijt} = \exp\left(\mathbf{X}_{ij}\boldsymbol{\beta}_t^T\right)$. Then, the joint likelihood function of y_{ijt} can be written as:

$$L(\boldsymbol{\lambda}; \mathbf{y}) = \prod_{i=1}^{n}\prod_{j=i+1}^{n} e^{-\lambda_{ijt}}\frac{\left(\lambda_{ijt}\right)^{y_{ijt}}}{y_{ijt}!} = \frac{\prod_{i=1}^{n}\prod_{j=i+1}^{n}\left(\lambda_{ijt}\right)^{y_{ijt}}}{\prod_{i=1}^{n}\prod_{j=i+1}^{n} y_{ijt}!} e^{-\sum_{i=1}^{n}\sum_{j=i+1}^{n}\lambda_{ijt}}, \quad t = 1, ..., T. \quad (7)$$

Taking the natural logarithm from Equation 7 leads to:

$$\ln[L(\boldsymbol{\lambda}; \mathbf{y})] = \sum_{i=1}^{n}\sum_{j=i+1}^{n} y_{ijt}\ln\left(\lambda_{ijt}\right) - \sum_{i=1}^{n}\sum_{j=i+1}^{n}\lambda_{ijt} - \sum_{i=1}^{n}\sum_{j=i+1}^{n}\ln\left(y_{ijt}!\right), t = 1, ..., T. \quad (8)$$

By replacing λ_{ijt} with $\mathbf{X}_{ij}\boldsymbol{\beta}_t^T$, Equation 8 at time t can be rewritten as:

$$\log[L(\boldsymbol{\beta}; \mathbf{y})] = \sum_{i=1}^{n}\sum_{j=i+1}^{n} y_{ijt}\log\left(\exp\left(\mathbf{X}_{ij}\boldsymbol{\beta}_t\right)\right) - \sum_{i=1}^{n}\sum_{j=i+1}^{n}\exp\left(\mathbf{X}_{ij}\boldsymbol{\beta}_t\right) - \sum_{j=1}^{n}\sum_{j=i+1}^{n}\log\left(y_{ijt}!\right), \quad t = 1, ..., T \quad (9)$$

then

$$\frac{\partial\log[L(\boldsymbol{\lambda}; \mathbf{y})]}{\partial\beta} = \mathbf{X}^T(\mathbf{y} - \boldsymbol{\lambda}). \quad (10)$$

The maximum likelihood estimation of $\boldsymbol{\beta}_t$ is the solution of $\mathbf{X}^T(\mathbf{y} - \boldsymbol{\lambda}) = \mathbf{0}$ where $\mathbf{0}$ is a p-dimensional zero vector. The vector of regression parameters can be estimated by utilizing the iterative weighted least squares method. For the tth time period, the estimated model parameters obtained by iterative weighted least square are denoted by $\widehat{\boldsymbol{\beta}}_{\mathbf{t}} = \left(\widehat{\beta}_{0t}, \widehat{\beta}_{1t}, .., \widehat{\beta}_{pt}\right)^T$. For more information about estimating parameters in Poisson regression profiles, refer to Amiri et al.[28]

4 | PROPOSED METHODS

To monitor a social network based on Poisson regression profiles, 3 methods, Hotelling's T^2, F, and SLRT, are extended and utilized in this section. Recall that there are T profiles (representing T time intervals) where each time period has l treatments (l pairwise relationships between nodes).

4.1 | Extended Hotelling's T^2

Yeh et al[29] proposed 5 charts based on Hotelling's T^2 to monitor logistic regression profiles in Phase I. They concluded that the T_I^2 method performed better than the others. This statistic was then used by Amiri et al[28] to monitor Poisson regression profiles. Due to the satisfactory performance of the T_I^2 statistic, it is used in the present study to monitor a social network. For the tth network, this statistic is computed as:

$$T_{I,t}^2 = \left(\widehat{\boldsymbol{\beta}}_t - \overline{\boldsymbol{\beta}}\right)^T \mathbf{S}_I^{-1}\left(\widehat{\boldsymbol{\beta}}_t - \overline{\boldsymbol{\beta}}\right); t = 1, 2, ..., T, \quad (11)$$

where

$$\overline{\boldsymbol{\beta}} = \frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{\beta}}_t \text{ and } \mathbf{S_I} = \frac{1}{T}\sum_{t=1}^{T}\text{var}\left(\hat{\boldsymbol{\beta}}_t\right) = \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{X^T}\hat{\mathbf{W}}_t\mathbf{X}\right)^{-1}, \quad (12)$$

and

$$\hat{\mathbf{W}}_{\mathbf{t}} = \begin{bmatrix} \hat{\lambda}_{12t} & 0 & \cdots & 0 \\ 0 & \hat{\lambda}_{13t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\lambda}_{(n-1)nt} \end{bmatrix}_{l\times l}, \hat{\lambda}_{ijt} = \exp\left(\mathbf{X}_{ij}\hat{\boldsymbol{\beta}}_t\right). \quad (13)$$

4.2 | Extended F-method

In this method, the indicator variables are used, and T profiles are converted to a profile of size $N = lT$ where l is the number of communications between any pairwise nodes. $T - 1$ indicator variables are defined as follows:

The relation $T' = T - 1$ is set, and the following regression model is fit to the data as follows:

$$Z_{it} = \begin{cases} 1 & \text{if } i\text{th observation is from the social network at time t} \\ 0 & \text{otherwise} \end{cases}; i = 1, 2, ..., N, t = 1, 2, ..., T-1. \quad (14)$$

$$\begin{aligned} g(\lambda_i) &= \beta_{10}x_{i1} + ... + \beta_{p0}x_{ip} \\ &\quad + Z_{1i}\left(\beta_{11}x_{i1} + ... + \beta_{p1}x_{ip}\right) + ... \\ &\quad \times + Z_{T'i}\left(\beta_{1T'}x_{i1} + ... + \beta_{pT}x_{ip}\right); \\ &\quad i = 1, 2, ..., N. \end{aligned} \quad (15)$$

Note that $g(\lambda_i)$ is the link function considering both the explanatory and indicator variables defined in Equation 15. To calculate these coefficients, the following regression model is fitted for each profile.

$$g(\lambda_i) = \beta_1 x_{i1} + ... + \beta_p x_{ip}, i = 1, 2, ..., N \quad (16)$$

The null hypothesis of above equation is as follows:

$$H_0{:}\beta_{11} = ... = \beta_{p1} = ... = \beta_{1T'} = ... = \beta_{pT'} = 0. \quad (17)$$

Then, the F statistic is calculated by using Equation 18, as

$$F = \frac{\{SSE(reduced) - SSE(full)\}}{(p(m-1))MSE(full)}, \quad (18)$$

where SSE (reduced) is equal to the reduced sum of squares error ($\sum_{i=1}^{N}\left(y_i - \widehat{\lambda}_i\right)^2/\widehat{\lambda}_i$) and SSE (full) (full sum of squares error) is calculated by $\sum_{t=1}^{T}\sum_{i=1}^{l}\left(y_{it} - \widehat{\lambda}_{it}\right)^2/\widehat{\lambda}_{it}$.

For large values of Poisson parameters, the F statistic in Equation 18 approximately follows a Fisher distribution with $(T-1)p$ and $T(n-p)$ degrees of freedom. When the value of this statistic exceeds $UCL = F_{\alpha,\,(T-1)p,\,T(n-p)}$, the null hypotheses with a $100(1-\alpha)\%$ confidence level is rejected.

4.3 | Standardized likelihood ratio test method

Suppose that a change occurs in 1 or more generalized linear model parameters, then:

$$\begin{aligned} g(\lambda_{ijt1}) &= \beta_{01} + \beta_{11}x_{1ij} + \beta_{21}x_{2ij} + ... + \beta_{p1}x_{pij}; \\ &\quad i = 1, 2, ..., n, j = 1, 2, ..., n, t = 1, 2, ..., T_1, \\ g(\lambda_{ijt2}) &= \beta_{02} + \beta_{12}x_{1ij} + \beta_{22}x_{2ij} + ... + \beta_{p2}x_{pij}; \\ &\quad i = 1, 2, ..., n, j = 1, 2, ..., n, t = T_1 + 1, T_1 + 2, ..., T. \end{aligned} \quad (19)$$

The equality of parameters λ_{ijt1} and λ_{ijt2} is tested by the following hypothesis test:

$$\begin{cases} H_0{:}\lambda_{ijt1} = \lambda_{ijt2} = \lambda_{ijt} \\ \mathrm{H}_1{:}\text{otherwise} \end{cases}. \quad (20)$$

The likelihood function is written as follows:

$$\begin{aligned} L(\mathbf{y};\boldsymbol{\lambda}) &= \prod_{i=1}^{n-1}\prod_{j=i+1}^{n} f\left(y_{ijt}\right) = \prod_{i=1}^{n-1}\prod_{j=i+1}^{n} e^{-\lambda_{ijt}}\frac{\left(\lambda_{ijt}\right)^{y_{ij}}}{y_{ijt}!} \\ &= \frac{\prod_{i=1}^{n-1}\prod_{j=i+1}^{n}\left(\lambda_{ijt}\right)^{y_{ijt}}}{\prod_{i=1}^{n-1}\prod_{j=i+1}^{n} y_{ijt}!} e^{-\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\lambda_{ijt}}. \end{aligned} \quad (21)$$

TABLE 3 People's attributes

	Attribute		
Node	**1**	**2**	**3**
V_1	50	25	10
V_2	42	14	6
V_3	38	10	6
V_4	34	8	4
V_5	30	3	2
V_6	28	2	4

TABLE 4 Indices of people's attributes

Edge Attribute	Attribute			
	Dummy Variable	First Attribute	Second Attribute	Third Attribute
x_{12}	1	0.84	0.56	0.60
x_{13}	1	0.76	0.40	0.60
x_{14}	1	0.68	0.32	0.40
x_{15}	1	0.60	0.12	0.20
x_{16}	1	0.56	0.08	0.40
x_{23}	1	0.90	0.71	1.00
x_{24}	1	0.81	0.57	0.67
x_{25}	1	0.71	0.21	0.33
x_{26}	1	0.67	0.14	0.67
x_{34}	1	0.89	0.80	0.67
x_{35}	1	0.79	0.30	0.33
x_{36}	1	0.74	0.20	0.67
x_{45}	1	0.88	0.38	0.50
x_{46}	1	0.82	0.25	1.00
x_{56}	1	0.93	0.67	0.50

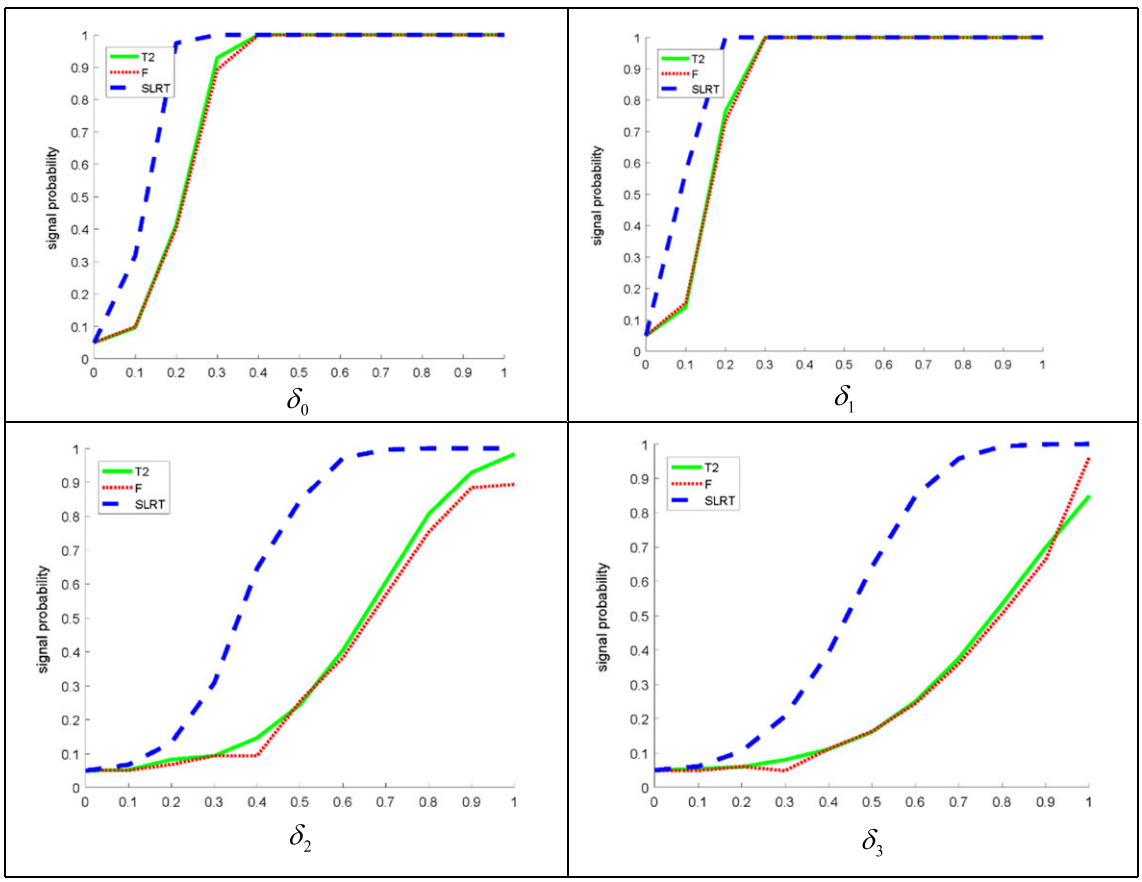

FIGURE 2 Performance of the proposed methods under step shifts in the vector $\boldsymbol{\beta}_t$when $\tau = 5$ [Colour figure can be viewed at wileyonlinelibrary.com]

l_0, l_1, and l_2 are used to construct the LRT statistic. Under H_0, the logarithm of the likelihood function for all observations is denoted by l_0 as follows:

$$l_0 = \sum_{t=1}^{l}\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} y_{ijt}\log\left(\hat{\lambda}_{ijt}\right) - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\hat{\lambda}_{ijt} - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\log\left(y_{ijt}!\right)\right]. \tag{22}$$

The logarithm of the likelihood function for observations before and after the change point is denoted byl_1, l_2 and given according to Equations 23 and 24, respectively.

$$l_1 = \sum_{t=1}^{T_1}\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} y_{ijt}\log\left(\hat{\lambda}_{ijt}\right) - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\hat{\lambda}_{ijt} - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\log\left(y_{ijt}!\right)\right], \tag{23}$$

$$l_2 = \sum_{t=T_1+1}^{T}\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} y_{ijt}\log\left(\hat{\lambda}_{ijt}\right) - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\hat{\lambda}_{ijt} - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\log\left(y_{ijt}!\right)\right]. \tag{24}$$

Under H_1, the likelihood function obtained by the sum of l_1 and l_2 is called l_a. Next, the generalized likelihood ratio statistic is calculated as follows:

$$lrt_{T_1} = -2(l_0 - l_a), T_1 = 1, ..., T-1. \tag{25}$$

Finally, the LRT statistic is standardized as follows:

$$SLRT_{T_1} = \frac{LRT_{T_1} - E(LRT_{T_1})}{\sqrt{Var(LRT_{T_1})}}, T_1 = 1, ..., T-1, \tag{26}$$

where$E(LRT_{T_1})$ and $Var(LRT_{T_1})$ are the expected value and variance of the LRT statistic, respectively, which are obtained by simulation runs. If the statistic max $\{\mathrm{SLRT}_{T_1}\}, T_1 = 1, ..., T-1$ exceeds an upper control limit (UCL), the social network is out of control. The UCL is obtained by simulation to achieve a specified probability of Type I errors.

5 | SIMULATION STUDIES

In this section, the performance of the extended methods in monitoring an attributed social network is evaluated in terms of signal probability criterion and compared under different kinds of changes (step shifts,

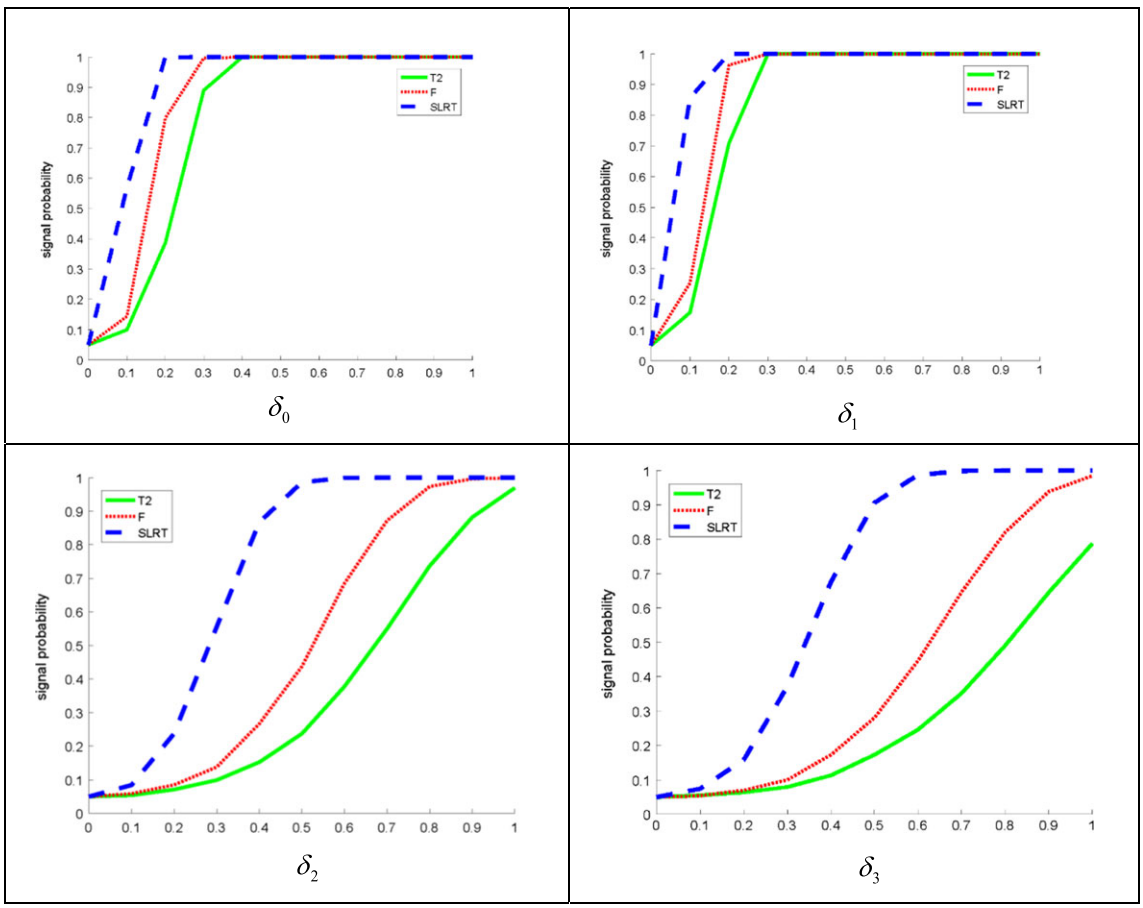

FIGURE 3 Performance of the proposed methods under step shifts in the vector $\boldsymbol{\beta}_t$when $\tau = 10$ [Colour figure can be viewed at wileyonlinelibrary.com]

drift, and outliers). Consider a social network with 6 nodes ($n = 6$), each with 3 attributes ($p = 3$). The values of the attributes for each node are given in Table 3.

Based on Table 3, the matrix of explanatory variables is determined by Equation 4 and summarized in Table 4.

Note that to adapt the proposed model to the literature of profile monitoring, the elements of the first column of matrix $\mathbf{X}$ are considered equal to one. Then, the first element of vector $\boldsymbol{\beta}_t$ is called a model intercept and denoted by β_{0t}.

It is assumed that when the process is in control, $\boldsymbol{\beta}_t = (\beta_{0t}, \beta_{1t}, \beta_{2t}, \beta_{3t}) = (1, 1, 1, 1)$. When a step change occurs, the vector of model parameters changes to $\boldsymbol{\beta}_t' = (\beta_{0t}, \beta_{1t}, \beta_{2t}, \beta_{3t}) + (\delta_0\sigma_0, \delta_1\sigma_1, \delta_2\sigma_2, \delta_3\sigma_3) \neq \boldsymbol{\beta}_t$ where $\boldsymbol{\delta} = (\delta_0, \delta_1, \delta_2, \delta_3) \neq \mathbf{0}$. The covariance matrix of model parameters can be computed as Equation 27.[29] $\left(\sigma_{\hat{\beta}_0}, \sigma_{\hat{\beta}_1}, \sigma_{\hat{\beta}_2}, \sigma_{\hat{\beta}_3}\right) = (0.7953, 1.2601, 0.4974, 0.3164)$, and is obtained for the numerical example.

$$\begin{bmatrix} \sigma^2_{\hat{\beta}_0} & \rho\sigma_{\hat{\beta}_0}\sigma_{\hat{\beta}_1} & \rho\sigma_{\hat{\beta}_0}\sigma_{\hat{\beta}_2} & \rho\sigma_{\hat{\beta}_0}\sigma_{\hat{\beta}_3} \\ \rho\sigma_{\hat{\beta}_1}\sigma_{\hat{\beta}_0} & \sigma^2_{\hat{\beta}_1} & \rho\sigma_{\hat{\beta}_1}\sigma_{\hat{\beta}_2} & \rho\sigma_{\hat{\beta}_1}\sigma_{\hat{\beta}_3} \\ \rho\sigma_{\hat{\beta}_2}\sigma_{\hat{\beta}_0} & \rho\sigma_{\hat{\beta}_2}\sigma_{\hat{\beta}_1} & \sigma^2_{\hat{\beta}_2} & \rho\sigma_{\hat{\beta}_2}\sigma_{\hat{\beta}_3} \\ \rho\sigma_{\hat{\beta}_3}\sigma_{\hat{\beta}_0} & \rho\sigma_{\hat{\beta}_3}\sigma_{\hat{\beta}_1} & \rho\sigma_{\hat{\beta}_3}\sigma_{\hat{\beta}_2} & \sigma^2_{\hat{\beta}_3} \end{bmatrix} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X}\right)^{-1}. \tag{27}$$

Concerning step changes, 3 kinds of out-of-control situations in terms of the location of shifts are considered. These states are as follow: (1) shifts in the second half of the data, ie, from the 16th to 30th profiles ($\tau = 15$), (2) shifts from the 11th to 30th profiles($\tau = 10$), and (3) shifts from the 6th to 30th profiles ($\tau = 5$). Note that τ is the location of shifts induced in the model parameters. To compare the performance of the proposed methods through simulation experiments, the UCL of each chart is set such that the probability of Type I errors is approximately equal to $\alpha = 0.05$. The UCL values of the extended T^2, F, and SLRT charts are equal to 18.00, 1.2752, and 4.12,

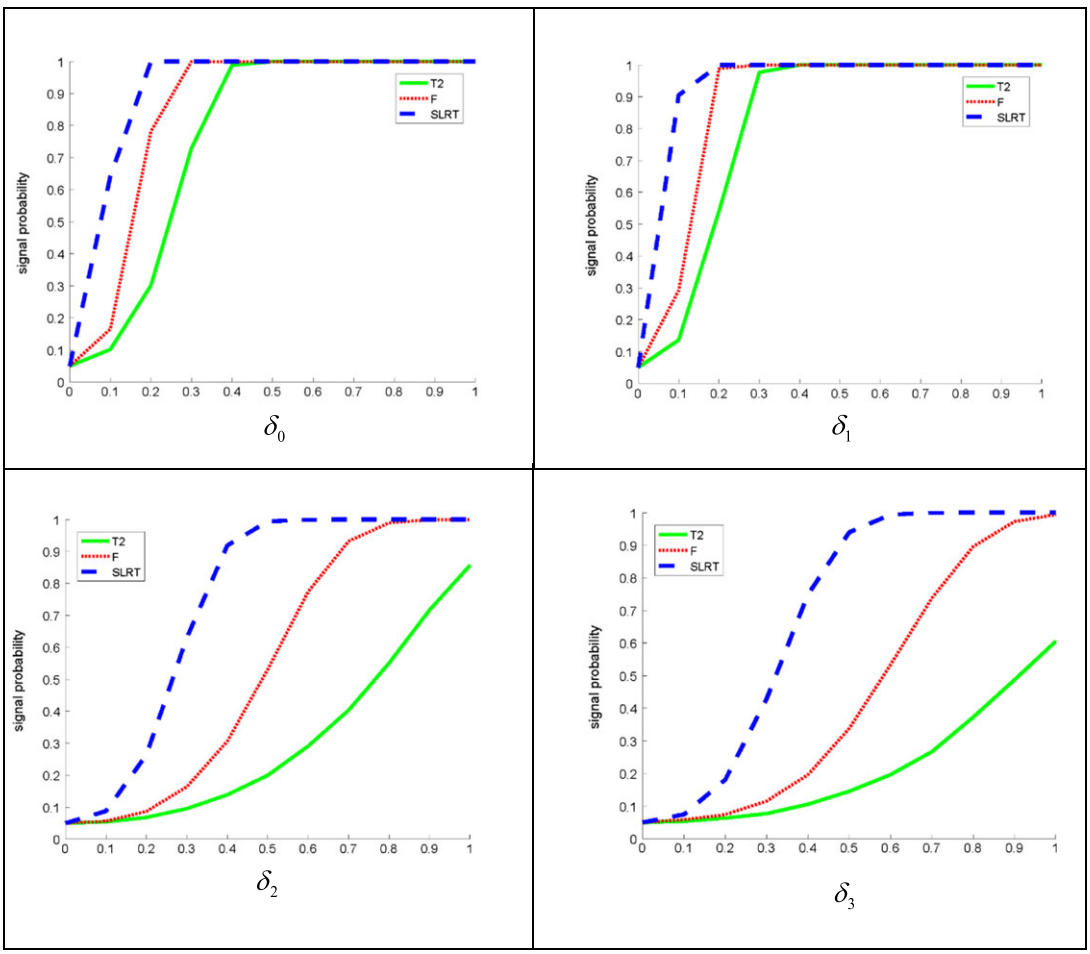

FIGURE 4 Performance of the proposed methods under step shifts in the vector $\boldsymbol{\beta}_t$when $\tau = 15$ [Colour figure can be viewed at wileyonlinelibrary.com]

respectively. It is worth mentioning that the simulations were performed with MATLAB software for 10 000 replications. The values of signal probability under different step changes in the model parameters when $\tau = 5$, 10, 15 are given in Figures 2-4.

As the magnitude of the shift in vector $\boldsymbol{\beta}_t$ increases, the capability of all proposed control charts to detect the step changes improves. The extended SLRT method outperforms the competing methods under all step changes in vector $\boldsymbol{\beta}_t$ and different values of parameter τ, especially when the shift occurs in the third and fourth model parameters (β_{2t} and β_{3t}). In the case of $\tau = 5$, under most step shifts, the signal probability values for T^2 are larger than those obtained by the F method. However, the F method outperforms T^2 when $\tau = 10$ and 15. Generally, it can be concluded that the performance of the extended F chart improves as the value of parameter τ increases, while τhas an inverse effect on the performance of T^2. In the other words, as parameter τincreases, the signal probability values for T^2 decrease. The reason for these trends is that the F method assesses whether the networks are similar to each other. For example, $\tau = 5$ implies that 25 out of 30 networks are similar. Hence, as the value of τ increases to 10 and 15, the similarities between networks, and consequently the power of the extended F method to detect out-of-control situations, decrease. In other words, when most of the networks are similar, the performance of the F method in detecting step shifts deteriorates.

Generating out-of-control profiles under drift is done by inducing a linear trend beginning from τ as:

$$\boldsymbol{\beta}_t^{'} = \begin{bmatrix} \beta_{0t} + ((i-1)/(m-1))\times\zeta_0\sigma_0, \beta_{1t} + ((i-1)/(m-1))\times\zeta_1\sigma_1, \beta_{2t} + \\ ((i-1)/(m-1))\times\zeta_2\sigma_2, \beta_{3t} + ((i-1)/(m-1))\times\zeta_3\sigma_3 \end{bmatrix}; t = \tau+1, \ldots, m, \tag{28}$$

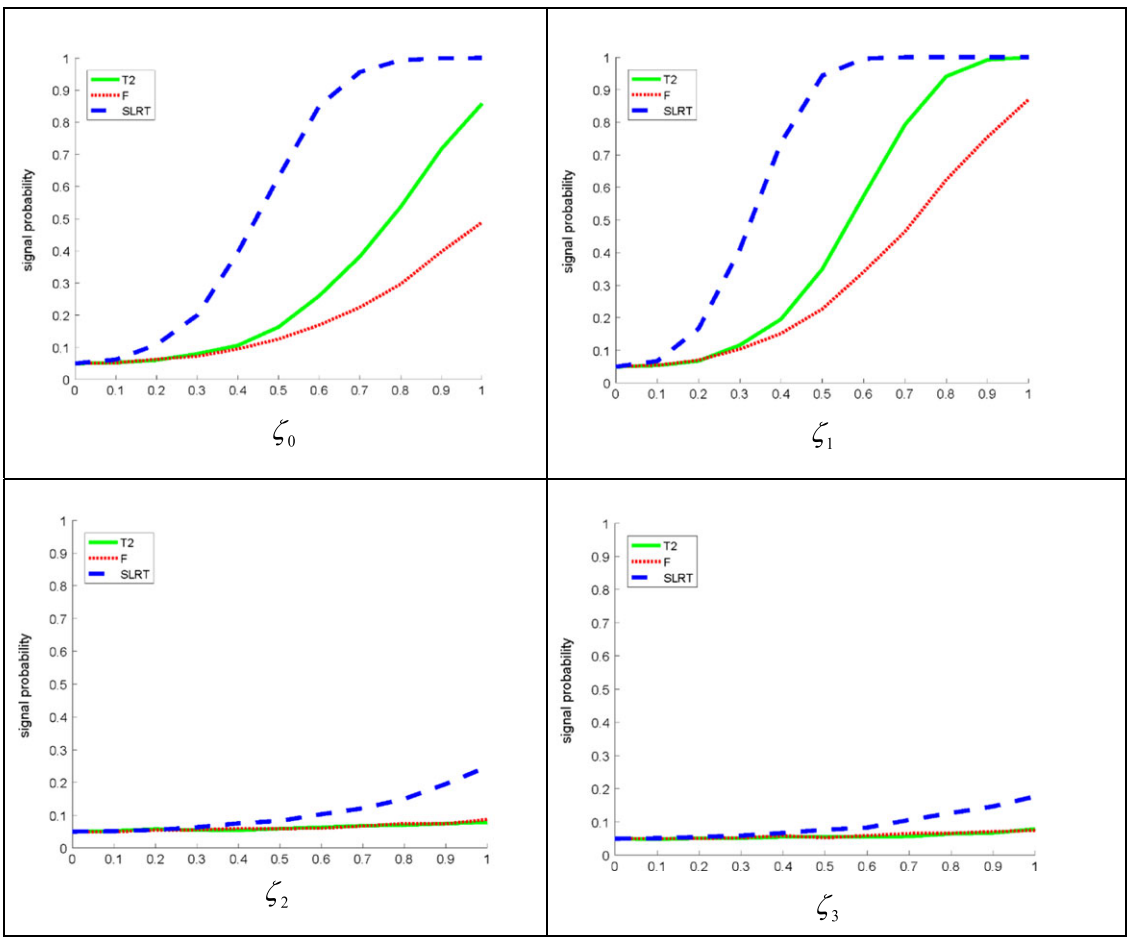

FIGURE 5 Performance of the proposed methods under drift shifts in the vector $\boldsymbol{\beta}_t$when $\tau = 5$ [Colour figure can be viewed at wileyonlinelibrary.com]

where $(\zeta_0, \zeta_1, \zeta_2, \zeta_3) \neq \mathbf{0}$. The signal probability values under a linear trend in vector $\boldsymbol{\beta}_t$ are summarized in Figures 5-7. When $\tau = 5$ and $\tau = 10$, the performance of the extended SLRT method in detecting out-of-control states with a linear trend in vector $\boldsymbol{\beta}_t$ is better than the other methods. In the case of $\tau = 5$ under linear trends in β_{0t} and β_{1t}, the T^2 chart outperforms F, while in the case of $\tau = 10$ and under the same situations, the performance of F is better than T^2. In out-of-control situations with linear trends in β_{2t} and β_{3t} under $\tau = 5$ and $\tau = 10$, all 3 methods cannot adequately detect out-of-control states. However, the SLRT method outperforms the other methods. The reason is that the standard deviations of $\hat{\beta}_{2t}$ and $\hat{\beta}_{3t}$ (0.4974 and 0.3164) are smaller than the standard deviations of $\hat{\beta}_{0t}$ and $\hat{\beta}_{1t}$ (0.7953 and 1.2601). The performance of the T^2 and F methods in detecting anomalies in β_{2t} and β_{3t} under change points $\tau = 5$ and $\tau = 10$ are almost the same. Under $\tau = 15$, none of the 3 methods can detect any changes in the vector of model parameters. This issue can be justified by noting that as the location of the induced linear trend increases, the probability of the chart statistic falling outside the control limit interval decreases considerably.

Another scenario involves k outliers among T profiles. In Figures 8-10, 1, 2, and 3 outliers are considered. Outliers are generated as follows:

$$\boldsymbol{\beta}_k' = (\beta_0, \beta_1, \beta_2, \beta_3) + (\theta_0\sigma_0, \theta_1\sigma_1, \theta_2\sigma_2, \theta_3\sigma_3) \neq \boldsymbol{\beta}; k = 10, 15, 20 \tag{29}$$

where $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3) \neq \mathbf{0}$. In order to apply outliers in the sampled profiles, 3 scenarios are considered: (1) shifts in the 5th, 10th, and 15th profiles; (2) shifts in the 10th and 15th profiles; and (3) shifts in the 15th profile. In all scenarios, the SLRT method shows the worst performance in detecting outliers. That is because the design of the SLRT method is based on the step shift; hence, the

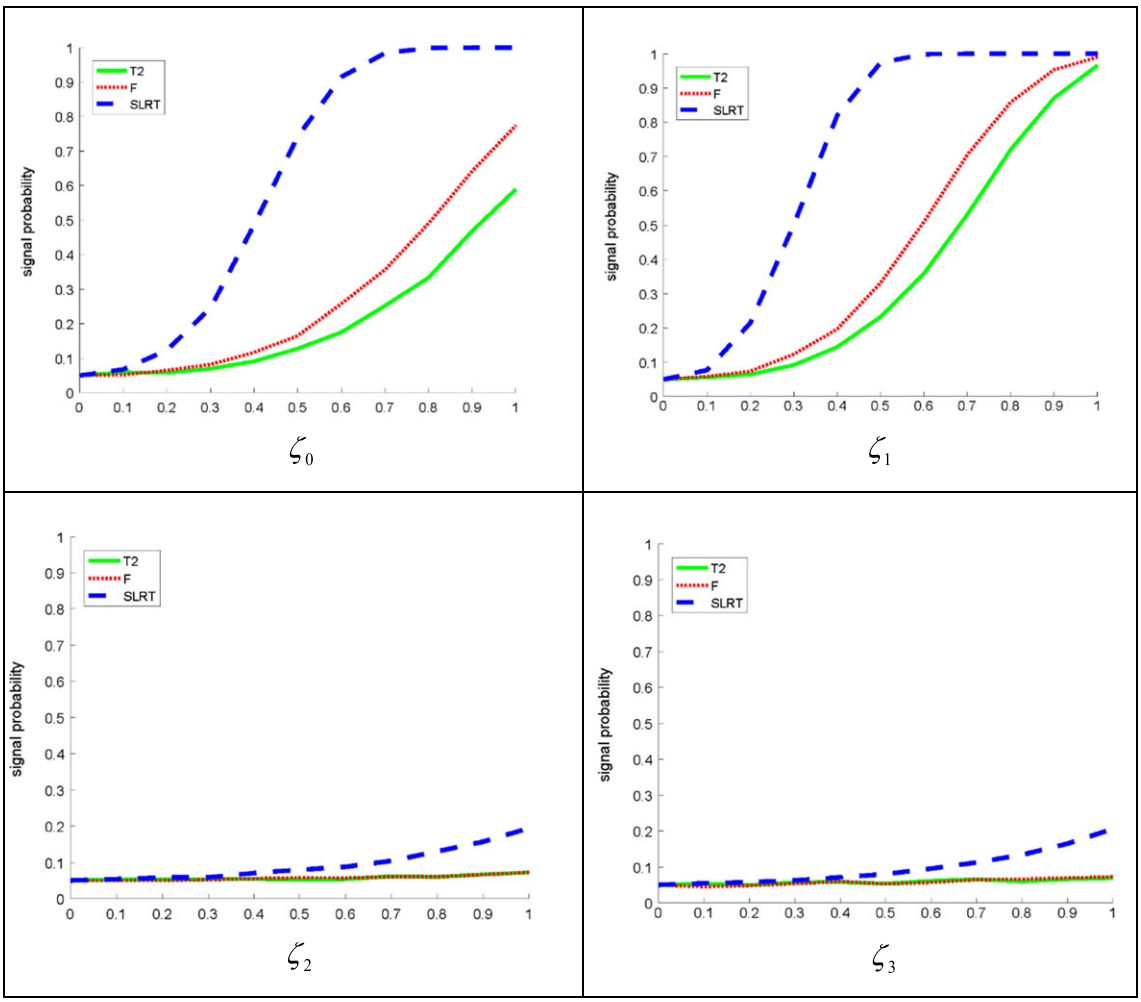

FIGURE 6 Performance of the proposed methods under drift shifts in the vector $\boldsymbol{\beta}_t$when $\tau = 10$ [Colour figure can be viewed at wileyonlinelibrary.com]

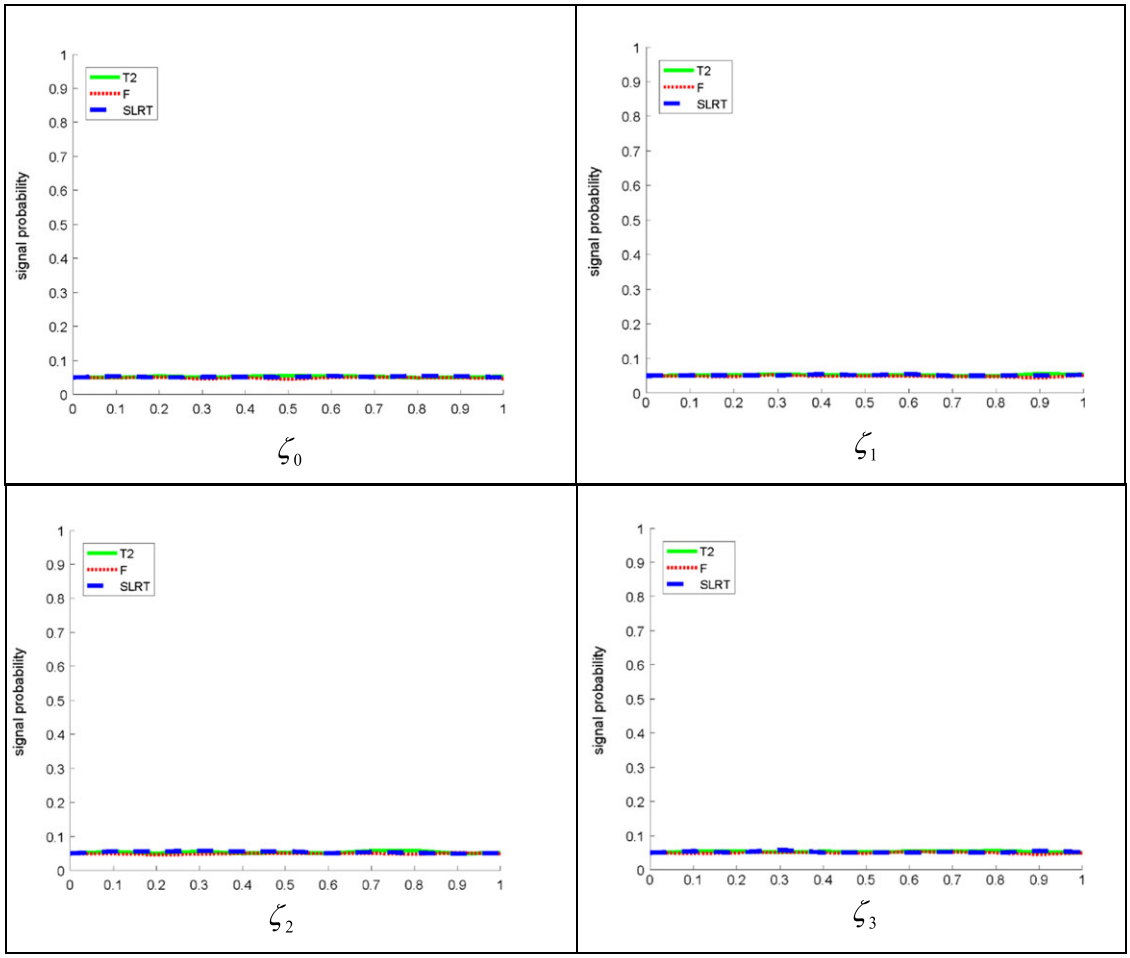

FIGURE 7 Performance of the proposed methods under drift shifts in the vector $\boldsymbol{\beta}_t$when $\tau = 15$ [Colour figure can be viewed at wileyonlinelibrary.com]

performance of this method under outliers is not satisfactory. In the first and second scenarios, the F statistic shows better performance than T^2, but in the third scenario T^2 outperforms the F method. As mentioned before, the larger the number of outliers, the better the performance of the F method.

6 | AN ILLUSTRATIVE EXAMPLE

Using the same data from Section 5, a numerical example is given in this section to illustrate the application of the extended methods for monitoring the counts of communications in social networks. The simulation includes 20 in-control profiles; then, the step change of $\boldsymbol{\delta} = (\delta_0 = 0.1, \delta_1 = 0.1, \delta_2 = 0.2, \delta_3 = 0.2)$ is induced to generate out-of-control data for the 21st to 30th profiles. The T^2 statistics for the generated samples are plotted in Figure 11 which shows that $\max_{t=1,\ldots,30}\{T_t^2\} = 20.9996$. This implies that the T^2 chart generally detects the corresponding out-of-control situation induced in the counts of communications. For the T^2 chart, a false alarm is received in the 7th sample. The T^2 chart triggers an out-of-control situation in the 23rd and 28th samples.

The values of F and $\{\max(SLRT_{T_1}), T_1 = 1, \ldots, T_1\}$ calculated for the 30 samples generated are equal to 2.7515 and 66.4821, respectively. These values show that both methods detect out-of-control situations.

7 | CASE STUDY: ENRON EMAIL DATA

As Woodall et al[30] stated the application of social network monitoring methods is often illustrated using 2 types of networks such as terroristic ones like the "al Qaeda" network or that based on Enron e-mail communications. Here, we select the real-world Enron email data set[31] to show the application of the proposed methods. Enron email communication network covers all the email communication within a data set of around half million

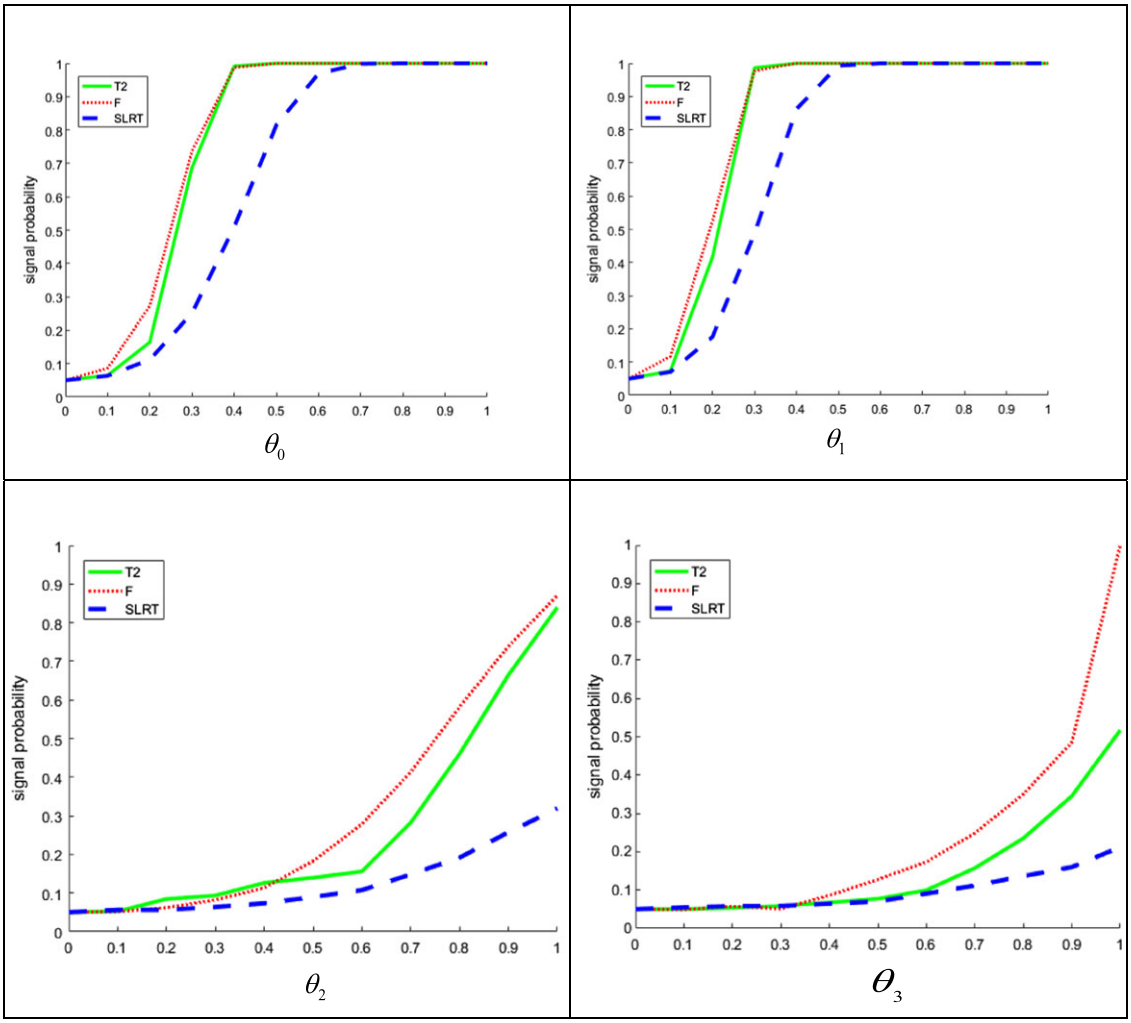

FIGURE 8 Performance of the proposed methods under 3 outlier profiles (5th, 10th, and 15th profiles) [Colour figure can be viewed at wileyonlinelibrary.com]

emails. These data were originally made public and posted to the web by the Federal Energy Regulatory Commission during its investigation. The nodes of the network are email addresses in which if the address i sent at least 1 email to address j, the graph contains an undirected edge from i to j. The Enron email data set has 150 folders containing email information such as inbox, sent mail, subjects, and mail contents from 1998 to 2002. The raw data set is available online in (http://www.cs.cmu.edu/~enron/). The SQL version of this data set is also presented by UC Berkeley and is available in http://bailando.sims.berkeley.edu/enron_email.html, which makes it easier to explore the communications between the persons. Each person is considered as a node, and monthly email communication is aggregated to form network edges at monthly time stamps. All people who send or receive emails exceed 250 persons which some of them have sent email just once. The email communication data set is shown in Figure 12. As seen, many nodes are contacted for 1 or 2 times and do not include majority of communications. Hence, for simplicity of computations and reducing network size, the nodes exchanging more than 10 emails with each other are selected for analysis. After applying this filter, 20 people (nodes) are remained, and then the monthly number of communications in 2001 is monitored in this section. These email communications are imported into Gephi software, which is a popular network analysis application. Two node attributes, gender and work experience, are also considered in this case study. The selected data network for 2001 is shown in Figure 13. Because, the sum of some Poisson random variables also follows a Poisson distribution, to check the suitability of Poisson regression model, the sum of the communications between the nodes for each time period is considered. Then, by using goodness of fitness test for Poisson distribution in Minitab

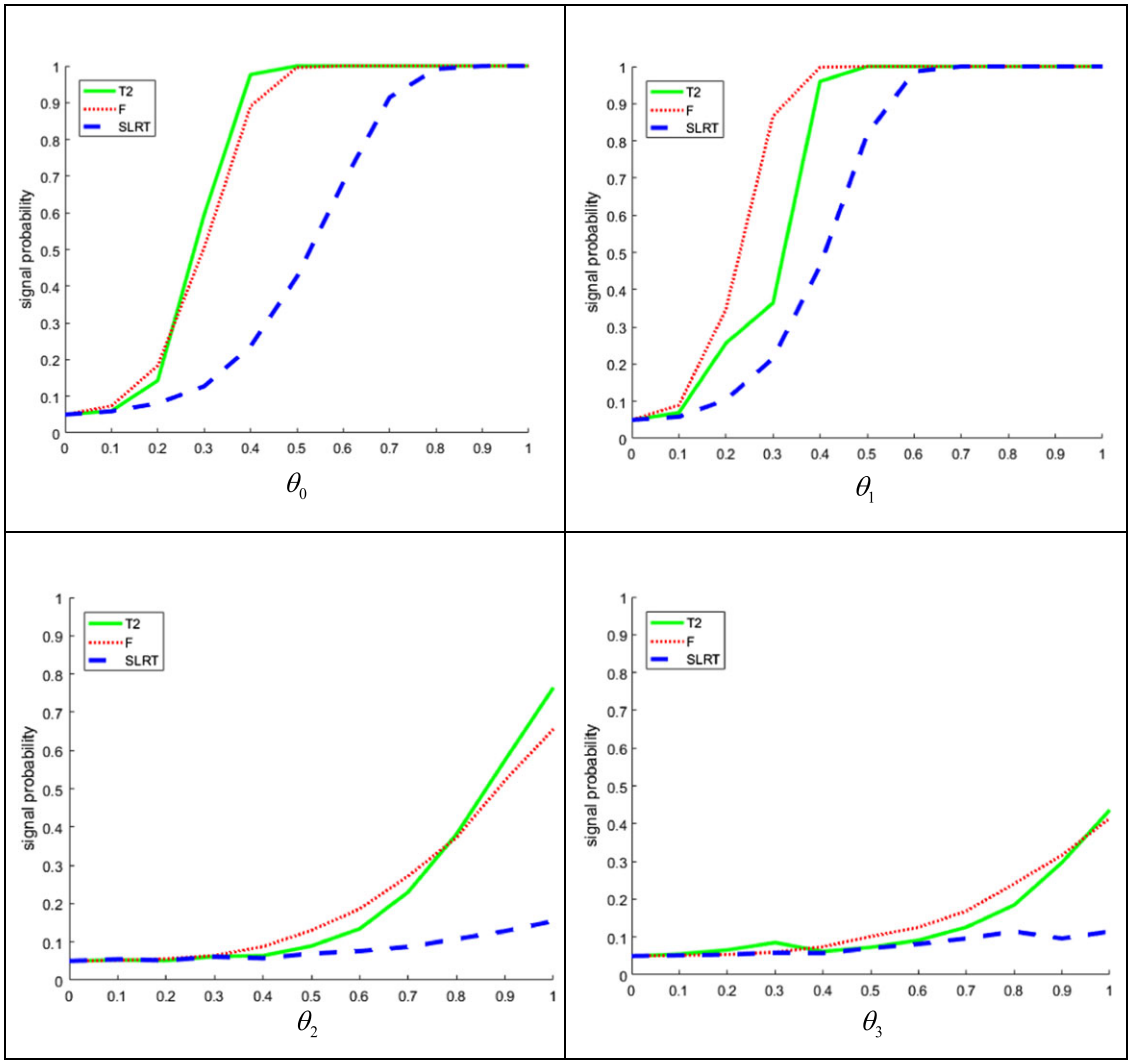

FIGURE 9 Performance of the proposed methods under 2 outlier profiles (10th and 15th profiles) [Colour figure can be viewed at wileyonlinelibrary.com]

software, the following hypothesis is checked; H_0: the response values follow Poisson distribution, H_1: otherwise. Then, the P-value is computed equal to 0.054. As the consequence, at significance level of $\alpha = 0.05$, H_0 is not rejected. The result of this test shows that the response variable follows a Poisson distribution. Hence, it is reasonable to use Poisson regression profile to model the relationship between the response variable and the explanatory variables.

The T^2 statistics for the selected Enron data set are plotted in Figure 14. T^2 control chart for indicating $\max_{t=1,\ldots,12}\{T_t^2\} = 32.03$. This implies that the T^2 chart detects the out-of-control situations induced in the average counts of communications. The T^2 chart triggers an out-of-control signal in the first time period, which is related to January. This is because of the annual holidays, which leads to decreasing the number of work emails.

The values calculated for F and $\{\max(SLRT_{T_1}), T_1 = 1, \ldots, T_1\}$ for the Enron data are equal to 29.37 and 14.43, respectively. Hence, all the proposed methods can detect out-of-control situations. After removing the out-of-control point, the remained data set will be in-control and can be used for estimating the process parameters.

8 | CONCLUSIONS AND A RECOMMENDATION FOR FUTURE RESEARCH

In this paper, first the relationship between the counts of communications and the attributes of network nodes was modeled by Poisson regression profiles. In previous studies, only the existence of communications between people has been modeled by logistic regression profiles.

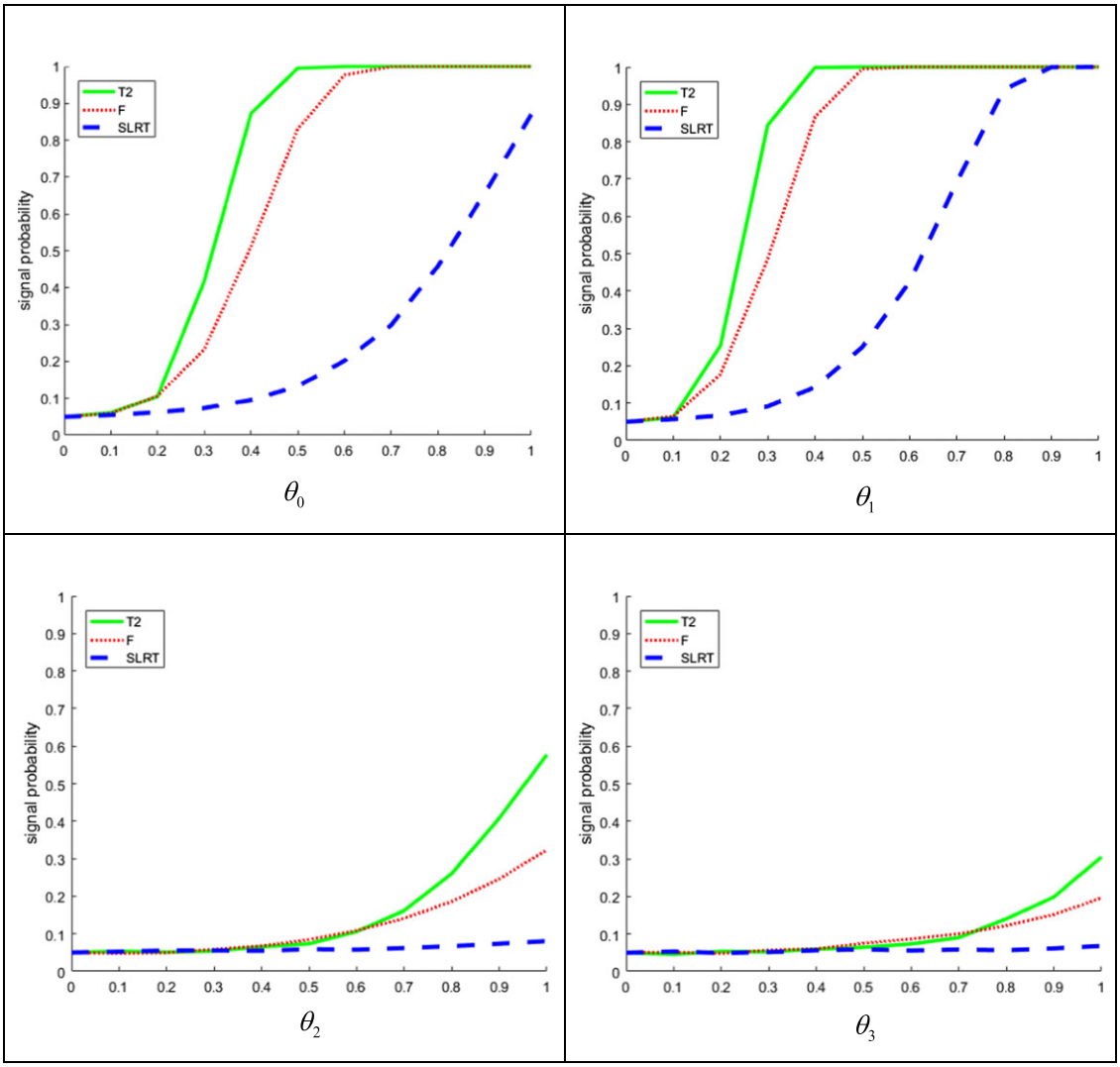

FIGURE 10 Performance of the proposed methods under 1 outlier profile (15th profiles) [Colour figure can be viewed at wileyonlinelibrary.com]

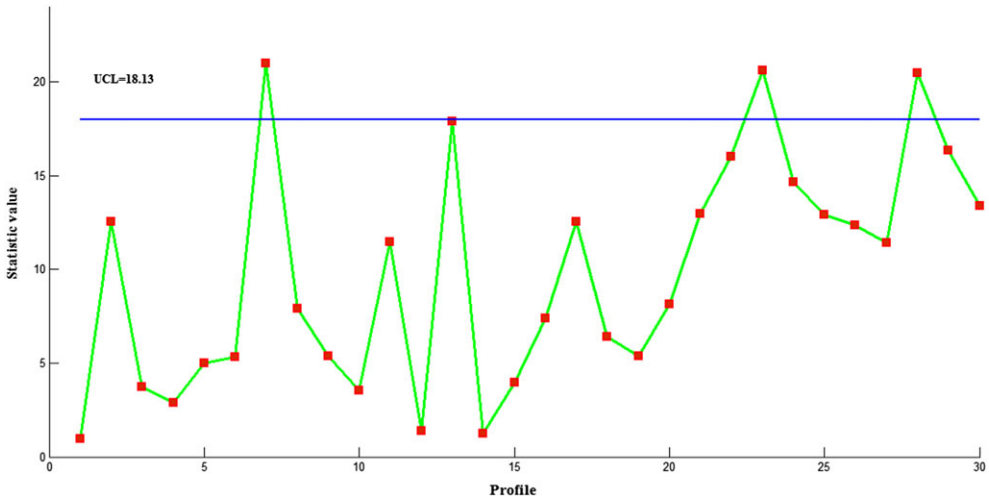

FIGURE 11 Sample statistics for T^2 chart [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 12 Enron email network in 2001

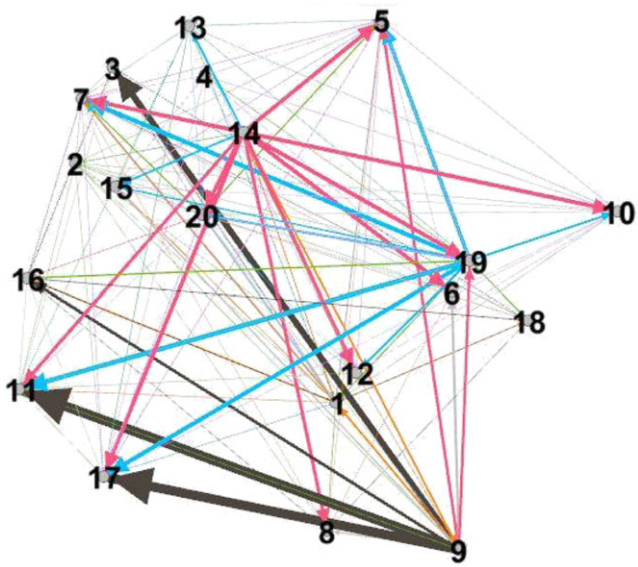

FIGURE 13 Enron email network case study considering the nodes with more than 10 emails in 2001 [Colour figure can be viewed at wileyonlinelibrary.com]

This paper discusses the superiority of modeling the number of communications rather than just the existence of communications among people. To monitor a social network modeled by Poisson regression 3 methods, T^2, F, and SLRT, were developed. Simulation studies were conducted for a simulated social network, and the performance of the extended methods in detecting out-of-control scenarios under different step

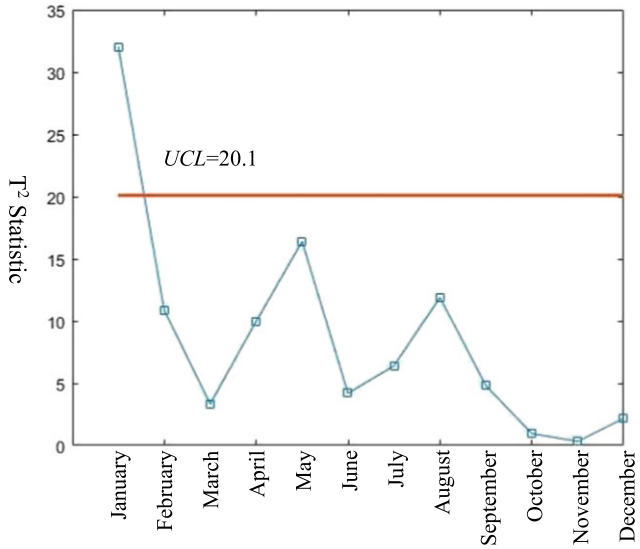

FIGURE 14 T^2 control chart for Enron email data [Colour figure can be viewed at wileyonlinelibrary.com]

changes, linear trends, and outliers were compared. The results of simulation studies in terms of signal probability criterion show that in most out-of-control situations, the SLRT method outperforms the T^2 and F methods except for detecting outlier scenarios. Note that as the number of outliers increases, the performance of the proposed F method improves. However, when the shift occurs at the initial or end time periods, the performance of the F method deteriorates. A numerical example was given to illustrate application of the proposed methods for monitoring the average counts of communications in social networks. Then, the proposed methods were applied to the real-world Enron email data set. In some cases, the counts of communications between people over time are autocorrelated. Hence, monitoring autocorrelated social networks in Phase I is a fruitful area for future research.

ORCID

Amirhossein Amiri http://orcid.org/0000-0002-2385-8910

REFERENCES

1. Erdos P, Renyi A. On random graphs. *Publ Math Debr*. 1959;6:290-297.
2. Barabasi A, Albert R. Emergence of scaling in random networks. *Science*. 1999;286(5439):509-512.
3. Leskovec J, Kleinberg J, Faloutsos C. Graph evolution: densication and shrinking diameters. *ACM Trans Knowl Discov Data*. 2007;1(1):1-41.
4. Chakrabarti D, Zhan Y, Faloutsos C. R-MAT: a recursive model for graph mining. Proceedings of SIAM International Conference on Data Mining 2004.

5. Pennock DM, Flake GW, Lawrence S, Glover EJ, Giles CL. Winners don't take all: characterizing the competition for links on the web. *Proc Natl Acad Sci*. 2002;5207-5211.
6. Frank O, Strauss D. Markov graphs. *J Am Stat Assoc*. 1986;81(395):832-842.
7. Hoff PD, Raftery AE, Handcock MS. Latent space approaches to social network analysis. *J Am Stat Assoc*. 2002; 97(460):1090-1098.
8. Kim M, Leskovec J. Multiplicative attribute graph model of real-world networks. Proceedings of International Workshop on Algorithms and Models for the Web-Graph 2010: 62–73.
9. Hoff PD. Random effects models for network data. Proceedings of Dynamic Social Network Modeling and Analysis 2003: 303–312.
10. Snijders TA. Models for longitudinal network data. In: *Models and Methods in Social Network Analysis*. Vol.1; 2005: 215-247.
11. Hanneke S, Fu W, Xing EP. Discrete temporal models of social networks. *Electron J Stat*. 2010;4:585-605.
12. Sarkar P, Moore AW. Dynamic social network analysis using latent space models. *SIGKDD Explor*. 2005;7(2):31-40.
13. Ho Q, Song L, Xing EP. Evolving cluster mixed-membership blockmodel for time-evolving networks. In: Proceedings of International Conference on Articial Intelligence and *Statistics*; 2011:342-350.
14. Xu KS, Hero A. Dynamic stochastic blockmodels: statistical models for time evolving networks. Proceedings of Social Computing, Behavioral-Cultural Modeling and Prediction 2013; 201–210.
15. McCulloh I, Carley KM. Detecting change in longitudinal social networks. *J Soc Struct*. 2011;12(3):1-37.
16. Park Y, Priebe C, Youssef A. Anomaly detection in time series of graphs using fusion of graph invariants. *IEEE J Sel Top Sign Proces*. 2012;7(1):67-75.
17. Priebe CE, Conroy JM, Marchette DJ, Park Y. Scan statistics on Enron graphs. *Comput Math Organ Theory*. 2005;11(3): 229-247.
18. Marchette D. Scan statistics on graphs. Wiley Interdisciplinary Reviews: Computational Statistics. 2012;4(5):466-473.
19. Neil J, Storlie C, Hash C, Brugh A, Fisk M. Scan statistics For the online detection of locally anomalous subgraphs. *Technometrics*. 2014;55(4):403-414.
20. Sparks R. Detecting periods of significant increased communication levels for subgroups of targeted individuals. *Qual Reliab Eng Int*. 2016;32(5):1871-1888.
21. Kang L, Albin SL. Online monitoring when the lrl process yields a linear profile. *J Qual Technol*. 2000;32(4): 418-426.
22. Mahmoud MA, Woodall WH. Phase I analysis of linear profiles with calibration applications. *Technometrics*. 2004;46(4): 380-391.
23. Amiri A, Jensen WA, Kazemzadeh RB. A case study on monitoring polynomial profiles in the automotive industry. *Qual Reliab Eng Int*. 2010;26(5):509-520.
24. Kim K, Mahmoud MA, Woodall WH. On the monitoring of linear profiles. *J Qual Technol*. 2003;35(3):317-328.
25. Mahmoud MA, Parker PA, Woodall WH, Hawkins DM. A change point method for linear profile data. *Qual Reliab Eng Int*. 2006;23(2):247-268.
26. Azarnoush B, Paynabar K, Bekki J, Runger G. Monitoring temporal homogeneity in attributed network streams. *J Qual Technol*. 2016;48(1):28-43.
27. Sharma S. *Applied Multivariate Techniques*. Wiley; 1996:7-10.
28. Amiri A, Koosha M, Azhdari A, Wang G. Phase I monitoring of generalized linear model-based regression profiles. *J Stat Comput Simul*. 2015;85(14):2839-2859.
29. Yeh AB, Huwang L, Li Y. Profile monitoring for a binary response. *IIE Trans*. 2009;41(11):931-941.
30. Woodall WH, Zhao MJ, Paynabar K, Sparks R, Wilson JD. An overview and perspective on social network monitoring. *IISE Trans*. 2017;49(3):354-365.
31. http://bailando.sims.berkeley.edu/enron/enron.sql.gz

Hatef Fotuhi is a PhD candidate in Industrial Engineering at Shahed University in Iran. His research interest is statistical process monitoring.

Amirhossein Amiri is an associate professor at Shahed University in Iran. He holds a BS, MS, and PhD in Industrial Engineering from Khajeh Nasir University of Technology, Iran University of Science and Technology, and Tarbiat Modares University in Iran, respectively. He is now vice chancellor of education in Faculty of Engineering at Shahed University in Iran and a member of the Iranian Statistical Association. His research interests are statistical process monitoring, profile monitoring, and change point estimation. He has published many papers in the area of statistical process control in high-quality international journals such as *Quality and Reliability Engineering International, Communications in Statistics, Computers and Industrial Engineering*, and so on. He has also published a book with John Wiley and Sons in 2011 titled *Statistical Analysis of Profile Monitoring*.

Mohammad Reza Maleki is a PhD graduate in Industrial Engineering at Shahed University in Iran. He has obtained his BS and MS degrees in Industrial Engineering from Isfahan University of Technology and Shahed University, respectively. His research interests include statistical process monitoring, profile monitoring, change point estimation, and measurement errors effect on SPM. He has been the author or coauthor of various papers published in high-ranked journals such as *Computers &*

Industrial Engineering, Quality and Reliability Engineering International, Communications in Statistics–Simulation and Computation, Communications in Statistics-Theory and Methods, Transactions of the Institute of Measurement and Control, and *Scientia Iranica*.

How to cite this article: Fotuhi H, Amiri A, Maleki MR. Phase I monitoring of social networks based on Poisson regression profiles. *Qual Reliab Engng Int*. 2018;34:572–588. https://doi.org/10.1002/qre.2273