Scientia Iranica E (2022) 29(3), 1581–1591

SCIENTIA IRANICA

Sharif University of Technology
Scientia Iranica
Transactions E: Industrial Engineering
http://scientiairanica.sharif.edu

Monitoring attributed social networks based on count data and random effects

H. Mogouie[a], Gh.A. Raissi Ardali[a], A. Amiri[b,*], and E. Bahrami Samani[c]

a. *Department of Industrial and Systems Engineering, Isfahan University of Technology, Isfahan, Iran.*
b. *Department of Industrial Engineering, Faculty of Engineering, Shahed University, Tehran, Iran.*
c. *Department of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran.*

Received 2 May 2018; received in revised form 21 April 2020; accepted 13 July 2020

KEYWORDS
Count data;
Random effects;
Social networks;
Statistical monitoring.

Abstract. This paper presents a novel approach to statistical monitoring of online social networks where the edges represent the number of communications between ties at each time stamp. Since the available methods in the literature are limited to the assumption that the set of all interacting individuals is fixed during the monitoring horizon and their corresponding attributes do not change over time, the proposed method of this study tackles these limitations due to the properties of the random effect concepts. The application of appropriate parameter estimation technique is involved in the Likelihood Ratio Testing (LRT) approach considering two different statistics and the longitudinal network data are monitored. The performance of the proposed method is verified using numerical examples including simulation studies as well as an illustrative example.

1. Introduction

Analysis of network data remains of interest to researchers since networks demonstrate a comprehensive image of complex systems. Preliminary studies, such as those by Erdös and Rényi [1], incorporate probabilistic concepts into graph theory to analyze network data. Although this approach was extended by other researchers such as Leskovec et al. [2], ignoring the dynamic property of network data necessitates more in-depth studies that are capable to analyze network streams over time.

The development and spread of the digital world and the occurrence of notable events, such as 2012 Arab uprisings and the role of online social networks in those movements, have made the dynamic network data analysis a hot topic for researchers [3]. The most crucial objectives of these investigations are to detect abrupt changes of interactions among members known as anomalies as soon as possible [4].

The surveillance of online networks has led to the introduction of methods, proposed by Krebs [5], for tracing central actors in terrorist groups. Similarly, Pandit et al. [6] argued that detection of anomalies, such as systematic cheating, in online networks was crucial to monetary systems and in this respect, insurance industries require further research into this issue.

A literature review of anomaly detection proposed by Savage et al. [7] employs categorization based on two criteria of whether the network under study is labeled or unlabeled and if the surveillance is conducted statically or dynamically. Aside from the studies dealing with static problems, the first criterion concentrates on the nodes of the networks to discriminate cases with known attributes of nodes (individuals) from those without node information such as the studies of Cheng and Dickinson [8] and McCulloh and Carely [9].

*. *Corresponding author.*
E-mail addresses: H.mogouie@in.iut.ac.ir (H. Mogouie); Raissi@cc.iut.ac.ir (Gh.A. Raissi Ardali); amiri@shahed.ac.ir (A. Amiri); E_bahrami@sbu.ac.ir (E. Bahrami Samani)

doi: 10.24200/sci.2020.50951.1933

Although dynamism is evaluated as a crucial property of social network studies, more recent developments necessitate gaining an insight into the underlying structure of communication among network members [10]. To this end, it is always of interest to determining how similarities between pairs of individuals affect the pattern of their communications. For instance, Bliss et al. [11] demonstrated how accurate the link predictions would be upon knowing the user's attributes. Hence, it can be concluded that further research opportunities in the field of social network analysis are likely required to prioritize dynamic labelled problems.

Woodall et al. [12] proposed a review investigation into classifying anomaly detection where notable studies that propose a certain approach were reviewed briefly. For instance, the research conducted by Heard et al. [13] was introduced as an instance for applying Bayesian methods for social network analysis and the scan statistic approach is discussed by referring to the works of Priebe et al. [14], Sparks [15], Neil et al. [16], and Marchette [17]. Another approach popular in the literature is the application of time series analysis for monitoring social networks, as can be seen in the study of Pincombe [18]. The last category of researches noted by Woodall et al. [12] is a control chart design-based hypothesis testing where network measures are used as the statistics for monitoring the application of Exponentially Weighted Moving Average (EWMA) charts. The works of McCulloh and Carley [19,20] are well-known studies that can be categorized in this class.

The extension of researches that used control charts is found in the research paper of Azarnoush et al. [21]. The main drawback to be addressed is that focusing on graph measures does not necessarily lead to the detection of structural anomalies. To overcome such a challenge, Azarnoush et al. [21] modeled the probability of communication of each pair of individuals in terms of similarity vector, which is derived from a predefined method of comparing the corresponding individuals of a certain pair. Then, by using Generalized Linear Modeling (GLM) concepts and an Likelihood Ratio Testing (LRT) approach, LRT-based statistics were computed and monitored. An important advantage of this method is focusing on the underlying structure of the network data that reveals a general insight into how individuals are likely to connect to each other based on similarities between attributes. This approach was extended by Farahani et al. [22] and Fotouhi et al. [23] with focus on Poisson distributed edge data rather than binary outcomes.

In the aforementioned researches, although significant progress results from the application of more applicable models for real-world networks, significant limitations remain to account for. An important notion that has been neglected by Azarnoush et al. [21], Farahani et al. [22], and Fotouhi et al. [23] is that the dynamism of network data necessitates more in-depth longitudinal modeling investigations. This fact has been recently restated by Reisi-Gahrooei and Peynabar [24] in the case of attributed social networks where data autocorrelations have been addressed for which Extended Kalman Filter (EKF) is employed for parameter estimation.

Woodall et al. [25] presented another review paper on social monitoring researches and chiefly classified statistical process monitoring approaches to detecting anomalies in social networks. Similarity, Sengupta and Woodall [26] briefly reviewed the statistical methods for computer and social networks.

Some other more recent papers include that of Hazrati-Maranagaloo and Noorossana [27] who monitored the probability of edge existence. In another study by Hosseini and Noorossana [28], the performance of EWMA and CUSUM control charts was evaluated for monitoring social networks. For dynamic networks, a method for detecting node propensity changes was investigated by Yu et al. [29]. Sparks [30] proposed a method for detecting periods of significant communication levels of targeted individuals. Komolafe et al. [31] reviewed use of spectral methods, and Mazrae Farahani et al. [32] employed Root Mean Square Error (RMSE) to improve the monitoring schemes for anomaly detection in social networks in terms of ARL criterion. Another interesting research by Fotuhi et al. [33] proposed a novel approach based on multiple correspondence analysis.

Nonetheless, the concepts of dynamism, randomness, and correlations in social network studies need more investigations due to the assumptions that available methods are limited.

One of these important limitations is that the available literature assumes that the whole set of interacting individuals (nodes) is known and fixed over time. For instance, consider the E-mail communications of a company in which the set of nodes is known from the list of staff members who have been assigned an E-mail address and the corresponding attributes of staffs can be easily accessible from the profile data. Moreover, the attributes of the individuals (nodes) such as gender, position in the company, nationality, etc. are chiefly fixed properties and do not change at time stamps. However, in many real-world social networks, individuals of the network may join or leave the networks easily; therefore, the set of nodes may vary at different time stamps. In addition, in many cases, the attributes of nodes may vary at different time stamps. For instance, in a social network of online gamers, each actor might have different levels of credit or rank at different time stamps.

Mogouie et al. [34] proposed a new approach for monitoring binary edge social networks considering random effects, and Najafi and Saghaei [35] worked on monitoring financial networks based on the concepts introduced by Mogouie et al. [34]. Noorossana et al. [36] also presented a review of statistical monitoring methods for social networks.

Such cases are very common in real-world applications. Another instance is the case of online auction in which members can join or leave the site readily, creating a condition that the set of nodes and their corresponding attributes may change.

In this paper, based on the properties of random effects for modeling the network data, a statistical monitoring procedure is proposed for cases where the vectors of attributes 'values of nodes' vary over time. Since the number of communications or the lengths of the messages sent and received between pairs is always of value for social network data analysis, Poisson distribution is considered for the data corresponding to ties. Hence, the modeling based on random effects is capable to work for monitoring the social network data when the set of nodes and their corresponding attributes may change.

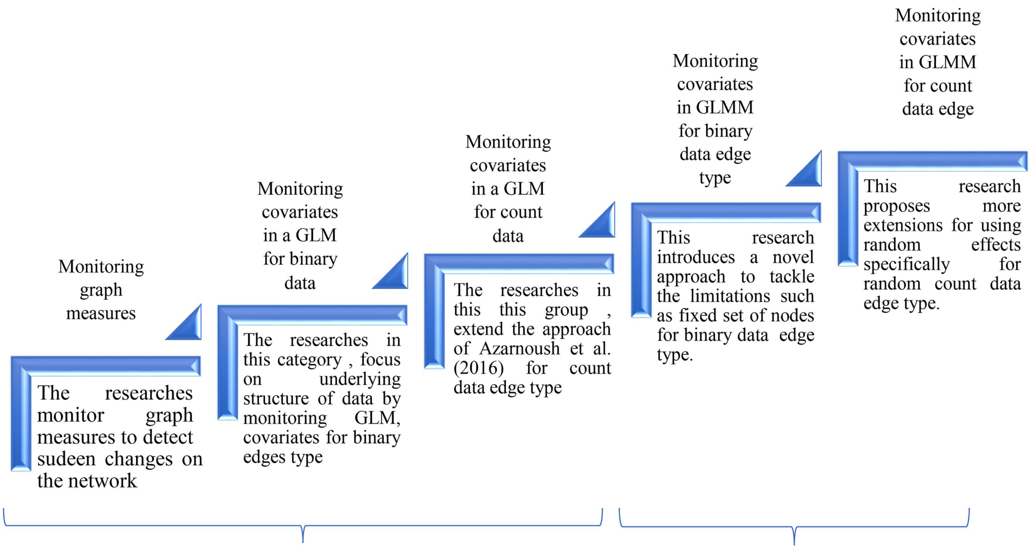
In the proposed approach, consideration of random effects would enable the model to work with cases where the set of nodes and the corresponding attributes vary over time. In the next subsection, the modeling and parameter estimation of count data of communications are discussed. Figure 1 positions the proposed approach more clearly in the most recent literature.

This research is organized as follows: In the second section, the concepts related to random effects modeling for count data are introduced. In the third section, the problem modeling and the parameter estimation method are discussed. The monitoring procedure and the related formulations are given in the fourth section, and numerical examples including simulation studies as well as illustrative examples are discussed in the fifth section. The conclusions and recommendations for future researches are presented in the last section.

2. Random effects model

In most statistical analyses, there are cases where some variables are characterized by levels that must be chosen randomly from a much larger set of levels [37]. Moreover, it is common that data be collected from different clusters and it is known that the data corresponding to a certain cluster have more similar properties while their characteristics differ from the data of other clusters. Ignoring such similarities or differences associated with cluster data may yield erroneous results [38].

Random effects summarize the similarities of data within a cluster. While the same value can be assigned to a certain cluster, it may vary for other clusters. The sources of these variables are not necessarily known or controllable and are called random effects because they are randomly distributed over the selected levels.


Monitoring graph measures
The researches monitor graph measures to detect sudeen changes on the network
Monitoring covariates in a GLM for binary data
The researches in this category , focus on underlying structure of data by monitoring GLM, covariates for binary edges type
Monitoring covariates in a GLM for count data
The researches in this this group , extend the approach of Azarnoush et al. (2016) for count data edge type
Monitoring covariates in GLMM for binary data edge type
This research introduces a novel approach to tackle the limitations such as fixed set of nodes for binary data edge type.
Monitoring covariates in GLMM for count data edge
This research proposes more extensions for using random effects specifically for random count data edge type.

This stream of research needs fixed set of known nodes. This means the set of individuals in the network is fixed and known from prior.

In the proposed approach, the set of nodes does not necessarily need to be fixed. This approach can be used in cases where, the individuals or their attributes change in different time stamps.

Figure 1. Positioning the proposed method in the latest literature.

Consider there are m panels and each panel is enumerated by the index of j, $j = 1, 2, 3, ..., m$. Also, if there are n individuals and each individual is characterized by a $1{\times}p$ vector of attributes, then the $(n \times p)$ covariate matrix of $\mathbf{X}$ will be available. Hence, the vector of coefficients denoted by $\boldsymbol{\beta}$, assuming the $n \times 1$ vector of model errors $\boldsymbol{\varepsilon}$, the fixed terms defining the $n \times 1$ response vector of $\mathbf{y}$ are obtained by a model such as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}{+}\boldsymbol{\varepsilon}$. However, to incorporate random effects to the model, the $(n \times qm)$ matrix of predictor variables $\mathbf{Z}$ multiplicated by the $(qm \times 1)$ vector of random effects denoted by $\boldsymbol{\delta}$ should be considered in the data structure. Note that q represents the number of random effects, while each of them has m levels.

The notations introduced above are the elements that are needed for most random effect models. Suitable models should be applied due to the data properties. While random effects models have been well studied for normal responses, their developments for non-normal outcomes have not been investigated extensively. A general representation of Generalized Linear Mixed Models (GLMM) introduced by Agresti [38] is given in Eq. (1) as follows:

$$g(\mu_{it}) = X'_{it}\boldsymbol{\beta} + Z'_{it}\boldsymbol{\delta}_i, \tag{1}$$

where $g(.)$ is the link function and $\boldsymbol{\delta}_i$ denotes the random effect vector that follows a multivariate normal distribution $N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the variance-covariance matrix of the random effects.

Random effects are incorporated into the modeling of count data multiplicatively. Consider a model with the multiplicative random effect of α_i; then, the conditional mean of the count data $E[y_{it}|\mathbf{x}_{it}, \alpha_i]$ can be modeled using Eq. (2) [39]:

$$E[y_{it}|\mathbf{x}_{it}, \alpha_i] = \mu_{it} = \alpha_i\lambda_{it} = \alpha_i \exp(\mathbf{X'}_{it}\boldsymbol{\beta}). \tag{2}$$

Note that we are mostly interested in estimating the coefficient parameters $\boldsymbol{\beta}$; to this end, the effect of α_i should be eliminated. Note that α_i represents iid random variables. Although the random effects are added to the model in Eq. (1) multiplicatively, they can be still interpreted as the cause of a shift in the intercept, as shown in Eq. (3):

$$\mu_{it} = \alpha_i \exp(\mathbf{X'}_{it}\boldsymbol{\beta}) = \exp(\delta_i + \mathbf{X'}_{it}\boldsymbol{\beta}), \tag{3}$$

where $\delta_i = \ln \alpha_i$. The notion that the random effects shift the intercept is related to the type of the link function of $g(\mu_{it})$, denoted by exp function, and it does not necessarily hold for all other types of link functions.

3. Problem modelling

This section discusses the problem modeling, gives the required formulations for network modeling, and evaluates the corresponding model parameters.

3.1. Network modelling

Consider a network graph of $G(t) = \{V(t), Y(t)\}$ at each time stamp $t = 1, 2, ..., T$ where $V(t) = \{v_1, v_2, ..., v_v\}$ and $Y(t) = \{y_{12}(t), ..., y_{ij}(t), ..., y_{v-1,v}\}$ denote the sets of individuals in terms of vertices and the links in terms of ties, respectively. Since the problem under study should be analyzed based on the attributed network data and the attributes of vertices collected for each pair of individuals i and j at time stamp t, the $p \times 1$ vector of similarities between these two individuals is $x_{ijt} = \{x_{ijt1}, x_{ijt2}, ..., x_{ijtp}\}$.

In this vector, x_{ijtp} for $p = 1, 2, 3, ..., p$ shows how individuals i and j are similar to each other and then, they are compared considering the pth attribute at time stamp t. In social network data analysis, common attributes of interest include sex, position, place or age, etc. and there should be reference criteria of how to compare individuals to determine x_{ijtp}. For instance, if the pth attribute is gender in case both individuals i and j are of the same sex, then $x_{ijtp} = 1$; else, 0.

In the available literature, the proposed methods assume that the set of vertices does not change over time and accordingly, the vectors of attributes are fixed. Moreover, the numerical examples that have been analyzed represent cases where the whole data set of individuals is available. However, in real-world applications, these assumptions mostly do not necessarily hold. As shown in Figure 1, in many cases, we may have access to only some of the randomly selected nodes and analyze them based on the available set of vertices, as shown by red dots in Figure 2.

3.2. Random effects count data modeling and parameters estimation

A commonly applicable random effects model for count data is the Poisson random effects where y_{kt} conditional on α_k and λ_{kt} follows Poisson distribution with the parameter of $\mu_{kt} = \alpha_k\lambda_{kt}$. Note that the index k is equivalent to the kth pair of individuals. In other words, if there are n individuals in a network, there

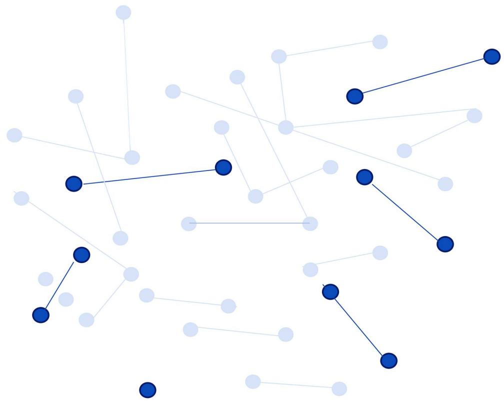


Figure 2. A sample network with randomly accessible nodes.

$$l_{R_0} = \log\left\{ \prod_{t\in R_0}^{\cdot}\prod_{i=1}^{\nu}\prod_{j\neq i}^{\cdot} \frac{[\exp\left((\mathbf{X'}_{ijt}\boldsymbol{\beta}_{R_0} + \boldsymbol{\delta}_{R_0\,ijt})\right]^{y_{ijt}}.[\exp(-(\mathbf{X'}_{ijt}\boldsymbol{\beta}_{R_0} + \boldsymbol{\delta}_{R_0\,ijt}))]}{y_{ijt}!} \right\}. \tag{6}$$

Box I

$$l_{R_\tau} = \log\left\{ \prod_{t\in \tau}^{\cdot}\prod_{i=1}^{\nu}\prod_{j\neq i}^{\cdot} \frac{[\exp\left((\mathbf{X'}_{ijt}\boldsymbol{\beta}_{R_\tau} + \boldsymbol{\delta}_{R_\tau\,ijt})\right]^{y_{ijt}}.[\exp(-(\mathbf{X'}_{ijt}\boldsymbol{\beta}_{R_\tau} + \boldsymbol{\delta}_{R_\tau\,ijt}))]}{y_{ijt}!} \right\}. \tag{7}$$

Box II

will be $\binom{n}{2}$ pairs of individuals, each denoted by a number from the index k.

The longitudinal analysis of the Poisson distributed data considering random effects is done based on the joint density of the kth pair for $t = 1, 2, ..., T$ time stamp, as shown in Eq. (4):

$$\prod_t \left(\frac{\lambda_{kt}}{y_{kt}!}\right)\left(\frac{\delta}{\sum\limits_t \lambda_{kt} + \delta}\right)^{\delta}\left(\sum_t \lambda_{kt} + \delta\right)^{-\sum\limits_t y_{kt}} \frac{\Gamma\left(\sum\limits_t y_{kt} + \delta\right)}{\Gamma(\delta)}. \tag{4}$$

The regression analysis of random effects count data is conducted under the influence of distribution of α_k; however, a suitable model for many applications such as social network data could be gamma (δ, δ). In this case, the maximum likelihood estimation of $\boldsymbol{\beta}$ and δ considering $\lambda_{kt} = \exp(x'_{kt}\boldsymbol{\beta})$ can be done through Eq. (5) as follows:

$$\sum_{k=1}^{\binom{n}{2}} \sum_{t=1}^{T} \mathbf{x}_{kt}\left(y_{kt} - \lambda_{kt}\frac{\bar{y}_k + \frac{\delta}{T}}{\bar{\lambda}_k + \frac{\delta}{T}}\right) = 0, \tag{5}$$

where $\bar{y}_k$ denotes the average of y_k's. This estimation has been performed based on the elimination of random effects; accordingly, $\hat{\boldsymbol{\beta}}$ is the main output of this procedure and the estimated values of δ are not necessarily of interest. In the next section, the monitoring procedure is explained in detail.

4. Monitoring procedure

This section proposes an LRT-based procedure to monitor the dynamic network considering static and dynamic reference set approaches. The static reference addresses an approach in which the computation of the LRT statistic applies a fixed set of network data as the reference set R_0, while the dynamic one updates the reference set by substituting the data of the last time stamp with the first ones. Hence, the most recent data are used as a dynamic window reference set. For this purpose, the likelihood functions at each time stamp are calculated as follows.

The parameters of the count model shown in Eq. (3) yield the log likelihood value of l_{R_0} represented in Eq. (6) as shown in Box I. At time stamp τ, the value of the log likelihood function is obtained using Eq. (7) as shown in Box II. Similarly, $R_1 = R_0 \cup \tau$; then, l_{R_1} is obtained using Eq. (8) as shown in Box III. When the network is in-control, the model parameters of the network at time stamp τ are not significantly different

$$l_{R_1} = \log\left\{ \prod_{t\in R_0}^{\cdot}\prod_{i=1}^{\nu}\prod_{j\neq i}^{\cdot} \frac{[\exp\left((\mathbf{X'}_{ijt}\boldsymbol{\beta}_{R_0} + \boldsymbol{\delta}_{R_0\,ijt})\right]^{y_{ijt}}.[\exp(-(\mathbf{X'}_{ijt}\boldsymbol{\beta}_{R_0} + \boldsymbol{\delta}_{R_0\,ijt}))]}{y_{ijt}!} \right.$$
$$\left. \prod_{t\in \tau}^{\cdot}\prod_{i=1}^{\nu}\prod_{j\neq i}^{\cdot} \frac{[\exp\left((\mathbf{X'}_{ijt}\boldsymbol{\beta}_{R_\tau} + \boldsymbol{\delta}_{R_\tau\,ijt})\right]^{y_{ijt}}.[\exp(-(\mathbf{X'}_{ijt}\boldsymbol{\beta}_{R_\tau} + \boldsymbol{\delta}_{R_\tau\,ijt}))]}{y_{ijt}!} \right\}. \tag{8}$$

Box III

$$l_0 = \log\left\{ \prod_{t\in R_0}^{\cdot}\prod_{i=1}^{\nu}\prod_{j\neq i}^{\cdot} \frac{[\exp\left((\mathbf{X'}_{ijt}\boldsymbol{\beta}_0+\boldsymbol{\delta}_{0\,ijt})\right]^{y_{ijt}}.[\exp(-(\mathbf{X'}_{ijt}\boldsymbol{\beta}_0+\boldsymbol{\delta}_{0\,ijt}))]}{y_{ijt}!}\right.$$
$$\left.\prod_{t\in \tau}^{\cdot}\prod_{i=1}^{\nu}\prod_{j\neq i}^{\cdot} \frac{[\exp\left((\mathbf{X'}_{ijt}\boldsymbol{\beta}_0+\boldsymbol{\delta}_{0\,ijt})\right]^{y_{ijt}}.[\exp(-(x'_{ijt}\boldsymbol{\beta}_0+\boldsymbol{\delta}_{0\,ijt}))]}{y_{ijt}!}\right\}. \tag{9}$$

Box IV

$$l_1 = \log\left\{ \prod_{t\in R_0}^{\cdot}\prod_{i=1}^{\nu}\prod_{j\neq i}^{\cdot} \frac{[\exp\left((\mathbf{X'}_{ijt}\boldsymbol{\beta}_0+\boldsymbol{\delta}_{0\,ijt})\right]^{y_{ijt}}.[\exp(-(\mathbf{X'}_{ijt}\boldsymbol{\beta}_0+\delta_{0\,ijt}))]}{y_{ijt}!}\right.$$
$$\left.\prod_{t\in \tau}^{\cdot}\prod_{i=1}^{\nu}\prod_{j\neq i}^{\cdot} \frac{[\exp\left((\mathbf{X'}_{ijt}\boldsymbol{\beta}_\tau+\delta_{\tau\,ijt})\right]^{y_{ijt}}.[\exp(-(\mathbf{X'}_{ijt}\boldsymbol{\beta}_\tau+\boldsymbol{\delta}_{\tau\,ijt}))]}{y_{ijt}!}\right\}. \tag{10}$$

Box V

from that of the reference set R_0 and the parameters can be considered as $\boldsymbol{\beta}_0$ for the whole time horizon of R_1. Hence, the in-control value of the log likelihood function can be formulated through Eq. (9) as shown in Box IV. However, when an assignable cause shifts the model parameters at time stamp τ, the parameters are not statistically equal to $\boldsymbol{\beta}_0$ and the corresponding log likelihood function should be calculated using Eq. (10) as shown in Box V. In Eqs. (9–11), the model parameters $\boldsymbol{\beta}_{R_0}$, $\boldsymbol{\beta}_{R_\tau}$, and $\boldsymbol{\beta}_0$ are unknown and they should be estimated using Eq. (5). However, software packages including MATLAB can be employed in this context. Finally, the LRT-based statistic is computed using Eq. (11):

$$\Lambda(\tau) = 2(l_1 - l_0). \tag{11}$$

The next step for the monitoring procedure is determination of the Upper Control Limit (UCL) for which a simulation approach is applied to satisfy the desired in-control ARL. The network is changed into the out-of-control state when $\Lambda(\tau)$ falls out of the obtained UCL.

5. Simulation studies

In this section, the performance of the proposed method is evaluated using simulation studies considering static and dynamic reference sets. To this end, a model is first defined for data generation. This model generates simulated data. However, for a monitoring procedure, we should determine in advance how to monitor the data considering a determined control limit. To have an in-control ARL value equal to 200 for both of dynamic and static reference sets, the corresponding UCLs are determined through 10000 simulation runs. For this reason, using a search approach, we set different control limits alternately until the determined limits satisfy the in-control ARL value equal to 200. For confirming the performance of the proposed method in detecting the out-of-control states, the parameters of the assumed model are shifted and the out-of-control ARLs are calculated for each shift. As the shifts increase, it is expected that the out-of-control ARLs decrease. In other words, given that the designed control chart detects shifts faster, it is implied that the proposed method is more reliable for real-time monitoring.

For simulation studies, we consider the model based on fixed and variable covariates over time. An important advantage of the random effects model for monitoring social networks is that the covariate vector may change over time. In other words, the mean of counts of links between individuals i and j, λ_{ijt}, is modeled in terms of the similarity vector of individuals i and j at time t. Hence, this model facilitates modeling the networks in which the attributes of nodes change at different time stamps. The following equation presents the assumed model for data generation of the fixed model:

$$\lambda_{ij} = \exp(\delta_{ij} + 0.5x_{ij1} + 0.5x_{ij2} + 0.5x_{ij3}).$$

Similarly, the variable covariate model is considered as follows:

$$\lambda_{ijt} = \exp(\delta_{ijt} + 0.3x_{ij1t} + 0.3x_{ij2t} + 0.3x_{ij3t}).$$

Note that δ_{ij} follows Gamma distribution with the parameters of $(\alpha = 4, \lambda = 4)$ and $(\alpha = 2, \lambda = 2)$ in the above models, respectively. The results of the simulation runs are presented in the next two subsections for the fixed and variable covariate models, respectively.

5.1. Simulation results for the fixed covariate model

Considering the aforementioned model properties for the fixed covariate model, the UCLs of the static and dynamic LRT-based methods are obtained equal to 370 and 392, respectively. By imposing shifts in the model parameters of β_1, β_2, and β_3, the corresponding ARL curves are obtained in Figures 3 to 5.

As shown in Figure 3, by shifting the β_1 value from 0.5 to 0.585, the ARL decreases considerably to 1, which means that the chart is able to detect small shifts effectively. The steeper the curve is, the more sensitive the designed chart is.

Similarly, for the parameter of β_2, shift from the value of 5.55 to 0.8 shows that the method performs well, even for step changes as small as 0.05. Also, the ARL values under different shifts in the parameter of β_3, also illustrate a similar performance that confirms the effectiveness of the method, as shown in Figure 5.

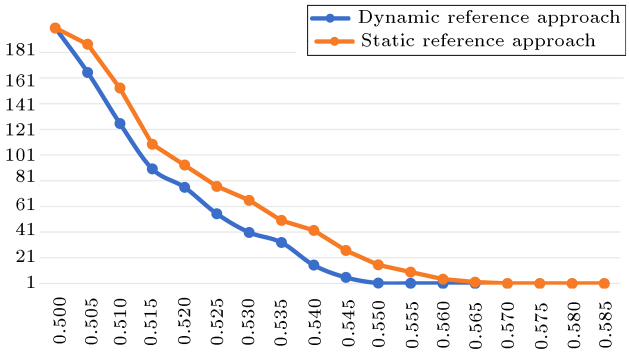


Figure 3. ARL curves under different shifts in β_1 (Fixed covariate model) considering static and dynamic reference set approaches.

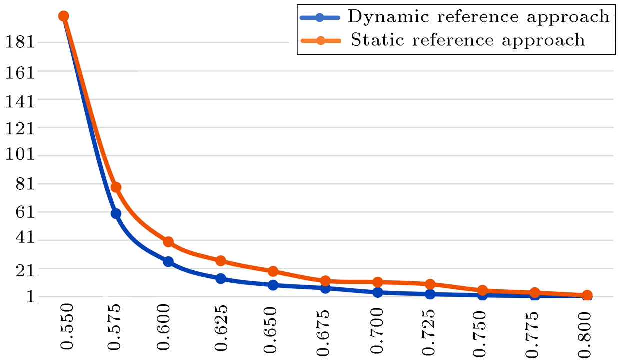


Figure 4. ARL curves under different shifts in β_2 (Fixed covariate model) considering static and dynamic reference set approaches.

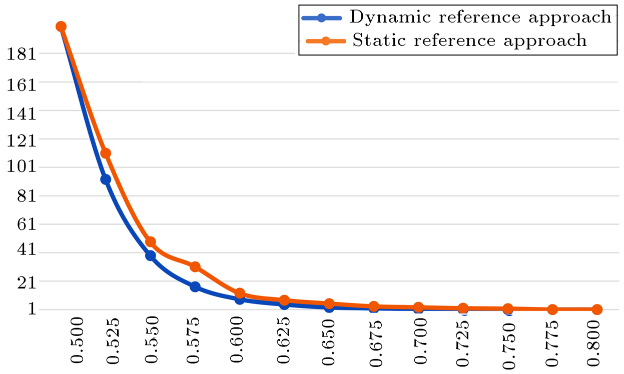


Figure 5. ARL curves under different shifts in β_3 (Fixed covariate model)considering static and dynamic reference set approaches.

In the next subsection, the performance of the proposed method for the variable covariate model is investigated.

5.2. Simulation results for variable covariate model

In the variable covariate model, the vectors of similarities are generated randomly at each time stamp. Considering the model properties for the variable covariate model, the UCLs of the dynamic and static LRT-based methods are obtained equal to 415 and 432, respectively. By shifting the model parameters of β_1, β_2, and β_3, the corresponding ARL values are obtained and shown in Figures 6 to 8.

The ARL values shown in Figures 6–8 demonstrate the efficacy of the proposed method in detecting shifts in the model parameters for the variable attributes. Similar to the discussion of the previous subsection, the dynamic LRT-based approach performs more accurately than the static approaches.

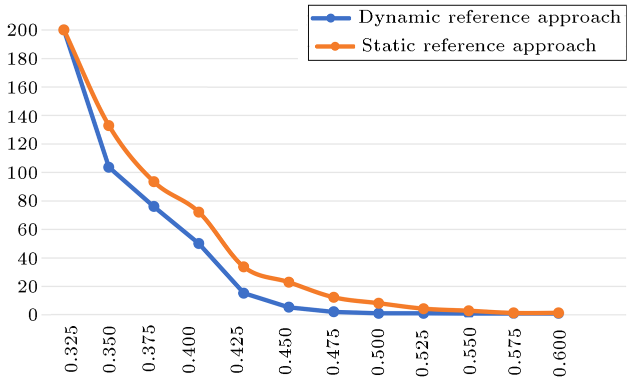


Figure 6. ARL curves under different shifts in β_1 (Variable covariate model) under static and dynamic reference set approaches.

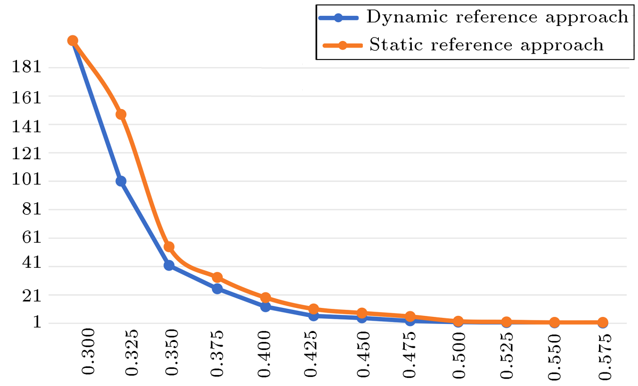


Figure 7. ARL curves under different shifts in β_2 (Variable covariate model) under static and dynamic reference set approaches.

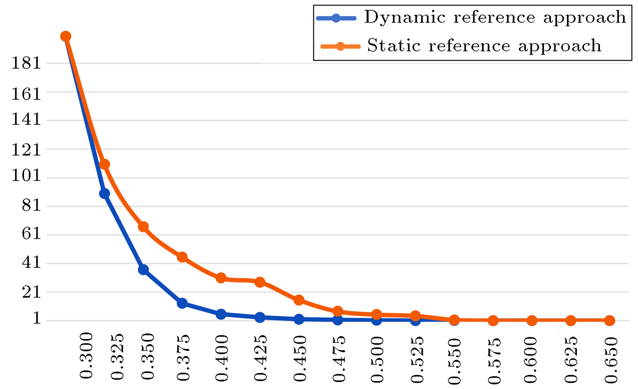


Figure 8. ARL curves under different shifts in β_3 (Variable covariate model) considering static and dynamic reference set approaches.

5.3. An illustrative example

Consider the example of an online market network of digital products to which different firms can register so as to have a profile on the website and present their products and search for other firms' profiles to make deals. Since the connections take place on a website and each firm is recognized by its profile, the corresponding attributes can be obtained based on profile information. Hence, each firm as a node can have several attributes such as the number of previous sales on the website, number of physical branches, and the rank of the firms in the market. Due to the policies of the owner of the website, the example under study is discussed anonymously.

Figure 9 demonstrates a schematic of the e-firms in the network under study. The thickness of the links is proportional to the number of the deals made between each pair of the firms.

When there is a trade between two firms, they are considered connected whether in terms of the number of deals at each time stamp, the value of the trade, or any other type of links as edges. To apply the proposed monitoring method to an illustrative example, one needs to estimate the model parameters and determine the corresponding control limit. In the example under study, by applying the data set of a 50-week time horizon, the model parameters are estimated in the first step model as follows:

$$\lambda_{ijt} = \exp(1 + 0.05x_{ij1t} + 0.07x_{ij2t} + 0.08x_{ij3t}).$$

In this model, the attributes are properties such as the number of previous sales on the website, number of physical branches, and the rank of the firms denoted by x_1, x_2, and x_3, respectively. By comparing each pair of firms together considering the specified attributes, the corresponding covariate vector of x_{ijt} is determined. Of note, the considered attributes follow a discrete uniform distribution of [1–20], [1–3], and [1–4], respectively.

In the next step, the LRT-based procedure given by Eq. (11) is applied to satisfy an in-control ARL equal to 200. For verification of the proposed method, a one-year time horizon of the data set is employed to evaluate the performance of the proposed method and the corresponding control chart is shown in Figure (9).

In the next step, using the model in Eq. (11), the UCL equals 4.4 that has been determined through 10000 simulation runs satisfying the in-control ARL equal to 200. For verification of the proposed method functionality, a 50-week time horizon of the data set was employed to evaluate the performance of the proposed method and the corresponding control chart is shown in Figure (10).

As is shown in Figure 8, results of shifts in the values of the first coefficient from 0.05 to 0.08 from Week 26 onwards led to a significant change in statistic value and an out-of-control state was presented from then on. Obtained evidence may confirm the satisfactory performance of the proposed method.

6. Conclusion and recommendations for future research

In this paper, a novel method was proposed for monitoring social networks with count data based on random effects concept. The applied modeling enabled the monitoring procedure to detect structural shifts in both of the networks with a fixed set of individuals and the variable attribute ones. Moreover, the incorporation of random effects concepts to the model improved the applicability of the monitoring procedure for the networks with variable covariates. The performance of the proposed method was evaluated in terms of ARL. Due to the improvements that the proposed method brings about, further research on network data with other distributions of ties such as ordinal data is recommended for future research studies.

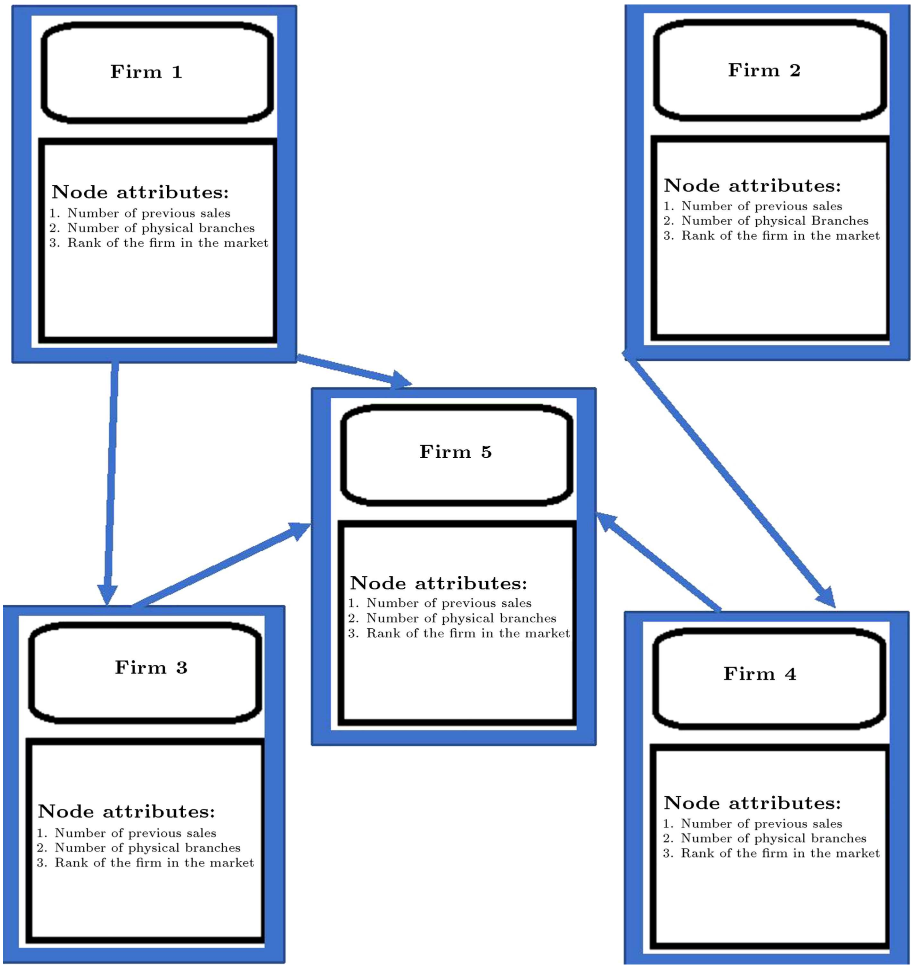

Firm 1
Node attributes:
1. Number of previous sales
2. Number of physical branches
3. Rank of the firm in the market
Firm 2
Node attributes:
1. Number of previous sales
2. Number of physical Branches
3. Rank of the firm in the market
Firm 5
Node attributes:
1. Number of previous sales
2. Number of physical branches
3. Rank of the firm in the market
Firm 3
Node attributes:
1. Number of previous sales
2. Number of physical branches
3. Rank of the firm in the market
Firm 4
Node attributes:
1. Number of previous sales
2. Number of physical branches
3. Rank of the firm in the market

Figure 9. Illustration of the firms' network.

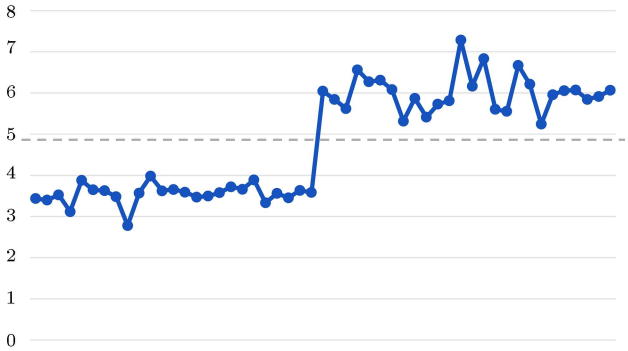


Figure 10. LRT-based control chart under the dynamic reference set approach.

References

1. Erdös, P. and Rényi, A. "On random graphs", *I. PUBL MATH-DEBRECEN (Debrecen)*, **6**, pp. 290–297 (1959).

2. Leskovec, J., McGlohon, M., Faloutsos, C., et al. "Patterns of cascading behaviour in large blog graphs", *Proceedings of the 2007 SIAM International Conference on Data Mining, Society for Industriies and Applied Mathematics*, USA, July (2007).

3. Tufekci, Z. and Wilson, C. "Social media and the decision to participate in political protest: Observations from tahrir square", *J. Commun*, **62**(2), pp. 363–379 (2012).

4. Shetty, J. and Adibi, J. "Discovering important nodes through graph entropy the case of Enron email database", In *Proceedings of the 3rd International Workshop on Link Discovery*, pp. 74–81, ACM (2005).

5. Krebs, V.E. "Mapping networks of terrorist cells", *Connections*, **24**(3), pp. 43–52 (2002).

6. Pandit, V., Modani, N., Mukherjea, S., Nanavati, A.A., et al. "Extracting dense communities from telecom call graphs", In *Communication Systems Software and Middleware and Workshops, COMSWARE 2008, 3rd International Conference*, pp. 82–89 IEEE (2008).

7. Savage, D., Zhang, X., Yu, X., et al. "Anomaly detection in online social networks", *Soc. Netw*, **39**, pp. 62–70 (2014).

8. Cheng, A. and Dickinson, P. "Using scan-statistical correlations for network change analysis", *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, Berlin, Heidelberg (2013).

9. Mcculloh, I. and Carley, K.M. "Detecting change

in longitudinal social networks", Journal of Social Structure, **12**(1), pp. 1–37 (2011).

10. Miller, B.A., Arcelano, N., and Bliss, N.T. "Efficient anomaly detection in dynamic, attributed graphs: Emerging phenomena and big data", *IEEE International Conference on Intelligence and Security Informatics (ISI)*, USA, June (2013).

11. Bliss, C.A., Frank, M.R., Danforth, C.M., et al. "An evolutionary algorithm approach to link prediction in dynamic social networks", *J Comput Sci-Neth*, **5**(5), pp. 750–764 (2014).

12. Woodall, W.H., Zhao, M.J., Paynabar, K., et al. "An overview and perspective on social network monitoring", *IEEE Trans. Ind. Appl.*, **49**(3), pp. 354–365 (2017).

13. Heard, N.A., Weston, D.J., Platanioti, K., et al. "Bayesian anomaly detection methods for social networks", *Ann. Appl. Stat.*, **4**(2), pp. 645–662 (2010).

14. Priebe, C.E., Conroy, J.M., Marchette, D.J., et al. "Scan statistics on enron graphs", *Comput. Math. Organ. Th.*, **11**(3), pp. 229–247 (2005).

15. Sparks, R. "Monitoring communications: aiming to identify periods of unusually increased communications between parties of interest", *Qual. Technol. Quant. M.*, **13**(1), pp. 39–57 (2016).

16. Neil, J., Hash, C., Brugh, A., et al. "Scan statistics for the online detection of locally anomalous subgraphs", *Technometrics*, **55**(4), pp. 403–414 (2013).

17. Marchette, D. "Scan statistics on graphs", *Wiley Interdisciplinary Reviews, Computation Stat*, **4**(5), pp. 466–473 (2012).

18. Pincombe, B. "Anomaly detection in time series of graphs using arma processes", *Australian Society for Operations Research Bulletin*, **24**(4), pp. 2–7 (2005).

19. McCulloh, I. and Carley, K. "Detecting change in human social behaviour simulation", *Center for Computational Analysis of Social and Organizational Systems*, Carnegie Mellon University, Pittsburgh, PA 15213 (2008a).

20. McCulloh, I. and Carley, K. "Social network change detection", *Institute for Software Research School of Computer Science Carnegie Mellon University Pittsburgh*, PA 15213 (2008b).

21. Azarnoush, B., Paynabar, K., Bekki, J., et al. "Monitoring temporal homogeneity in attributed network streams", *J Qual Technol*, **48**(1), pp. 28–43 (2016).

22. Farahani, E.M., Baradaran Kazemzadeh, R., Noorossana, R., et al. "A statistical approach to social network monitoring", *Commun Stat-Theor M.*, **46**(22), pp. 11272–11288 (2017).

23. Fotuhi, H., Amiri, A., and Maleki, M.R. "Phase I monitoring of social networks based on Poisson regression profiles", *Qual Reliab Eng Int*, **34**(4), pp. 1–17 (2018).

24. Reisi-Gahrooei, M. and Peynabar, K. "Change detection in a dynamic stream of attributed networks", *Journal of Quality Technology*, **50**(4), pp. 418–430 (2018). DOI: 10.1080/00224065.2018.1507558

25. Woodall, W.H., Zhao, M.J., Paynabar, K., et al. "An overview and perspective on social network monitoring", *IEEE Trans. Ind. Appl.*, **49**(3), pp. 354–365 (2017).

26. Sengupta, S. and Woodall, W.H. "Discussion of statistical methods for network surveillance", *Appl Stoch Model Bus*, **34**(4), pp. 446–448 (2018).

27. Hazrati-Marangaloo, H. and Noorossana, R. "Detecting outbreaks in temporally dependent networks", *Qual Reliab Eng Int.*, **35**(6), pp. 1753–1765 (2019).

28. Hosseini, S.S. and Noorossana, R. "Performance evaluation of EWMA and CUSUM control charts to detect anomalies in social networks using average and standard deviation of degree measures", *Qual Reliab Eng. Int.*, **34**(4), pp. 477–500 (2018).

29. Yu, L., Woodall, W.H., and Tsui, K.L. "Detecting node propensity changes in the dynamic degree corrected stochastic block model", *Social Networks*, **54**, pp. 209–227 (2018).

30. Sparks, R. "Detecting periods of significant increased communication levels for subgroups of targeted individuals", *Qual Reliab Eng Int*, **32**(5), pp. 1871–1888 (2016).

31. Komolafe, T., Quevedo, A.V., Sengupta, S., et al. "Statistical evaluation of spectral methods for anomaly detection in networks", *Network Science*, **7**(3), pp. 319–352 (2019).

32. Mazrae Farahani, E., Baradaran Kazemzade, R., Albadvi, A., et al. "Modeling and monitoring social network in term of longitudinal data", *Int. J. Ind. Eng. Comput.*, **29**(3), pp. 247–259 (2018).

33. Fotuhi, H., Amiri, A., and Taheriyoun, A.R. "A novel approach based on multiple correspondence analysis for monitoring social networks with categorical attributed data", *J Stat Comput Sim.*, **89**(16), pp. 3137–3164 (2019).

34. Mogouie, H., Raissi-Ardali, G.A., Bahrami-Samani, E., et al. "Statistical monitoring of binary response attributed social networks considering random effects", *Communications in Statistics-Simulation and Computation*, **51**(3), pp. 973–992 (2022). DOI: 10.1080/03610918.2019.1661471

35. Najafi, H. and Saghaei, A. "Statistical monitoring for change detection of interactions between nodes in networks: with a case study in financial interactions network, *Communications in Statistics-Theory and Methods*, **50**(20), pp. 4900–4911 (2021). DOI: 10.1080/03610926.2020.1725830

36. Noorossana, R., Hosseini, S.S., and Heydarzade, A. "An overview of dynamic anomaly detection in social networks via control charts", *Qual Reliab Eng Int.*, **34**(4), pp. 641–648 (2018).

37. Myers, R.H., Montgomery, D.C., Vining, G.G., et al., *Generalized Linear Models: with Applications in Engineering and the Sciences*, John Wiley & Sons (2012).

38. Agresti, A., *Categorical Data Analysis*, John Wiley & Sons (2013).

39. Cameron, A.C. and Trivedi, P.K. “Regression analysis of count data”, **53**, Cambridge University Press, (2013).

Biographies

Hamed Mogouie is a PhD candidate in Industrial Engineering at Isfahan University of Technology in Iran. His fields of research include statistical process monitoring, design of experiments, and quality management. He is a member of Iranian elite’s community of ministry of energy and has experienced collaboration with notable research centers in Australia. He is recently working on monitoring social networks considering statistical models.

Gholam Ali Raissi Ardali is an Associate Professor of Industrial Engineering and is the Head of the Industrial and Systems Engineering Faculty at Isfahan University of Technology in Iran. He has a wide range of experiences: founding educational institutes, reengineering large-scale industries, giving consultation in different public and private sectors, and academic activities in the area of Total Quality Management.

Amirhossein Amiri is an Associate Professor at Shahed University in Iran. He holds BS, MS, and PhD degrees 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 the Director of Postgraduate Education 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 international high-quality 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.

Ehsan Bahrami Samani is an Associate Professor in Statistics at Shahid Beheshti University in Iran. His research area is developing novel complex models for social network, health, and longitudinal analyses. His recent activities are about introducing zero inflated models in social network data which can be pioneering in this field of science.