Monitoring Attributed Social Networks Based on Count Data and Random Effects

Hamed Mogouie[1], Gholam Ali Raissi Ardali[2], Amirhossein Amiri[3*], Ehsan Bahrami Samani[4]

[1,2] Department of industrial and Systems Engineering, Isfahan University of Technology, Isfahan, Iran.

[1]**Email** : Raissi@cc.iut.ac.ir **Phone** : +98 09131161815

[2]**Email** : H.mogouie@in.iut.ac.ir **Phone** : +98 09124646089

[3] Department of Industrial Engineering, Faculty of Engineering, Shahed University, Tehran, Iran.

[3*]**Email** : amiri@shahed.ac.ir **Phone** : +0989123236909

[4] Department of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran.

[4]**Email** : E_bahrami@sbu.ac.ir **Phone** : +0989127237214

Abstract

This paper presents a novel approach for the statistical monitoring of online social networks where the edges represent the count 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 tackles these limitations due to the properties of the random effect concepts. Applying appropriate parameters estimation technique involved in a likelihood ratio testing (LRT) approach considering two different statistics, 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.

Keywords: Count data, Random effects, Social networks, Statistical monitoring.

1. Introduction

The analysis of network data has been always of the interest of researchers since networks demonstrate a comprehensive image of complex systems. The initial studies such as those by Erdos and Renyi [1] incorporated probabilistic concepts to graph theory to analyze network data. Although such an approach was extended by other researchers such as Leskovec *et al*. [2], ignoring the dynamic property of network data necessitated more in-depth studies which are capable to analyze network streams over time.

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

The surveillance of online networks has led to introduction of methods for tracing central actors in terrorist groups proposed by Kerbs [5]. Similarly, the detection of anomalies such as systematic cheating in online networks which is very crucial for monetary systems and the insurance industries showed the need to the studies like that of by Pandit *et al*. [6].

The literature review of anomaly detection proposed by Savage *et al.* [7] applies a categorization based on two criteria of whether the network under study is labeled or unlabeled and if the surveillance is conducted statically or dynamically. Far apart from 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], McCulloh and Carely [9].

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

Woodall *et al.* [12] proposed a review paper for classifying the anomaly detection where the notable studies, which propose a certain approach, are reviewed briefly. For instance, the research conducted by Heard *et al*. [13] is introduced as an instance for applying Bayesian methods in social network analysis, and the scan statistic approach is discussed by referring to the works of Priebe *et al*. [14], Sparks [15], Neil [16], Marchetti [17]. The other approach that has been focused in the literature is the application of time series analysis for monitoring social networks such as the study presented by 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 the monitoring applying exponentially weighted moving average (EWMA) charts. The articles by McCulloh and Carley [19] and [20] are the well-known studies that can be categorized in this class.

The extension of the researches which used control charts can be found in the article of Azarnoush *et al*. [21]. The main draw backs which this study has addressed is that focusing

on graph measures does not necessarily lead to detecting structural anomalies. To overcome such a challenge, Azarnoush *et al*. [21]. modelled the probability of communication of each pair of individuals in terms of a similarity vector which is derived from a predefined method of comparing the corresponding individuals of a certain pair. Then using generalized linear modelling (GLM) concepts and a LRT approach, LRT-based statistics are computed to be monitored. An important advantage of this method is focusing on the underlying structure of the network data which facilitates a general insight of how individuals are likely to connect each other based on their similarities of attributes. Such an approach has been extended by Farahani *et al.* [22] and Fotouhi *et al.* [23] focusing on Poisson distributed edge data instead of binary outcomes.

In the researches aforementioned, although significant progresses can be traced to propose more applicable models for real world networks, significant limitations still exist that should be accounted for. An important notion that has been ignored in the research by Azarnoush *et al*. [21], Farahani *et al.* [22] and Fotouhi *et al.* [23] is paying attention to the fact that dynamism of network data necessitates more in-depth longitudinal modelling investigations.

This fact has been recently stated by Reisi-Gharoei and Peynabar [24] for attributed social networks in which autocorrelations of data have been addressed for which Extended Kalman Filter (EKF) is used for parameters estimation.

Another review article that introduces the social monitoring researches is the work by Woodall *et al*. [25] which chiefly classifies statistical process monitoring approaches for detecting anomalies in social networks. Similarity Sengupta and Woodall [26] briefly review the statistical methods for computer and social networks.

Some other more recent papers include the study by Hazrati-Maranagaloo and Noorossana [27] in which the probability of edge existence is monitored. In another study by Hosseini and Noorossana [28] the performance of EWMA and CUSUM control charts is evaluated for monitoring social networks. For dynamic networks, a method for detecting node propensity changes is investigated by Yu *et al*. [29]. Sparks [30] proposed a method for detecting periods of significant communication levels of targeted individuals. Spectral methods were also reviewed by Komolafe *et al*. [31] and Mazrae Farahani *et al.* [32] used 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 to.

One of these important limitations is that the available literature assumes that the whole set of interacting individuals (nodes) are known and fixed over time. For instance, consider the email communications of a company in which the set of nodes is known from the list of staffs who have been assigned an email 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 and etc. are chiefly fixed properties and do not change in time stamps. However, in many real-world social networks, individuals of the network may join or a leave the networks easily so the set of nodes may vary in different time stamps. In addition, in many cases also the attributes of nodes may vary in different time stamps for instance, in a social network of online gamers, each actor might have different level 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 article on statistical monitoring methods for social networks.

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

In this paper, based on the properties of random effects for modelling 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 the interest in social network data analysis, Poisson distribution is considered for the data corresponding to ties. Hence, the modelling that is 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 modelling and parameters estimation of count data of communications are discussed. Figure 1, positions the proposed approach more clearly in the most recent literature.

Insert Figure 1 about here.

This paper is organized as follows: In the second section the random effects concepts modelling for count data are introduced. In the third section the problem modelling and the parameter estimation method is 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 have levels which have to be chosen randomly from a much larger set of levels (Myers *et al.* [37]). Moreover, it is common that data are collected from different clusters and it is known that the data corresponding to a certain cluster have more similar properties while their characteristics are different from the data of other clusters. The ignorant of such similarities of within cluster data and between cluster differences may lead to erroneous results (Agresti, [38]).

Random effects summarize the similarities of data within a cluster and while having the same value for a certain cluster, its' value varies 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.

For modelling a set of data clustered in the panels of $j = 1,2,3,...,m$, the $(n \times p)$ covariate matrix of $\mathbf{X}$ and the $p \times 1$ 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 built up. However, for incorporation of 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, however suitable models should be applied due to the properties of data. 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 Equation (1) as follows:

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

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 incorporate in to the modelling of count data multiplicatively. Consider a model with the multiplicative random effect of α_i, then the conditional mean of the count data $E[y_{it} \mid \mathbf{x}_{it}, \alpha_i]$, can be modelled using Equation (2), (Cameron and Trivedi, [39]):

$$\begin{aligned} E[y_{it} \mid \mathbf{x}_{it}, \alpha_i] &= \mu_{it} \\ &= \alpha_i \lambda_{it} \\ &= \alpha_i \exp(\mathbf{x}_{it}^{'}\boldsymbol{\beta}). \end{aligned} \tag{2}$$

Note that we are mostly interested in estimation of the coefficients parameters $\boldsymbol{\beta}$ for the aim of which the effect of α_i should be eliminated. Note that α_i are iid random variables. Although the random effects enter the model in Equation (1) multiplicatively, they can be still interpreted to be the cause of a shift in the intercept as shown in Equation (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 $g(\mu_{it})$, herein set as $\exp$ function and it does not necessarily hold for all other types of link functions.

3. Problem modelling

In this section the problem modelling is discussed and the required formulations are given for network modelling and the corresponding model parameters estimation.

3.1.Network modelling

Consider a network graph of $G(t) = \{V(t), Y(t)\}$ at each time stamp $t = 1, 2, ..., T$ where $V(t) = \{\upsilon_1, \upsilon_2, ..., \upsilon_\upsilon\}$ and $Y(t) = \{y_{12}(t), ..., y_{ij}(t), ..., y_{\upsilon\text{-}1,\upsilon}\}$ 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 through an attributed network data, having collected the attributes of vertices for each pair of individuals *i* and *j* at time stamp *t* the $p \times 1$ vector of similarities between these two individuals would be $\mathbf{x}_{ijt} = \{x_{ijt1}, x_{ijt2}, ..., x_{ijtp}\}$.

In this vector, x_{ijtp} for $p = 1, 2, 3, ..., p$ shows that how individuals *i* and *j* are similar to each other being compared considering the *pth* attribute at time stamp *t*. In social network data analysis, the common attributes of the interest are categories such as the sex, position, place or age and etc., and there should be reference criteria of how to compare individuals to determine x_{ijtp}. For instance, if the *pth* attribute is the gender, in case that individuals *i* and *j* are both the same sex, $x_{ijtp} = 1$ 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 mostly these assumptions do not necessarily hold. As it is shown in Figure 1, in many cases we may have access to only some of the randomly selected nodes and we have to analyse data based on the available set of vertices as shown with red dots in Figure 2.

Insert Figure 2 about here

3.2. Random effects count data modelling and parameters estimation

The very 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 by the parameter of $\mu_{kt} = \alpha_k \lambda_{kt}$. Note that the index of *k* is equivalent to the *kth* pair of individuals. In other words, if there are *n* individuals in a network, there would be $\binom{n}{2}$ pairs of individuals each denoted by a number from the index of *k*.

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

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

The regression analysis of random effects count data is 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 $\lambda_{kt}=\exp(\mathbf{x}_{kt}^{'}\boldsymbol{\beta})$ can be obtained from Equation (5) as

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

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

4. Monitoring procedure

In this section, we propose a LRT based procedure to monitor the dynamic network considering static and dynamic reference sets 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 computed as follows:

The parameters of the count model shown in Equation (3) lead to the log likelihood value of l_{R_0} represented in Equation (6),

$$l_{R_0}=\log\{\prod_{t\in R_0}\prod_{i=1}^{v}\prod_{j\neq i}\frac{[\exp((\mathbf{x}_{ijt}^{'}\boldsymbol{\beta}_{R_0}+\delta_{R_0 ijt})]^{y_{ijt}}.[\exp(-(\mathbf{x}_{ijt}^{'}\boldsymbol{\beta}_{R_0}+\delta_{R_0 ijt}))]}{y_{ijt}!}\}. \quad (6)$$

At time stamp τ, the value of the log likelihood function is obtained using Equation (8) as follows:

$$l_{R_\tau} = \log\{\prod_{t\in\tau}\prod_{i=1}^{v}\prod_{j\neq i}\frac{[\exp((\mathbf{x}'_{ijt}\boldsymbol{\beta}_{R_\tau}+\delta_{R_\tau ijt})]^{y_{ijt}}.[\exp(-(\mathbf{x}'_{ijt}\boldsymbol{\beta}_{R_\tau}+\delta_{R_\tau ijt}))]}{y_{ijt}!}\}, \quad (7)$$

Similarly, $R_1 = R_0 \cup \tau$, then l_{R_1} would be obtained using Equation (8).

$$l_{R_1} = \log\{\prod_{t\in R_0}\prod_{i=1}^{v}\prod_{j\neq i}\frac{[\exp((\mathbf{x}'_{ijt}\boldsymbol{\beta}_{R_0}+\delta_{R_0 ijt})]^{y_{ijt}}.[\exp(-(\mathbf{x}'_{ijt}\boldsymbol{\beta}_{R_0}+\delta_{R_0 ijt}))]}{y_{ijt}!}$$
$$\prod_{t\in\tau}\prod_{i=1}^{v}\prod_{j\neq i}\frac{[\exp((\mathbf{x}'_{ijt}\boldsymbol{\beta}_{R_\tau}+\delta_{R_\tau ijt})]^{y_{ijt}}.[\exp(-(\mathbf{x}'_{ijt}\boldsymbol{\beta}_{R_\tau}+\delta_{R_\tau ijt}))]}{y_{ijt}!}\}. \quad (8)$$

When the network is in-control, the model parameters of the network at time stamp τ are not significantly different from that of from 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 by using Equation (9),

$$l_0 = \log\{\prod_{t\in R_0}\prod_{i=1}^{v}\prod_{j\neq i}\frac{[\exp((\mathbf{x}'_{ijt}\boldsymbol{\beta}_0+\delta_{0ijt})]^{y_{ijt}}.[\exp(-(\mathbf{x}'_{ijt}\boldsymbol{\beta}_0+\delta_{0ijt}))]}{y_{ijt}!}$$
$$\prod_{t\in\tau}\prod_{i=1}^{v}\prod_{j\neq i}\frac{[\exp((\mathbf{x}'_{ijt}\boldsymbol{\beta}_0+\delta_{0ijt})]^{y_{ijt}}.[\exp(-(\mathbf{x}'_{ijt}\boldsymbol{\beta}_0+\delta_{0ijt}))]}{y_{ijt}!}\}. \quad (9)$$

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 by using Equation (10) as follows:

$$l_1 = \log\{\prod_{t\in R_0}\prod_{i=1}^{v}\prod_{j\neq i}\frac{[\exp((\mathbf{x}'_{ijt}\boldsymbol{\beta}_0+\delta_{0ijt})]^{y_{ijt}}.[\exp(-(\mathbf{x}'_{ijt}\boldsymbol{\beta}_0+\delta_{0ijt}))]}{y_{ijt}!}$$
$$\prod_{t\in\tau}\prod_{i=1}^{v}\prod_{j\neq i}\frac{[\exp((\mathbf{x}'_{ijt}\boldsymbol{\beta}_\tau+\delta_{\tau ijt})]^{y_{ijt}}.[\exp(-(\mathbf{x}'_{ijt}\boldsymbol{\beta}_\tau+\delta_{\tau ijt}))]}{y_{ijt}!}\}. \quad (10)$$

In Equations (9), (10) and (11) the model parameters $\boldsymbol{\beta}_{R_0}$, $\boldsymbol{\beta}_{R_\tau}$ and $\boldsymbol{\beta}_0$ are unknown and they should be estimated using Equation (5), however software packages such as MATLAB can be used for this regard. Finally, the LRT-based statistic is computed using Equation (11),

$$\Lambda(\tau) = 2(l_1 - l_0). \quad (11)$$

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 to 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 by using simulation studies considering static and dynamic reference sets. For this aim, we first define a model for data generation. This model generates simulated data, however for a monitoring procedure we need to know how to monitor the data considering a determined control limit. To have an in-control ARL value equal to 200 for the both of dynamic and static reference sets, the corresponding UCLs are determined by using 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, while the designed control chart detects shifts faster, it means 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 that random effect model proposes for monitoring social networks, is that the covariates vector may change over time. In other words, the mean of counts of links between individual *i* and *j*, λ_{ijt}, is modelled in terms of the similarity vector of individual *i* and *j* at time *t*. Hence, this model enables us to model the networks in which the attributes of nodes change in different time stamps. The below 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,

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

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

5.2.Simulation results for the fixed covariate model

Considering the aforementioned model properties for the fixed covariate model, the UCL of the static and the 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.

Insert Figure 3 about here.

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.

Insert Figure 4 about here.

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 illustrates a similar performance that confirms the effectiveness of the method as shown in Figure 5.

Insert Figure 5 about here.

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

5.3. 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 UCL of the dynamic and the 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, 7 and 8 show 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 in comparison with the static approaches.

Insert Figure 6 about here.

Insert Figure 7 about here.

Insert Figure 8 about here.

5.4. An Illustrative Example

Consider an example from an online market network of digital products where different firms can register to have a profile in this site to present their product and search the profiles of other firms to make deals together. Since the connections take place within a site, and each firm is recognized by its profile, the corresponding attributes can be obtained by using profile information. Hence, each firm as a node; can have several attributes such as the number of previous sales in the site, number of physical branches and the rank of the firms in the market. Due to the policies of the owner of the site, the example under study is discussed anonymously.

Insert Figure 9 about here.

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 in each time stamp, the value of the trade or any other type of links as edges. For applying the proposed monitoring method on an illustrative example, one needs to estimate the model parameters and determine the corresponding control limit. In the example under study, applying the data set of a 50-week time horizon the model parameters are estimated in the first step model as,

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

In this model, the attributes can be properties such as the number of previous sales in the site, number of physical branches and the rank of the firms noted 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 $\mathbf{x}_{ijt}$ is determined. it is noted that the considered attributes follow discrete uniform distribution of [1-20], [1-3] and [1-4] respectively.

In the next step, the LRT based procedure given by Equation (12) 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

used 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 Equation (16), the UCL equals to 4.4 which is determined by applying 10000 simulation runs satisfying an in-control ARL equal to 200. For verification of the proposed method, a 50-week time horizon of the data set was used to evaluate the performance of the proposed method and the corresponding control chart is shown in Figure (10) as follows.

Insert Figure 10 about here.

As it is shown in Figure 8, results of shifts in the values of the first coefficient from 0.05 to 0.08 from the week 26 forward, has led to significant change of statistic value and it presented an out-of-control state from then, the evident that can approve 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 considering random effects concept. The applied modelling enables the monitoring procedure to be capable for detecting structural shifts in both of the networks with fixed set of individuals and the variable attribute ones. Moreover, the incorporation of random effect concepts to the model would improve the applicability of the monitoring procedure for the networks with variable covariates. The performance of the proposed method was evaluated by using in terms of ARL. Due to the improvements that the proposed method suggests, more studies considering network data with other distributions of ties such as ordinal data can be recommended for future research studies.

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., Glance, N. and Hurst, M., "Patterns of cascading behaviour in large blog graphs", *Proceedings of the 2007 SIAM International conerence. 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., August. "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., Roy, S. and Agarwal, A., January, "Extracting dense communities from telecom call graphs", In *Communication Systems Software and Middleware and Workshops, 2008. COMSWARE 2008. 3rd International Conference*, pp. 82-89. IEEE, (2008).

[7] Savage, D., Zhang, X., Yu, X., Chou, P. and Wang, Q., "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 Military Academy", West Point NY Network Science Center (NSC), (2011).

[10] Miller, B.A., Arcelano, N., and , 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. and Dodds, P.S, "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., Sparks, R. and Wilson, J.D., "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., and Hand, D. J, "Bayesian anomaly detection methods for social networks", *Ann. Appl. Stat*. 4(2), (2010).

[14] Priebe, C.E., Conroy, J.M., Marchette, D.J. and Park, Y, "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., Fisk, M., and Storlie, C.B., "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., and Runger, G., "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. and Rahimian, G., "A statistical approach to social network monitoring", *Commun Stat-Theor M*, 46(22): pp. 11272-11288, (2017).

[23] Fotuhi H, Amiri A, Maleki MR. "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., Peynabar, K., "Change Detection in a Dynamic Stream of Attributed Networks", arXiv preprint arXiv:1711.04441 (2018).

[25] Woodall WH, Zhao MJ, Paynabar K, Sparks R, Wilson JD. "An overview and perspective on social network monitoring", *IEEE Trans. Ind. Appl.* 2017; 49(3):354- 365.

[26] Sengupta, S, Woodall, WH. "Discussion of Statistical methods for network surveillance", *Appl Stoch Model Bus*, 34(4): pp. 446-448, (2018).

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

[28] Hosseini SS, 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., & 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., & Woodall, W. H., "Statistical evaluation of spectral methods for anomaly detection in networks", arXiv preprint arXiv:1711.01378, (2017).

[32] Mazrae Farahani E, Baradaran Kazemzade R, Albadvi A, Teimourpour B. "Modeling and monitoring social Network in term of longitudinal data", *Int. J. Ind. Eng. Comput.*, 29 (3): pp. 247-259, (2018).

[33] Hatef Fotuhi, Amirhossein Amiri & Ali Reza Taheriyoun A novel approach based on multiple correspondence analysis for monitoring social networks with categorical attributed data, *J Stat Comput Sim*, 89(16), 3137-3164, (2019).

[34] Mogouie, H., Raissi-Ardali, G. A., Bahrami-Samani, E., Amiri A., Statistical monitoring of binary response attributed social networks considering random effects, Published online in Commun Stat-Sim C, (2019). doi: 10.1080/03610918.2019.1661471.

[35] Najafi, H., Saghaei, A., Statistical monitoring for change detection of interactions between nodes in networks: with a case study in financial interactions network, Published online in *Commun Stat-Theor M*, (2020). doi: 10.1080/03610926.2020.1725830.

[36] Noorossana R, Hosseini SS, 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., and Robinson, T.J., "Generalized linear models: with applications in engineering and the sciences", John Wiley & Sons, (2012).

[38] Agresti, A., "Categorical data analysis", John Wiley & Sons, (2013).

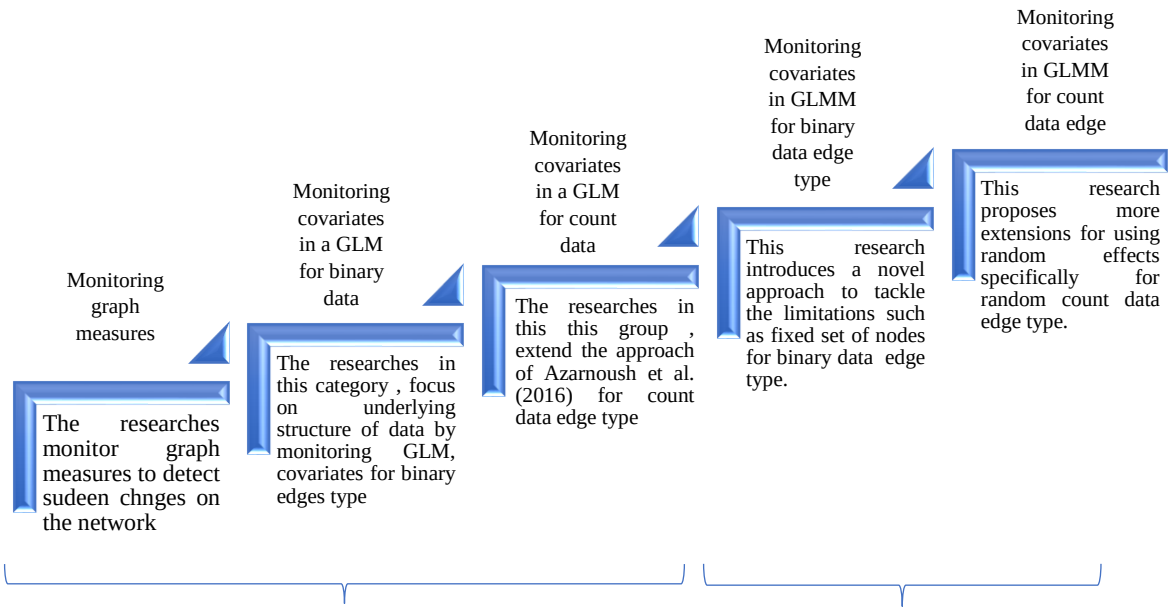
[39] Cameron, A.C. and Trivedi, P.K., "Regression analysis of count data", (Vol. 53). Cambridge university press, (2013).

Hamed Mogouie is a Ph.D. candidate in Industrial Engineering at Isfahan University of Technology in Iran. His field of research is 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 including founding educational institutes, reengineering large scale industries, consulting in different public and private sectors as well as academic activities in the area of Total Quality Management.

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 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 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.

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.



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 more recent literature

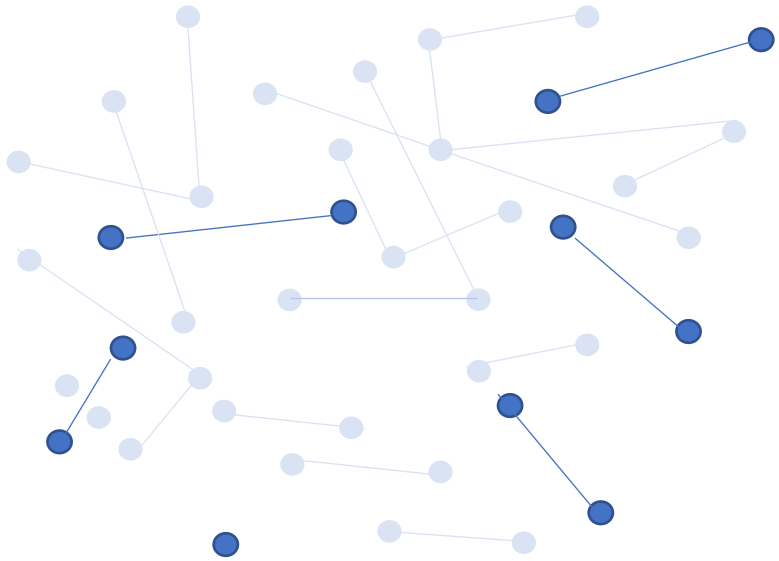


Figure 2. A sample network with randomly accessible nodes

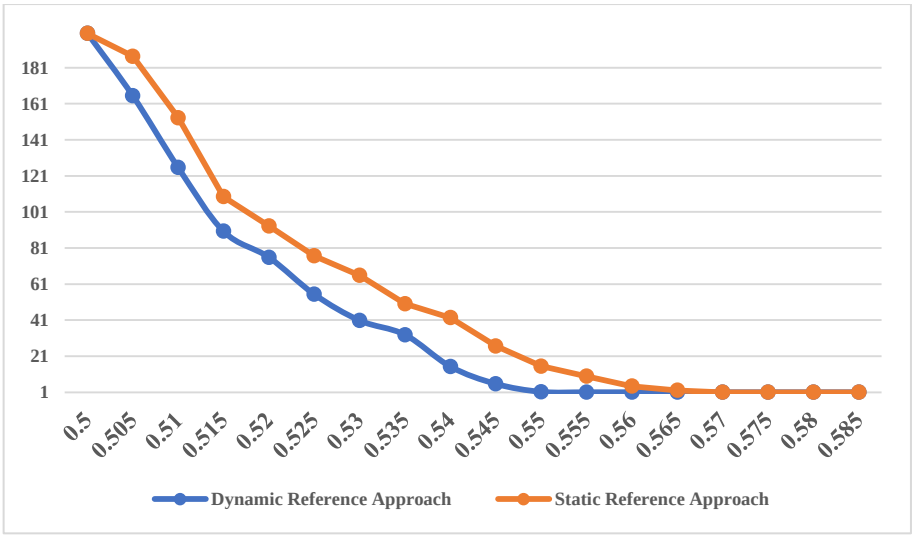


Figure 3- ARL curves under different shifts in β_1 (Fixed covariate model) considering static and dynamic reference set approaches

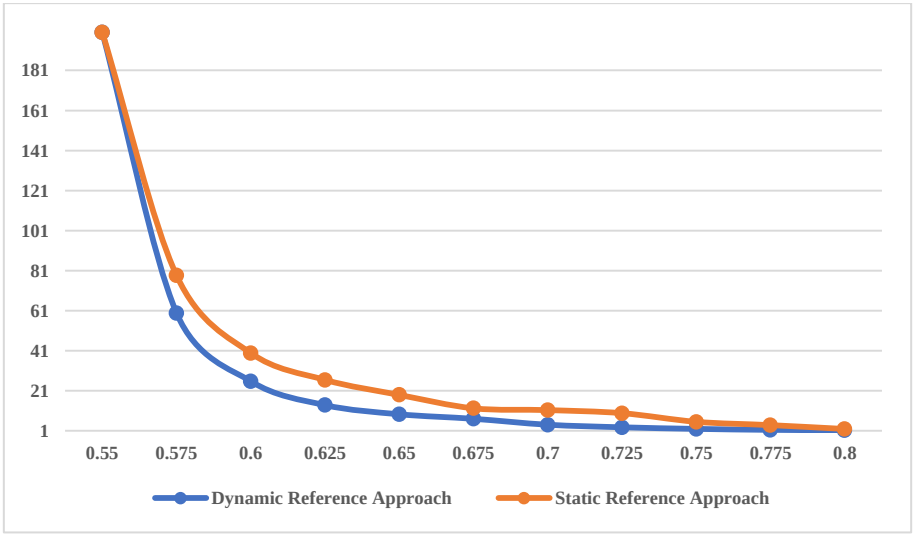


Figure 4- ARL curves under different shifts in β_2 (Fixed covariate model) considering static and dynamic reference set approaches

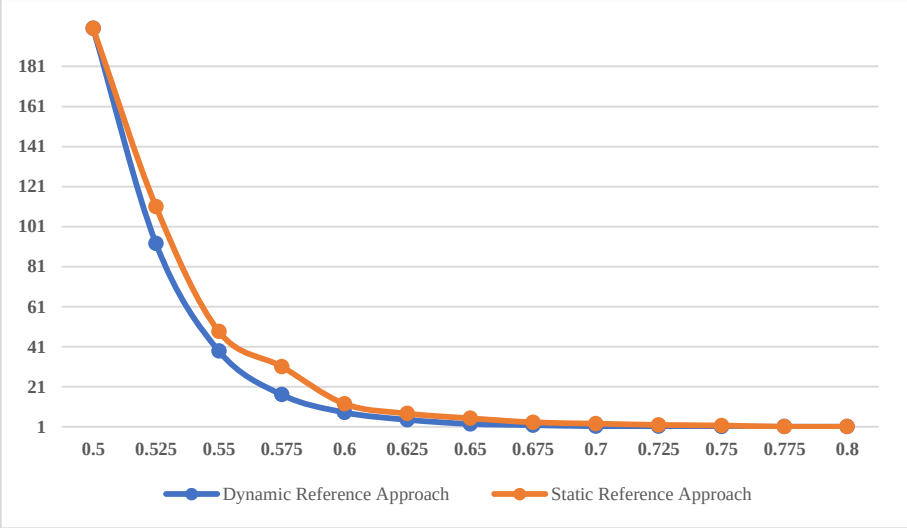


Figure 5- ARL curves under different shifts in β_3 (Fixed covariate model) considering static and dynamic reference set approaches

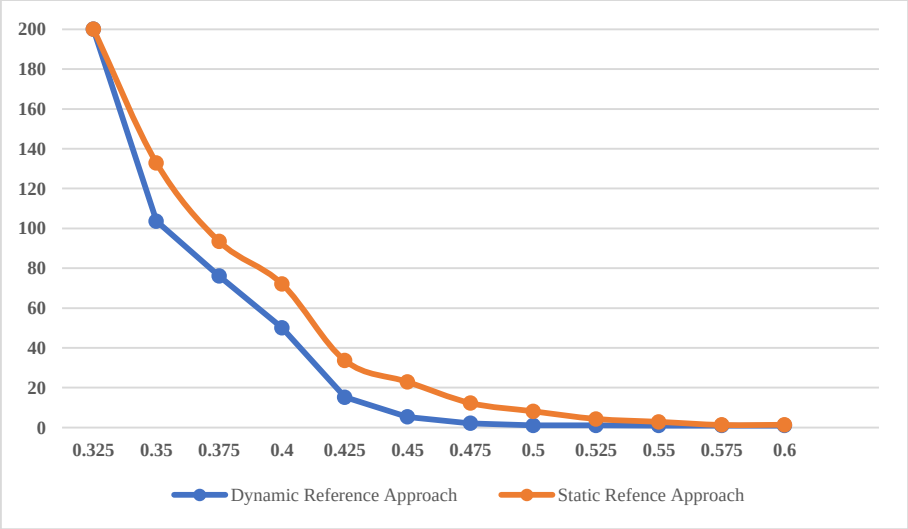


Figure 6- ARL curves under different shifts in β_1 (Variable covariate model) under static and dynamic reference set approaches

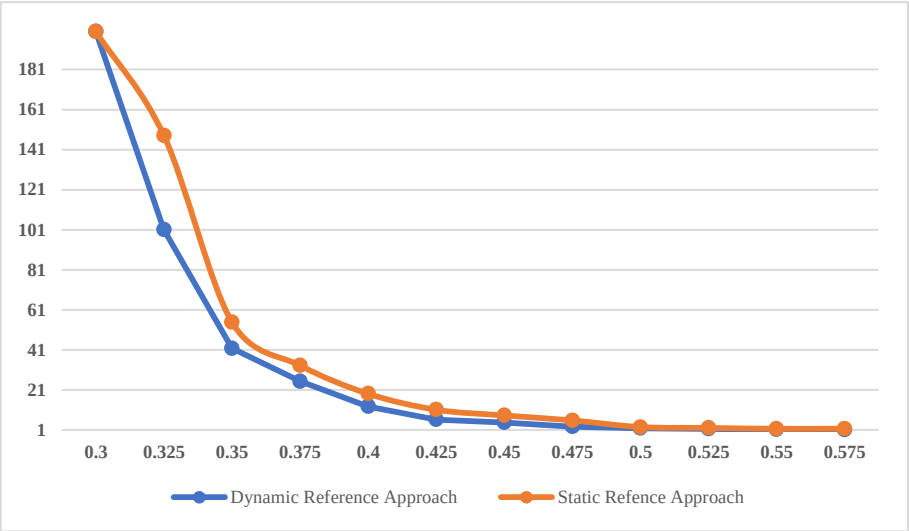


Figure 7- ARL curves under different shifts in β_2 (Variable covariate model) under static and dynamic reference set approaches

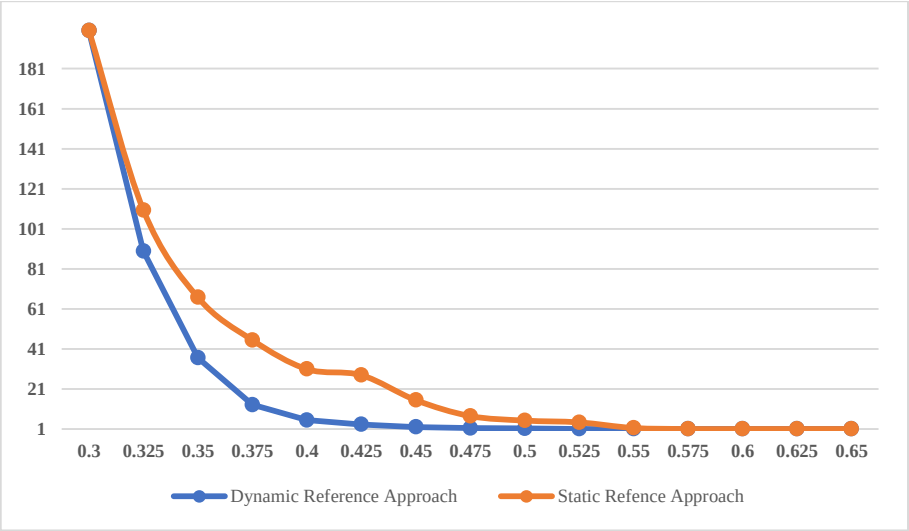


Figure 8- ARL curves under different shifts in β_3 (Variable covariate model) considering static and dynamic reference set approaches

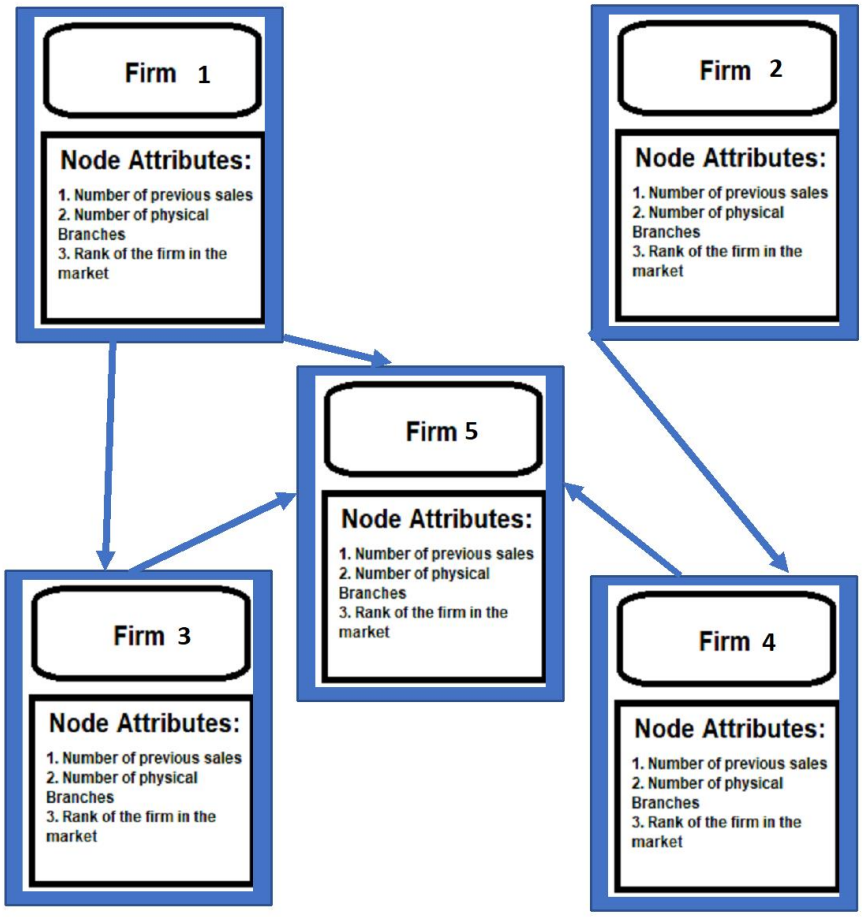


Figure 9 - An illustration of the firms' network

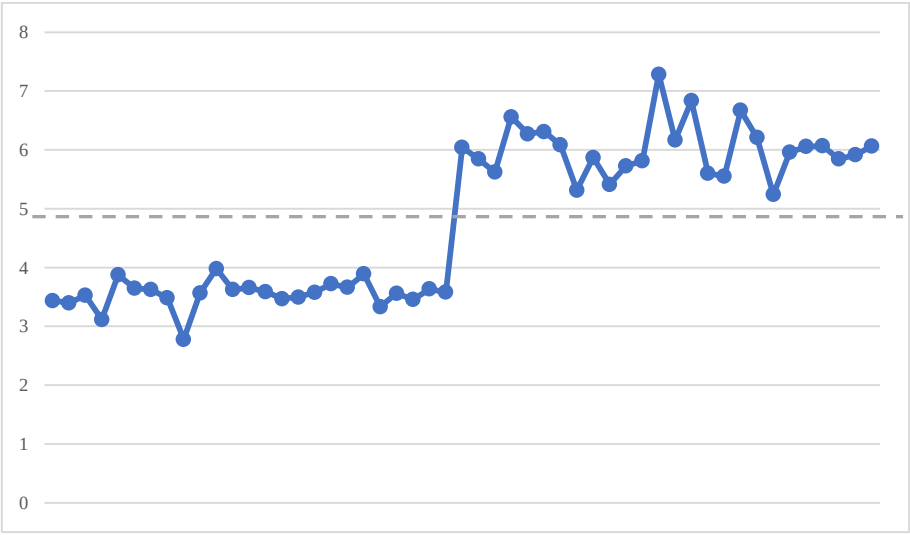


Figure 10. LRT based control chart under the dynamic reference set approach