International Journal of Industrial Engineering & Production Research (2018)

September 2018, Volume 29, Number 3
pp. 247- 259

DOI: 10.22068/ijiepr.29. 3.247 http://IJIEPR.iust.ac.ir/

GLMM-Based Modeling and Monitoring Dynamic Social Network

Ebrahim Mazrae Farahani, Reza Baradaran Kazemzadeh*, Amir Albadvi & Babak Teimourpour

Ebrahim Mazrae Farahani, *Industrial Engineering Department, Tarbiat Modares University, Tehran, Iran*
Reza Baradaran Kazemzadeh, *Industrial Engineering Department, Tarbiat Modares University, Tehran, Iran*
Amir Albadvi, *Industrial Engineering Department, Tarbiat Modares University, Tehran, Iran*
Babak Teimourpour, *Industrial Engineering Department, Tarbiat Modares University, Tehran, Iran*

KEYWORDS

Social network monitoring;
Generalized Linear Mixed Models;
Likelihood ratio test (LRT);
Average Run Length (ARL).

ABSTRACT

Social network monitoring (SNM) can play a significant role in everyone's life. Recent studies show the importance and increasing interests in the subject by modeling and monitoring the communications between network members over time by treating the collected observations as longitudinal data. Typically, the tendency for modeling social networks, considering the dependency of an outcome variable on the covariates, is growing recently. However, these studies fail to incorporate the possible correlation between responses in the proposed models. In this paper, we use generalized linear mixed models (GLMMs), also referred to as random effects models, to model a social network according to the attributes of nodes in which the nodes take a role of random effect or hidden effect in modeling. In order to estimate the regression parameters, Monte Carlo expectation maximization (MCEM) algorithm is used to maximize the likelihood function. In our simulation studies, we applied root mean square error (RMSE) and standard deviation criteria to select an appropriate model for the simulated data. Results indicate zero inflated Poisson mixed as an appropriate model for the data. In addition, compared to the other studies, our simulation study demonstrates an improvement in the average run length (ARL).

1. Introduction

In general, social networks represent the patterns of ties between social actors in which ties or relationships can be elucidated as exchanging valuable items [1], and social network analysis(SNA) is the method for studying social networks analytically, which is based on the identification of characteristics of individuals and constructed communities in networks, modeling network behaviors, and predicting relationships between members[2]. To control and test networks homogeneity, Azarnoush et al. [3] considered two major categories of homogeneity, namely static homogeneity and temporal homogeneity. Test of static homogeneity attempts to identify networks which are comprised of relationships or edges that are anomalous to the rest of current network. The purpose of testing temporal homogeneity is to diagnose the period of time when the network is composed of different structures and edges due to specific reasons in comparison to previous networks when normal relationships or edges were in place.

* **Corresponding author:** ***Reza Baradaran Kazemzadeh***
Email: rkazem@modares.ac.ir
Received 26 January 2018; revised 19 May 2018; accepted 5 August 2018

Due to the nature of social networks, external attributes can affect the formation process of relationships between people. In this regard, Miller et al. [16], Azarnoush et al. [3], and Mazrae Farahani et al. [17] monitored social networks considering these attributes. Azarnoush et al. [3] used logistic regression to model the probability of edge formation between vertices by considering attribute variables such as gender and age. Then, they monitored the coefficients of regression model in order to detect anomalies in phase II and demonstrated the ability of the proposed method through detecting anomalies in Enron network. The type of regression model that they used is known as generalized linear model (GLM). Woodall et al. [18] provided an extensive literature review of social network monitoring with emphasis on the statistical methods such as control charts and hypothesis testing's, Bayesian analysis, scan statistics, and time series.

Mazrae Farahani et al. [17] modeled a social network using Poisson regression model and used MEWMA and MCUSUM control charts to monitor the average degree, average betweenness, and average closeness measurements simultaneously. Sparks and Wilson [19] proposed the neighborhood-based search method to find the unknown candidate team in the network. Their monitoring plan covers collaborative teams, teams with a dominant leader, and global outbreak of communications. First, they used a multivariate method to smooth the communication counts and compared total number of communications in a team with the expected mean of communications in the same team to detect anomalies. Wilson et al. [20] used dynamic version of the degree-corrected stochastic block model (DCSBM) in order to model network. From the perspective of block model, the network includes several communities in which communications within them are denser than those between them. They monitored three parameters belonging to the mentioned model using Shewhart control charts. These parameters are probability of nodes belonging to a specific community, probability of connection between and within different communities, and tendency of nodes to make connection. Zou and Li [21] proposed network state space model (NSSM) to describe the natural evolution of dynamic networks and, then, used a singular value decomposition (SVD) based method to integrate NSSM and statistical process control (SPC) to detect changes. Mazrae Farahani et al. [17] modeled the baseline periods of a social network using the probability density profile (PDP) method. Then, they applied Poisson regression to monitor a social network in phase I.

Savage et al.[13] described four kinds of anomalies in online social networks: static labeled, static unlabeled, dynamic labeled, and dynamic unlabeled. In addition, they provided an overview of existing methods for detecting the aforementioned anomalies. Sparks [14,15] monitored the departure of smoothed communications level from their expected mean and expected median, respectively. The first study used multivariate EWMA control chart and the second applied adaptive CUSUM chart.

In order to model a social network, we intend to consider the hidden effect that each node has in the networks. This hidden effect, which leads to a correlation structure among response variables at different time steps, is modeled using generalized linear mixed model (GLMM). Generalized linear mixed models are the extensions of generalized linear models and include both fixed and random effects and are often used in order to describe longitudinal data or correlated data. Longitudinal data arise when the same sample is tracked in different time periods. Longitudinal data allow for the measurement of within-sample changes over time. For more information on the analysis of longitudinal data using random effects model, see Laird and Ware [22].

In social science literature, some researchers, including Katz and Proctor [5] and Wasserman [6], studied social network data over time. Moreover, in the realm of longitudinal social networks, various studies were done by researchers including Sampson[7], Newcomb[8], Frank [4],McCulloh et al.[9,10], Wasserman and Faust [2], and Stokman and Doreian [11]

After modeling social network with the mentioned GLMM models, we need to estimate their parameters. Many researchers, including Davidian and Giltinan [24], McCulloch [25], and McCulloch et al. [26], used the method of maximum likelihood estimation to estimate parameters of the aforementioned GLMM models. It is common to use numerical methods to maximize the likelihood function, and we took a approach similar to that of McCulloch et al. [12,26] which used MCEM to estimate the model parameters. Next, we used likelihood ratio test (LRT) to monitor changes in a social network in phase II. Many researchers, including Sullivan and Woodall [30], Paynabar et al.[31], Zou and Tsung [35], and Azarnoush et al.[3], used the same method for monitoring purpose.

The third section is devoted to our simulation studies to compare the performance of the proposed method with those of the previous methods using ARL. Our concluding remarks are provided in the final section.

2. Proposed Methodology

2-1. Modeling the social network:

In this section, we intend to describe the details of the proposed method to monitor and detect abnormal behaviors among actors in the social network streams. Our intended method incorporates the hidden effect of each node that has influence on making communication. We used GLMMs model and applied the longitudinal concept to the response variable, which can have the Binary, Poisson, and zero inflated Poisson (ZIP) distributions. We defined $G(t)=(V(t),Y(t)),\ t=1,2,..,m\ (m \geq 2)$ as the indicator of the network at time t. This network is composed of individuals as nodes $V(t)$ and edge $Y(t)$ as the communication between the nodes. Indeed, $V(t)$ is a subset of nodes, i.e., $V(t) = V = \{ v_1, v_2, ..., v_i, ..., v_s \}$ where v_i is the node or actor of the network, such that s is the number of all nodes which we assume this number is fixed in the network streams. In addition, $Y(t) = \{ y_{12}(t), ..., y_{ij}(t), ..., y_{(s-1)s} \}$ in which $y_{ij}(t)$ is the communication or the edge between nodes i and j at time t where i and j have the range from 1 to s and, for $i \neq j$, communication $y_{ij}(t)$ is equivalent to $y_{ji}(t)$. The observed network is represented as a graph with the adjacency matrix and, for each time t, $y_{ij}(t)$ is the component of this matrix and its diagonal, i.e. s, for all y_{ii}, is defined to be 0. As mentioned previously, we assume that, in the social network, $y_{ij}(t)$ can have different distributions: the Binary, Poisson, and zero inflated Poisson (ZIP). For instance, if $y_{ij}(t)$ has Bernoulli distribution, it takes two possible values 1 or 0, and in the existing communication mode between nodes i and j, it takes 1, otherwise accepts 0; if $y_{ij}(t)$ follows Poisson distribution, it indicates the number of communications between nodes i and j at time t. At each time snapshot, $y_{ij}(t)$ as a response variable which denotes connection between nodes i and j at time $t$is modeled through "Generalized Linear Mixed Model" as follows:

$$Y_{ij}(t) = \mathbf{X}_{ij}^T \boldsymbol{\beta}(t) + u_{ij} + \varepsilon_{ij}. \quad (1)$$

It is noteworthy that the type of regression model that should be selected depends on the kind of the response variable. For example, if the response variable is a Binary, the Logistic Regression model is usually used with the random error; likewise, for other distributions, the respective model should be implemented.

In Equation (1), $\mathbf{x_{ij}} = (x_{1ij}, ..., x_{pij})$ is the p-vector of covariates related to the fixed effects. In other words, in the realm of the social network is a vector of associated attributes; then, x_{pij} is the p^{th} attribute of the edge between nodes i and j. Al Hasan et al. [27] presented an overview of mentioned attributes. For instance, in the email communication network related to a university, age difference of the two users, having common major and the number of courses that two users are enrolled in, are effective attributes. Moreover, u_{ij} is defined as the random effect associated with pair nodes i and j which, as prevalently assumed, has normal distribution with mean 0 and variance D. Finally, ε_{ij} represents the random errors of repeated measurement of the communication within actors (i, j) which follows the normal distribution with mean 0 and variance σ^2. Note that ε_{ij} and u_{ij} are independent variables.

Now, one important question happens after representing the GLMM model like Eq. (1). The question is that "which one of the models GLMM or GLM is better fitted for social network data?''. In other words, "Does the hidden or random effect of nodes has an important role in the nature of the social network data?" To answer this question, we used likelihood ratio test (LRT). For acquiring this purpose, it is assumed that model I is a regression model with not random effect element (GLM), and model II encompasses this element (GLMM). Then, we designed one statistical hypothesis testing as the following procedure:

$$\begin{aligned} H_0 &= \text{Model I and Model II fit equally} \\ H_1 &= \text{model I and model II dont fit equally.} \end{aligned} \quad (2)$$

For testing this hypothesis, the LRT statistic is defined as follows:

$$T = -2\log(\frac{L_1(\hat{\varphi}_1)}{L_2(\hat{\varphi}_2)}), \quad (3)$$

where L_1 and L_2 are the likelihoods while $\hat{\varphi}_1$ and $\hat{\varphi}_2$ are MLEs of parameter φ related to models I and II, respectively. The LRT statistic T asymptotically follows χ^2_r distribution in which r is the degree of freedom equal to the difference in the number of parameters being in two models [28]. If the null hypothesis is rejected, it means that the variance of random effect has positive value; then, GLMM model is better than GLM for describing the social network data.

Generally, by assuming that the expectation value of $y_{ij}(t)$ in Eq.(1) is given as $E(y_{ij}(t) \mid u_{ij}) = \mu_{ij}(t)$, we can rewrite the regression model in another form with link function g and η as the linear predictor:

$$g(\mu_{ij}(t)) = \eta_{ij}(t) = \mathbf{X}_{ij}\boldsymbol{\beta}_{ij}(t) + u_{ij}. \quad (4)$$

In addition, the GLMMs model can cover the exponential family such as the Binomial, Poisson, and zero inflated Poisson distributions. Therefore, the general probability density function for y_{ij} is defined as

$$f(y_{ij} \mid \beta, D, \phi) = \exp\{\frac{y_{ij}\eta_{ij} - c(\eta_{ij})}{a_{ij}(\phi)} + d_{ij}(y_{ij}, \phi)\}, \quad (5)$$

where $a_{ij}(0), c(0), d_{ij}(0,0)$ are known functions, and ϕ is a dispersion parameter, which may or may not be known [29]. For instance, we present these functions and parameters for two distributions in Table (1) as given in [32].

For the estimation of parameters β and D in Eq. (1), we can implement the maximum likelihood estimation method for $t = 1, 2, \ldots, m$ as follows:

$$f_{ij}(y_{ij} \mid \beta, D, \phi) = \int \prod_{t=1}^{m} f_{ijt}(y_{ij}(t) \mid u_{ij}, \beta(t), \phi) f(u_{ij} \mid D) du_{ij} \quad (6)$$

Then, the likelihood function is

$$L(\beta, D, f) = \prod_i \prod_j f_{ij}(y_{ij} \mid \beta, D, f) = \prod_i \prod_j \int \prod_{t=1}^{m} f_{ijt}(y_{ij}(t) \mid u_{ij}, \beta(t), f) f(u_{ij} \mid D) du_{ij} \quad (7)$$

Because of the complicated calculations in the maximization of the likelihood function for the

Tab. 1. The value of parameters and functions which are used for the Binomial, and Poisson distributions

Distribution	η	ϕ	$a(\phi)$	$c(\eta)$	$d(y,\phi)$
Binomial (n,μ)	$\log(\frac{\mu}{1-\mu})$	1	1	$-n\log(1-\mu)$	$\log\binom{n}{y}$
Poisson (μ)	$\log(\mu)$	1	1	e^{η}	$-\log(y!)$

estimation of the parameters in Eq. (7). It is common to use numerical iterative methods. Here, we applied the Monte Carlo-Expectation Maximization (MCEM) and to prevent the troublesome calculations of this method, we refer readers to [32] to discuss the numerical methods for GLMMs model.

2-2. Monitoring the social network:

In this section, we applied the monitoring process for social network. Due to the similarity of procedure of monitoring the social network for different distributions of $y_{ij}(t)$, we just mentioned this method and its calculations for Bernoulli distribution. In the situation where $y_{ij}(t)$ has the Bernoulli distribution, we can use the logistic regression to model the probability of the communication between the pair of nodes at time t. Thus, in Eq. (4), link function g is logit as follows:

$$g(\mu_{ij}(t)) = \eta_{ij}(t) = \log it(\mu_{ij}(t)) = \boldsymbol{X}_{ij}\boldsymbol{\beta}(t) + u_{ij}. \quad (6)$$

Then, we have

$$\mu_{ij}(t) = P(y_{ij}(t) = 1 \mid \boldsymbol{X}_{ij}, \boldsymbol{\beta}(t), u_{ij}) = \log it^{-1}(\boldsymbol{X}_{ij}\boldsymbol{\beta}(t) + u_{ij}), \quad (7)$$

where $\mu_{ij}(t)$ is the probability of the communication between nodes i and j at time t that has a direct relationship with the attribute vector and the random effect. We assume that the random effect has lognormal distribution with scalar variance *D*. The probability density function for the random effect with parameters l, k as mentioned is estimated by MCEM methods.

In the network streams, reference network R is a set of networks $\{G(t), t \in R\}$ with size *q,* which are collected during the normal mode and without any out-of-control situation, and is considered as a criterion for comparing the current network streams. In order to test whether the mechanism that forms the structure of incoming current network $G(\tau)$ with the time parameter $\tau = 1, 2, \ldots m$ is the same as the reference networks, we compared the current incoming network with reference network. If we suppose that mechanisms generating the current incoming network $G(\tau)$ have different structures in comparison with the reference set, then we can write

$$\mu_{ij}(t) = \begin{cases} \log it^{-1}(\boldsymbol{X}_{ij}\boldsymbol{\beta}^0(t) + u_{ij}) \ for \ t \in R, \\ \log it^{-1}(\boldsymbol{X}_{ij}\boldsymbol{\beta}^1(t) + u_{ij}) \ for \ t \in \tau, \end{cases} \quad (8)$$

where $\boldsymbol{\beta}^0$ (with element β_p^0) and $\boldsymbol{\beta}^1$ (with element β_p^1) are the vectors of the regression coefficients for the reference sets and current incoming network, respectively. In addition, u_{ij} is the random effect with the element of variance D. To investigate the possibility of any change, we should test

$$H_0 : \boldsymbol{\beta}^1 = \boldsymbol{\beta}^0 \quad H_1 : \boldsymbol{\beta}^1 \neq \boldsymbol{\beta}^0. \quad (9)$$

in which, to this end, we used the likelihood-ratio test (LRT). Many studies in different fields for the monitoring purpose have used LRT-based method such as [31], [32], [33] and [3]. If we assume that $P(0)$ indicates the Bernoulli probability mass function, the log-likelihood under the alternative is as follows:

$$\begin{aligned} l_1 &= \log\{(\prod_i \prod_j \int \prod_{t=1}^{R} P(y_{ij}(t); \boldsymbol{X}_{ij}, \boldsymbol{\beta}^0, u_{ij}) \\ &\times (\prod_i \prod_j \int P(y_{ij}(\tau); \boldsymbol{X}_{ij}, \boldsymbol{\beta}^1, u_{ij})\} = \\ &\log\{(\prod_i \prod_j \int \prod_{t=1}^{R} (\mu_{ij}^0(t))^{y_{ij}(t)} (1-\mu_{ij}^0(t))^{1-y_{ij}(t)} f(u_{ij}|D) du_{ij}) \\ &\times (\prod_i \prod_j \int (\mu_{ij}^1(\tau))^{y_{ij}(\tau)} (1-\mu_{ij}^1(\tau))^{1-y_{ij}(t)} f(u_{ij}|D) du_{ij})\} \\ &= \sum_i \sum_j \log(\int \prod_{t=1}^{R} (\mu_{ij}^0(t))^{y_{ij}(t)} (1-\mu_{ij}^0(t))^{1-y_{ij}(t)} f(u_{ij}|D) db_{ij}) \\ &+ \sum_i \sum_j \log(\int (\mu_{ij}^1(\tau))^{y_{ij}(\tau)} (1-\mu_{ij}^1(\tau))^{1-y_{ij}(t)} f(u_{ij}|D) du_{ij}). \end{aligned} \quad (10)$$

If we substitute vector parameters $\boldsymbol{\beta}^0, \boldsymbol{\beta}^1$ in Equation (7), quantities μ_{ij}^0 and μ_{ij}^1 as the probability of communication between nodes i and j at time t can be obtained, respectively. We denote the estimated parameters related to the reference network and the current network as $\hat{\boldsymbol{\beta}}^R, \hat{\boldsymbol{\beta}}^\tau$, respectively, each of which was obtained by the mentioned MCEM algorithm. Then, if we substitute these estimated parameters in Equation (7), $\hat{\mu}_{ij}^R$, $\hat{\mu}_{ij}^\tau$ are estimated under alternative hypothesis. In the same way, if no change occurs, we can write the log-likelihood function under the null as

$$
\begin{aligned}
l_0 &= \log\{(\prod_i \prod_j \int \prod_{t=1}^{R} P(y_{ij}(t); \boldsymbol{X}_{ij}, \boldsymbol{\beta}^0, u_{ij}) \\
&\times (\prod_i \prod_j \int P(y_{ij}(\tau); \boldsymbol{X}_{ij}, \boldsymbol{\beta}^0, u_{ij})\} \\
&= \log\{(\prod_i \prod_j \int \prod_{t=1}^{R} (\mu_{ij}^0(t))^{y_{ij}(t)} (1-\mu_{ij}^0(t))^{1-y_{ij}(t)} f(u_{ij}|D) du_{ij}) \\
&\times (\prod_i \prod_j \int (\mu_{ij}^0(\tau))^{y_{ij}(\tau)} (1-\mu_{ij}^0(\tau))^{1-y_{ij}(t)} f(u_{ij}|D) du_{ij})\} \\
&= \sum_i \sum_j \log(\int \prod_{t=1}^{R} (\mu_{ij}^0(t))^{y_{ij}(t)} (1-\mu_{ij}^0(t))^{1-y_{ij}(t)} f(u_{ij}|D) du_{ij}) \\
&+ \sum_i \sum_j \log(\int (\mu_{ij}^0(\tau))^{y_{ij}(\tau)} (1-\mu_{ij}^0(\tau))^{1-y_{ij}(\tau)} f(u_{ij}|D) du_{ij}).
\end{aligned} \tag{11}
$$

The estimated probability under null $\mu_{ij}^0(t)$, which is denoted by $\hat{\mu}_{ij}^{R'}$, is obtained by substituting the estimated $\hat{\boldsymbol{\beta}}^{R'}$, $R' = \cup(R,\tau)$ into Equation (7). If we replace the parameters with their estimations and simplify l_1, l_0, the negative of the log-likelihood ratio can be reached :

$$
\begin{aligned}
l_1 - l_0 = \sum_i \sum_j [&\log(\frac{(\int \prod_{t=1}^{R} (\hat{\mu}_{ij}^{R}(t))^{y_{ij}(t)} (1-\hat{\mu}_{ij}^{R}(t))^{1-y_{ij}(t)} f(u_{ij}|D) du_{ij})}{(\int \prod_{t=1}^{R} (\hat{\mu}_{ij}^{R}(t))^{y_{ij}(t)} (1-\hat{\mu}_{ij}^{R}(t))^{1-y_{ij}(t)} f(u_{ij}|D) du_{ij})}) \\
&+ \log(\frac{(\int (\hat{\mu}_{ij}^{\tau}(\tau))^{y_{ij}(\tau)} (1-\hat{\mu}_{ij}^{\tau}(\tau))^{1-y_{ij}(\tau)} f(u_{ij}|D) du_{ij})}{(\int (\hat{\mu}_{ij}^{R'}(\tau))^{y_{ij}(\tau)} (1-\hat{\mu}_{ij}^{R'}(\tau))^{1-y_{ij}(\tau)} f(u_{ij}|D) du_{ij})})].
\end{aligned} \tag{12}
$$

Under the null hypothesis, the approximate distribution of the LRT statistics, $\Lambda(\tau) = 2(l_1 - l_0)$ is chi-square with degrees of freedom equal to the number of coefficient and parameter in logistic regression model [34]. The value of $\Lambda(\tau)$ is calculated in order to receive network and is plotted against time to monitor changes.

3. Experimental Evaluation

This section intends to evaluate the performance of our proposed method by numerical simulation. In the first subsection, the mechanism of generating networks is described. The accuracy of parameters estimation by different models is discussed in the second subsection. In subsection three, the power of change detection in the models used in the analysis is compared. Finally, in the last subsection, power of change detection in GLMM and GLM is compared.

3-1. Generation of networks

Considering the sparse nature of social networks and also the number of communications affected by the independent variables, we generated 100 sample networks for each of m=60 time steps. Each network includes 50 vertices, and each node has two external attributes as independent

variables. We assume that x_{1ij} and x_{2ij} are defined as gender and age difference attributes, which follow Bernoulli distribution with probability equal to 0.5 and uniform (0, 30), respectively. The number of communications in the network affected by independent variable generated with inflated parameter 0.5 is zero by Eq. 14 in the following form:

$$\log(\lambda_{ij}) = 0.1 - 0.2x_{1ij} + 0.3x_{2ij} + u_{ij}, \quad (14)$$

where u_{ij} defines the random effect. We assume that the correlation between time steps is as follows:

$$R = [r_{t_1 t_2}]_{m*m},$$
$$r_{t_1 t_2} = \begin{cases} 1 & \text{if } |t_1 - t_2| = 0 \\ 0.7^{|t_1 - t_2|} & \text{if } 0 < |t_1 - t_2| \le 10 \end{cases}. \quad (13)$$

where t_1 and t_2 are defined as time steps. It is assumed that correlation disappears after ten consecutive time steps.

3-2. Comparison of the accuracy results for estimating parameters in different models

The generated dataset in Section 4.1 is modeled using three mixed models, namely zero inflated Poisson mixed model (ZIP Mixed), Poisson mixed model, and Logistic mixed model. The value of coefficients and variance of the random effect related to these three models are estimated using Monte Carlo expectation maximization (MCEM) algorithm. The plot of variance of random effect for these models is shown in Figure 1. According to this figure, variance values of random effect for all these models have significant values in various time steps, indicating the importance of random effects in network modeling. In addition, we use standard deviation, relative bias, and root-mean-square-error (RMSE) indices to compare the accuracy of estimates for the coefficients in different regression models. The results are presented in Table 3. These results indicate that value of RMSE index for ZIP mixed regression model has low deviation from the real value of coefficients. It can be deduced that due to sparse nature of the social network data, if non-inflated models are used, the accuracy of the estimation of the parameters can be decrease and affect the speed of change detection. The change detection performance for these models will be investigated in the next subsection.

3-3. Comparison of change detection performances for the models

In this section, we investigate the performance of change detection for the three models. Therefore, χ^2 control chart with upper control limit equal to $\chi^2_{4,\ .0027}$ is used. First, the ability of change detection of each model is studied by inducing a shift in the coefficients at time step 30 using

$$\log(\lambda_{ij}) = 0.1 - 0.2x_{1ij} + (0.3 + \delta_1)x_{2ij} + \delta_2(1 - x_{2ij}) + u_{ij}, \quad (13)$$

where δ_1 and δ_2 are the magnitudes of changes in the coefficients considering Table 2 for three different cases. Results shown in Figure 1 indicate that ZIP mixed model can detect the shift in the coefficients effectively. Run length (RL) criterion is computed for each scenario presented in Table 4. The results in Table 4 and Figure 2 indicate the superior performance of ZIP mixed model in comparison to the other models. Results demonstrate that the inappropriate selection of a model for social network, considering the sparse nature of its data, can affect chart performance.

Tab. 2. The magnitudes of the changes in the coefficients.

Change	δ_1	δ_2
C_0	no change	
C_1	0.3	-0.3
C_2	0.3	-0.5
C_3	0.3	0.5
C_4	0.3	0

3-4. Change detection performance for GLMM and GLM.

In this section, numerical simulation results in terms of run length are used to compare performances of the proposed general linear mixed model against general linear model. The ZIP model with a better performance in terms of run length is considered under the two model. The results in Table 5 and Figure 3 show descriptive statistics for run length. In most cases, results indicate the superiority of GLMM over GLM.

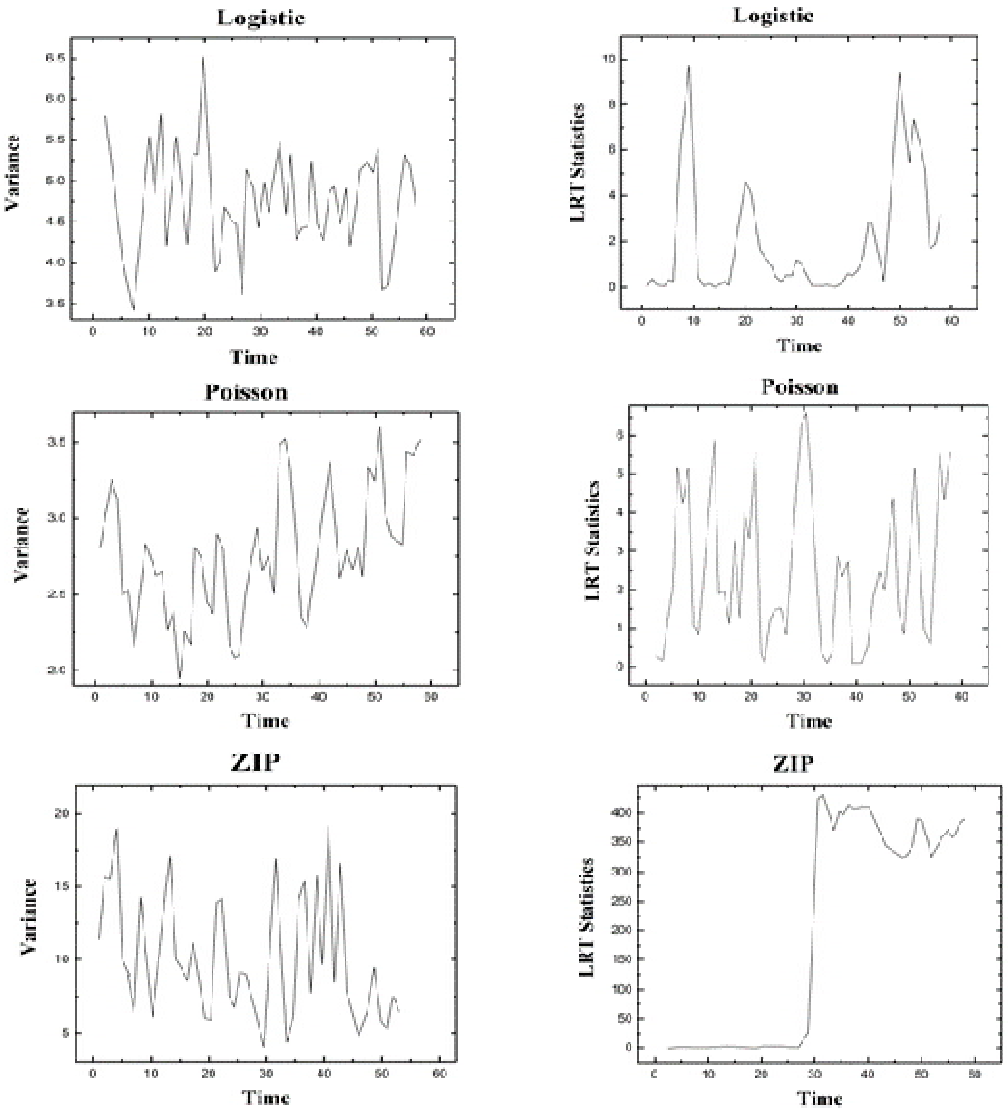

Fig. 1. The value of Variance of random effect and LRT statistics for different models

Tab. 3. Standard deviation, Standard error, Relative bias, and Root mean square error (RMSE) for estimating parameters related to each model

distribution		Beta0	Beta 1	Beta 2
	Real	0.1	-0.2	0.3
Logistic	Est	0.0064	0.0011	0.0127
	SE	0.6755	0.0542	0.5982
	Rel.Bias	-0.9360	-1.0057	-0.9576
	RMSE	0.6816	0.2083	0.6633

ZIP	Est	0.5051	-0.1838	0.2816
	SE	0.0871	0.0061	0.0648
	Rel.Bias	4.0511	-0.0808	-0.0612
	RMSE	0.4144	0.0173	0.0673
Poisson	Est	-0.3683	-0.1498	0.2348
	SE	0.5756	0.0715	0.3959
	Rel.Bias	-4.6834	-0.2511	-0.2173
	RMSE	0.7418	0.0873	0.4011

Tab. 4. The mean and standard deviation of RL for each model considering the different shift Scenario

Distribution	Change	RL	
		Mean	Standard Deviation
Logistic	C0	13.11	10.75
	C1	11.64	10.52
	C2	11.02	10.85
	C3	12.72	12.41
	C4	11.15	10.96
Poisson	C0	19.25	17.84
	C1	5.64	5.03
	C2	3.52	2.71
	C3	15.11	
	C4	13.38	11.01
ZIP	C0	27.60	25.62
	C1	1.00	.00
	C2	1.00	.00
	C3	1.00	.00
	C4	1.00	.00

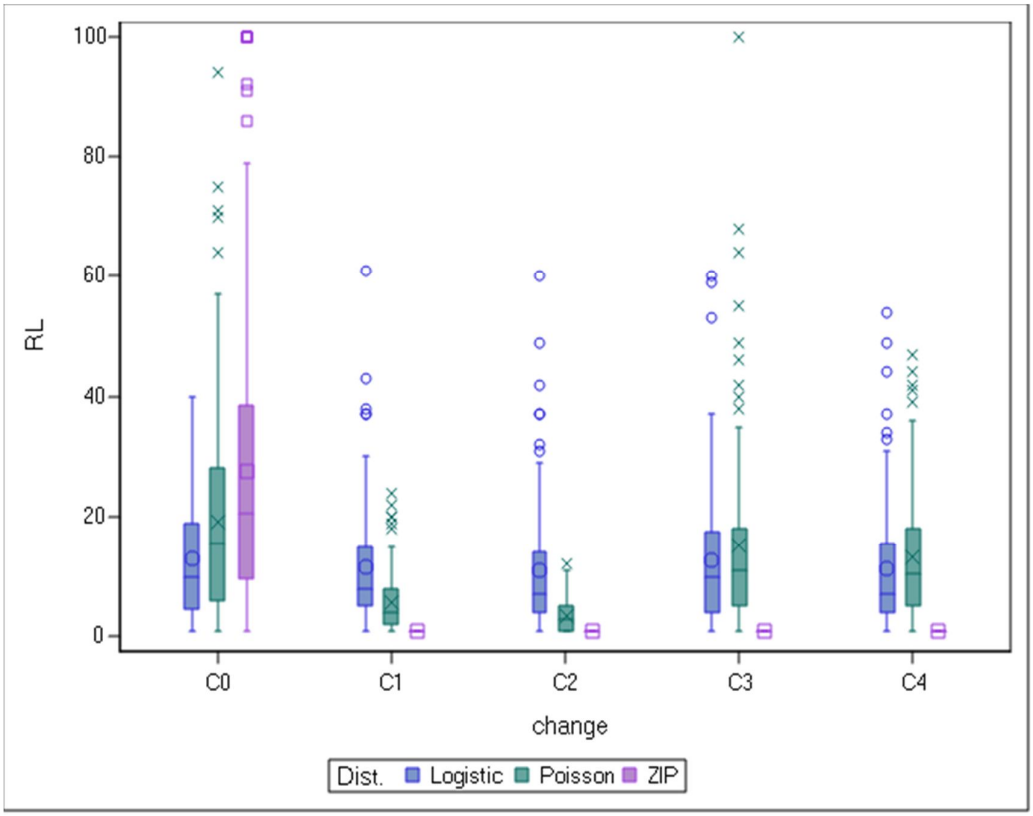

Fig. 2. Box Plot of RL for the three models under different shift scenarios

Tab. 5. RL properties for GLMM and GLM model

Shift	Model	Run Length	
		Mean	Standard Deviation
C0	GLM	42.21	37.69
	GLMM	58.96	38.67
C1	GLM	1.00	.00
	GLMM	1.00	.00
C2	GLM	1.00	.00
	GLMM	1.00	.00
C3	GLM	1.68	4.88
	GLMM	1.09	.67
C4	GLM	1.00	.00
	GLMM	1.00	.00

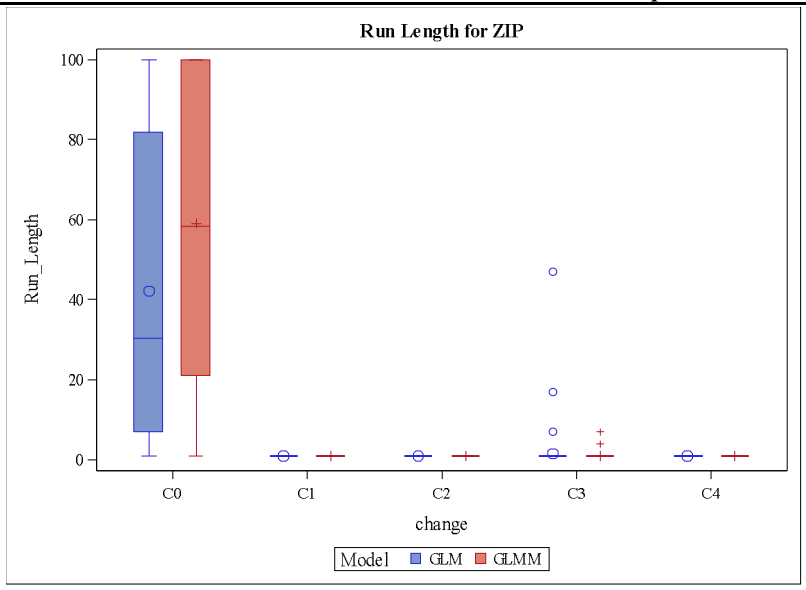

Fig. 3. Box plot of RL criterion for ZIP model when GLMM and GLM are applied

4. Conclusion

This paper discussed social network using the generalized linear mix models when count response variables were considered. Monte Carlo expectation maximization algorithm was used to estimate parameters of models. Our results demonstrated the preference of inflated count model to non-inflated count models in the change detection with respect to sparse nature of social network data. In addition, the outcomes show the superiority of GLMM over GLM considering ARL index. The future investigations may focus on the other parameter estimation approaches in the generalized linear mix models in order to improve change detection performance. In addition, using other models that can incorporate the effect of interrelationship besides the node random effect at each time snapshot can be the subject of another future study.

Reference

[1] Cook, K.S. and J.M. Whitmeyer, *Two approaches to social structure: Exchange theory and* network *analysis*. Annual review of Sociology, Vol. 18, No. 1, (1992), pp. 109-127.

[2] Wasserman, S. and K. Faust, *Social network analysis: Methods and applications*. Vol. 8. (1994), Cambridge university press.

[3] Azarnoush, B., et al., *Monitoring Temporal Homogeneity in Attributed Network Streams*. Journal of Quality Technology, Vol. 48, No. 1, (2016), p. 28.

[4] Frank, O. and D. Strauss, *Markov Graphs*. Journal of the American Statistical Association, Vol. 81, No. 395, (1986), pp. 832-842.

[5] Katz, L. and C.H. Proctor, *The concept of configuration of interpersonal relations in a group as a time-dependent stochastic process*. Psychometrika, Vol. 24, No. 4, (1959), pp. 317-327.

[6] Wasserman, S., *Stochastic models for directed graphs*. (1977), Harvard University.

[7] Sampson, S.F., *Crisis in a cloister*. (1969), Ph. D. Thesis. Cornell University, Ithaca.

[8] Newcomb, T.M., *The acquaintance process*. (1961).

[9] McCulloh, I., et al., *IkeNet: Social network analysis of e-mail traffic in the Eisenhower Leadership Development Program*. (2007), DTIC Document.

[10] McCulloh, I., J. Lospinoso, and K. Carley. *Social network probability mechanics*. in *Proceedings of the World Scientific Engineering Academy and Society 12 th International Conference on Applied Mathematics*. (2007).

[11] Stokman, F.N. and P. Doreian, *Evolution of social networks: processes and principles*. Evolution of social networks, Vol. 1, (1997), p. 997.

[12] McCulloh, I., *Detecting changes in a dynamic social network*. (2009), ProQuest.

[13] Savage, D., et al., *Anomaly detection in online social networks.* Social Networks, Vol. 39, (2014), pp. 62-70.

[14] Sparks, R., *Social Network Monitoring: Aiming to Identify Periods of Unusually Increased Communications Between Parties of Interest*, in *Frontiers in Statistical Quality Control* Vol. 11, (2015), pp. 3-13.

[15] Sparks, R., *Detecting Periods of Significant Increased Communication Levels for Subgroups of Targeted Individuals*. Quality and Reliability Engineering International, (2015).

[16] Miller, B.A., N. Arcolano, and N.T. Bliss. *Efficient anomaly detection in dynamic, attributed graphs: Emerging phenomena and big data*. in *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*. (2013), IEEE.

[17] Mazrae Farahani, E., et al., *A Statistical Approach to Social Network Monitoring.* Communications in Statistics-Theory and Methods, (2016), (just-accepted).

[18] Woodall, W.H., et al., *An overview and perspective on social network monitoring.* arXiv preprint arXiv:1603.09453, (2016).

[19] Sparks, R. and J.D. Wilson, *Monitoring communication outbreaks among an unknown team of actors in dynamic networks.* arXiv preprint arXiv:1606.09308, (2016).

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

[21] Zou, N. and J. Li, *Modeling and change detection of dynamic network data by a network state space model.* IISE Transactions, Vol. 49, No. 1, (2017), pp. 45-57.

[22] Laird, N.M. and J.H. Ware, *Random-effects models for longitudinal data.*Biometrics, (1982), pp. 963-974.

[23] Hall, D.B., *Zero-inflated Poisson and binomial regression with random effects: a case study*. Biometrics, Vol. 56, No. 4, (2000), pp. 1030-1039.

[24] Davidian, M. and D.M. Giltinan, *Nonlinear models for repeated measurement data*. Vol. 62. (1995), CRC press.

[25] Davidian, M. and D.M. Giltinan, *Nonlinear models for repeated measurement data: an overview and update.* Journal of agricultural, biological, and environmental statistics, Vol. 8, No. 4, (2003), pp. 387-419.

[26] McCulloch, C.E., *Maximum likelihood variance components estimation for binary data.* Journal of the American Statistical Association, Vol. 89, No. 425, (1994), pp. 330-335.

[27] McCulloch, C.E., *Maximum likelihood algorithms for generalized linear mixed models*. Journal of the American statistical Association, Vol. 92, No. 437, (1997), pp. 162-170.

[28] Al Hasan, M., et al. *Link prediction using supervised learning*. in *SDM06: workshop on link analysis, counter-terrorism and security*. (2006).

[29] Wu, L., *Mixed effects models for complex data*. (2009), CRC Press.

[30] Jiang, J., *Linear and generalized linear mixed models and their applications*. (2007), Springer Science & Business Media.

[31] Sullivan, J.H. and W.H. Woodall, *A control chart for preliminary analysis of individual observations.* Journal of Quality Technology, Vol. 28, No. 3, (1996), pp. 265-278.

[32] Paynabar, K., J. Jin, and A.B. Yeh, *Phase I risk-adjusted control charts for monitoring surgical performance by considering categorical covariates.* Journal of Quality Technology, Vol. 44, No. 1, (2012), pp. 39-53.

[33] Myers, R.H., et al., *Generalized linear models: with applications in engineering and the sciences*. Vol. 791, (2012), John Wiley & Sons.

[34] Paynabar, K. and A.B. Yeh., Phase I risk-adjusted control charts for monitoring surgical performance by considering categorical covariates. Journal of Quality Technology, Vol. 44, No. 1, (2012), p. 39.

[35] Stokman, F.N. and P. Doreian, *Evolution of social networks: processes and principles.* Evolution of social networks, Vol. 1, (1997), p. 997.

[36] Zou, C. and F. Tsung, *Likelihood ratio-based distribution-free EWMA control charts*. Journal of Quality Technology, Vol. 42, No. 2, (2010), p. 174.

Follow This Article at The Following Site

Mazrae Farahani E., Baradaran Kazemzade R., Albadvi A., Teimourpour B. Modeling and Monitoring Social Ntwork in term of Longitudinal Data. IJIEPR. 2018; 29 (3) :247-259
URL: http://ijiepr.iust.ac.ir/article-1-808-en.html