Advances in Industrial Engineering, Winter 2022, 56(1): 57-72
DOI: 10.22059/AIE.2022.338423.1824

RESEARCH PAPER

Monitoring Social Networks in Phase I using Zero-Inflated Poisson Regression Model

Narges Motalebi [a], Mohammad Saleh Owlia [*,a], Amirhossein Amiri [b], Mohammad Saber Fallahnezhad [a]

a. *Department of Industrial Engineering, Yazd University, Yazd, Iran.*
b. *Department of Industrial Engineering, Shahed University, Tehran, Iran.*

Received: 02 February 2022, Revised: 26 February 2022, Accepted: 01 March 2022

Abstract

In this paper, zero-inflated Poisson (ZIP) regression was assumed as an underlying model to generate network data. This model can be an appropriate model if the network data is sparse and produced with two processes, one generates only zeros and the other generates count data that follow the Poisson model, the two parameters of the model are functions of variables here referred to as similarity variables. The performance of the Likelihood Ratio Test (LRT), a Combined Residual-Square Residual (R-SR), and Hotelling's T^2 control charts was investigated in networks based on the ZIP regression model in Phase I. Traditionally in Phase I, the parameters of the model are unknown and need to be estimated. One needs to be sure the process is stable and the changes are detected and removed. The performance of our proposed methods is compared using simulation when parameters slope and intercept are under step changes. Signal probability was recorded as a comparison measure. The simulation results show that the LRT outperforms two other methods significantly in terms of signal probability. The efficiency of methods was also examined using the real Enron data set.

Keywords:
Likelihood Ratio Test;
Hotelling's T^2;
Residual;
Phase I;
Social Networks

Introduction

Social network analysis (SNA) includes theories, models, and methods based on relational data, we can analyze the people and their relationships using the powerful tools introduced in SNA. Detecting normal patterns of network data is interesting for researchers due to its applications in real-world, various statistical models such as latent variable models [1,2], Stochastic Block Models [3], Exponential random graphs [4] are introduced to model the normal pattern of interaction between pair of individuals. These models try to capture properties and characteristics of networks such as reciprocation, homophily [5], transitivity, small world, power low distribution, etc. [6]. On the other hand, finding out when the normal behavior has changed is also important. Spammers, frauds, and vandal groups are expected not to follow normal patterns, although changes are not limited to those that have adverse effects, for example, the announcement of performing a famous artist after years can also cause the normal patterns changed in social networks.

The performance of various control charts has been investigated to detect such changes while considering different models to represent networks data since Woodall et al. [7] mentioned in

[*] Corresponding author: (M.S. Owlia)
Email: owliams@yazd.ac.ir

their article that tools and methods in statistical control can benefit this area of research and recommended the use of these tools.

In general, statistical control charts have been applied in two Phases. In retrospective Phase I [8], one aim is to remove the special causes to estimate the parameters of the model precisely so the estimated parameters can be used to set control limits in a prospective study (Phase II).

In literature, there are studies that investigated various methods such as likelihood ratio test, Hotelling's T^2 to monitor weighted social networks under the assumption that the underlying model is Poisson regression. Fotuhi et al. [9] developed extended Hotelling's T^2, F, and standardized likelihood ratio test methods to monitor networks in phase I. Farahani et al. [10] developed Hotelling's T^2 and Likelihood ratio test statistics to monitor the network based on Poisson regression in Phase I.

However, the social networks are sparse and this assumption may lead to underestimating the parameters of the model and have detrimental effects on the performance of methods.

While the sparsity of unweighted networks has been considered in models such as logistic regression, in weighted networks we need to consider parameters to model the extra zeros. Mainly, two approaches have been introduced to handle the sparsity of data, hurdle models and zero-inflated models, in zero-inflated Poisson models, zeroes are produced by the two processes, in the hurdle Poisson model zeroes are produced by only one process, and truncated Poisson distribution produce the count data. Ebrahimi et al. [11] considered the sparsity of attributed networks in time with a hurdle state model. Motalebi et al. [12] adopted the ZIP regression model to represent network data. They introduce a latent variable named similarity variable which is a function of nodes attribute and considers the relationship between nodes a function of these variables. They investigated the performance of several methods in a Phase II study.

Motivated by these works, we investigate the power of the Likelihood ratio test and combined residual-square residual chart and Hotelling's T^2 in Phase I, when the underlying mechanism to generate the weighted adjacency matrix is zero-inflated Poisson regression. To our knowledge, no study evaluates the performance of mentioned methods in a Phase I study considering the ZIP regression model. With the ZIP regression model, we can take into account the characteristics such as sparsity as well as a nodal covariate.

The remainder of this article is organized as follows. The next section briefly introduces the model followed by Section 3 which explains the likelihood ratio test and combined residual-square residual and Hotelling's T^2 control charts in detail. Section 4 compares the performance of proposed methods using simulation. Section 5 presents a case study. Section 6 discusses the managerial implication of the methods. Finally, Section 7 provides conclusions and suggestions for future research.

Model

Social networks are sparse, this means that the number of zeros in the adjacency matrix of weighted networks is more than what we expect from processes such as Poisson distribution. The missing data, the interactions with weight equal to zero, and no interactions between nodes are replaced by zero in the adjacency matrix. On the other hand, in the ZIP model, it is assumed that the data is produced with two processes, one process generates zero and the other process generates data with the Poisson distribution, for data with extra zeros the model could be more suitable. To consider the sparsity of networks as well as nodal covariates we assume the ZIP regression model as an underlying model to generate network data. In the following, we explain social networks based on ZIP regression. Assume some undirected weighted networks are available, in each network there are n nodes, the interaction between pairs $(i, j), i, j = 1, \dots, n$

denoted by $y_{ij}(t)$ which is a function of the similarities vector indicated by S_{ij} . $y_{ij}(t)s$ following ZIP regression model. It means that:

$$y_{ij}(t)=\begin{cases} 0 & \text{with probability} \quad \theta_{ij}(t)+(1-\theta_{ij}(t))e^{-\lambda_{ij}(t)} \\ C>0 & \text{with probability} \quad (1-\theta_{ij}(t))\dfrac{e^{-\lambda_{ij}(t)}\lambda_{ij}(t)^{C}}{C!}, \end{cases} \tag{1}$$

The similarity variable is defined as a function of nodal covariates e.g., gender, position, and religion, the parameters of the model are $\lambda(t)$, $\theta(t)$ and satisfy:

$$\log(\lambda_{ij}(t))=\sum_{p}\beta_{p}(t)s_{pij}$$
$$\text{logit}(\theta_{ij}(t))=\log(\frac{\theta_{ij}(t)}{(1-\theta_{ij}(t))})=\sum_{r}\gamma_{r}(t)s_{rij}, \tag{2}$$

To understand the social networks based on ZIP regression we example a hypothetical company with *n* employees. In this company, every two employees (i,j) do not interact with probability θ_{ij} ; i.e., the value of corresponding row *i* and column *j* in its adjacency matrix is zero, or with probability $1-\theta_{ij}$ they interact with the rate which follows Poisson distribution; in other words, the value of corresponding row *i* and column *j* in its adjacency matrix can be zero with probability $(1-\theta_{ij}(t))e^{-\lambda_{ij}(t)}$ and positive value *c* with probability $\dfrac{(1-\theta_{ij}(t))\lambda_{ij}(t)^{c}}{c}$. In this example, at the time *t*, the probability to interact or not to interact can be influenced as $\log it(\theta_{ij}(t))=\gamma_{0}(t)+\gamma_{1}(t)s_{ij}$, the similarity variable s_{ij} can be defined as any function e.g.,

$$\mathrm{s}_{\mathrm{ij}}=\begin{cases} 0 & \text{if the two employees are not in same project} \\ 1 & \text{if the two employees are in same project,} \end{cases}$$

the rate of communication of two employees (i,j) can also be dependent to any function of employments attributes such as work experience differences. Note that there is no information on these zeros i.e., we don't know if they don't interact or interact with weight zero. To find more details on the ZIP regression model in social netwok see Motalebi et al. [12].

The parameters of the model can be estimated by maximizing the likelihood (ML) function or expectation-maximization (EM) algorithm. Here, we assume the θ_{ij} and λ_{ij} are functions of different latent variables which are constant in times.

The mean and variance of ZIP regression are calculated as follows:

$$E(y_{ij}\,|\,\theta_{ij},\lambda_{ij})=(1-\theta_{ij})\lambda_{ij}$$
$$Var(y_{ij}\,|\,\theta_{ij},\lambda_{ij})=(1-\theta_{ij})(\lambda_{ij}+\theta_{ij}\lambda_{ij}^{2}). \tag{3}$$

Where θ and λ follow Eq. 2.

Methods

Control chart based on Likelihood Ratio Test

In this section, we explain the LRT method [13] to detect changes in parameters of the ZIP regression model in Phase I [14]. There are m networks as described in the previous section, a change at the time τ has happened and vectors of model parameters changes form $\beta^{In} \rightarrow \beta^{Out}$ and $\gamma^{In} \rightarrow \gamma^{Out}$, we are interested to detect changes in parameters of the model at a particular τ, this is equal to test Hypothesis below:

$$\begin{cases} H_0: & \theta(t)=\theta^{In}, \gamma(t)=\gamma^{In}, \quad t=1,\dots,m \\ H_1: & \begin{cases} \theta(t)=\theta^{In}, \gamma(t)=\gamma^{In} & t=1,\dots,\tau \\ \theta(t)=\theta^{Out}, \gamma(t)=\gamma^{Out} & t=\tau+1,\dots,m, \end{cases} \end{cases} \tag{4}$$

Under the H$_0$ assumption, the logarithm of likelihood function is written as:

$$l_0 = z_0(t)\{\, log(\theta_{ij}^{In}(t)+(1-\theta_{ij}^{In}(t))exp(-\lambda_{ij}^{In}(t))\}+\sum_{y_{ij}(t)=0}(y_{ij}(t))\log(\lambda_{/}^{In}ij(t))-(\frac{n(n-1)}{2}-z_0(t))$$
$$\{\lambda_{ij}^{In}(t)-\log(1-\theta_{ij}^{In}(t))\}-\sum_{y_{ij}(t)>0}\log(y_{ij}(t)!), \tag{5}$$

Under the H$_1$ assumption, logarithms of likelihood function before and after the changes are written as:

$$l_1 = z_0(t)\{\, log(\theta_{ij}^{In}(t)+(1-\theta_{ij}^{In}(t))exp(-\lambda_{ij}^{In}(t))\}+\sum_{y_{ij}(t)=0}(y_{ij}(t))\log(\lambda_{ij}^{In}(t))-(\frac{n(n-1)}{2}-z_0(t))$$
$$\{\lambda_{ij}^{In}(t)-\log(1-\theta_{ij}^{In}(t))\}-\sum_{y_{ij}(t)>0}\log(y_{ij}(t)!), \tag{6}$$
$$l_2 = z_0(t)\{\, log(\theta_{ij}^{Out}(t)+(1-\theta_{ij}^{Out}(t))exp(-\lambda_{ij}^{Out}(t))\}+\sum_{y_{ij}(t)=0}(y_{ij}(t))\log(\lambda_{ij}^{Out}(t))-(\frac{n(n-1)}{2}-z_0(t))\{\lambda_{ij}^{Out}(t)-\log(1-\theta_{ij}^{Out}(t))\}-\sum_{y_{ij}(t)>0}\log(y_{ij}(t)!),$$

The likelihood ratio statistic for τ is calculated as follows:

$$LRT(\tau) = -2(\hat{l}_0 - \hat{l}_a), \tag{7}$$

so that:

$$l_a = l_1 + l_2.$$

If we know the value of τ, we can estimate the parameters of the model under null and alternative hypotheses and calculate the statistic of LRT according to Eq. 8, but the value of τ is also unknown and we should estimate this point too. Because there is no closed-form solution for estimating parameters of the model and τ simultaneously, a grid search on all possible values $1 \le \tau \le m-1$ is applied. Then, the value of $\hat{\tau}$ which in the maximum of LRT is achieved and its corresponding LRT is considered as an estimated statistic.

In the simulation, we repeat this procedure a number of times when there are no changes in parameters and set the Upper Control Limit (UCL) for desired false passive rate.

Control chart based on Residual-Square Residuals

For each pair *(i,j)* in network at time *t*, the Pearson Residual (PR) is calculated as follows:

$$PR_{ij}(t) = \frac{y_{ij}(t) - (1-\hat{\theta}_{ij})\hat{\lambda}_{ij}}{\sqrt{(1-\hat{\theta}_{ij})(\hat{\lambda}_{ij} + \hat{\theta}_{ij}\hat{\lambda}_{ij}^2)}}, \qquad t = 1,\dots,m \tag{8}$$

where $y_{ij}(t)$ is the number of communications between node *i* and *j* at time *t*, $\hat{\lambda}_{ij}$ and $\hat{\theta}_{ij}$ are estimated when networks are in-control.

The average of residuals over network is calculated.

$$\overline{PR}(t) = \frac{\sum_{ij} PR_{ij}(t)}{\frac{n(n+1)}{2}}, \qquad t = 1,\dots,m \tag{9}$$

where *n* is the number of nodes, we normalize the statistics by subtracting its expected value and dividing its standard deviation, a Shewhart control chart is developed to monitor the statistics below:

$$PR(t) = \frac{\overline{PR(t)} - E(\overline{PR})}{Sd(\overline{PR})}. \qquad t = 1,\dots,m \tag{10}$$

Simultaneously, the square of residual is monitored to detect shifts in the process variance, in case the magnitudes of the residuals are large but very small values for the average of the residuals is achieved because of the signs of the residuals.

$$SPR^2(t) = \sum_{ij} PR_{ij}^2(t), \qquad t = 1,\dots,m \tag{11}$$

The combined Residual-Square Residual control chart signals when either *PR(t) or SPR*^{2}*(t)* falls outside the Shewhart control limits.

The Hotelling's T^2 Control Chart

The Hotelling's T^2 is one of the primary statistics of multivariate literature in Phase I. The statistics is as follows:

$$T_t^2 = (\phi_t - \phi_0)^T \Sigma_0^{-1} (\hat{\phi}_t - \phi_0), \tag{12}$$

ϕ_0 and Σ are estimated when the process is in control, $\hat{\phi}_t$ is a vector of parameters which estimated from the data in time *t*. Ye et al. [15] introduced five methods to estimate parameters ϕ_0 and Σ_0. Here, we investigate Hotelling's T^2 based on a sample average and moving ranges in the context of social network, we assume there are *m* networks. Let $\hat{\phi}_t = (\hat{\beta}_t, \hat{\gamma}_t)$ be the MLE estimator of parameters of network $1 \le t \le m$.

The statistics of Hoteling T^2 is calculated as:

$$T_{Rt}^{2} = (\hat{\phi}_t - \bar{\phi})^T \; S_R^{-1} \; (\hat{\phi}_t - \bar{\phi}), \tag{13}$$

where:

$$\bar{\phi} = \frac{1}{m}\Sigma_{t=1}^{m}\hat{\phi}_j \quad \text{and} \quad S_R = \frac{1}{2(m-1)}\Sigma_{t=1}^{m-1}(\hat{\phi}_{t+1} - \hat{\phi}_t)(\hat{\phi}_{t+1} - \hat{\phi}_t)^T. \tag{14}$$

Performance comparisons

To evaluate and compare the performance of the methods we simulate the environment of a company with $n = 100$ employees, we model the communication between employees with a weighted undirected network so that the communications follow ZIP regression model explained in Section 2. Mathematically, we show a set of networks as $G(t) = (V, Y(t)), t = 1, \ldots, m$ so that in each network y_{ij} is an edge and it's weight is equal to the number of interactions between members *i* and *j*. There are $\binom{100}{2}$ possible y_{ij}'s which are either positive with probability $(1-\theta_{ij})\frac{e^{-\lambda_{ij}}\lambda_{ij}^{y_{ij}}}{y_{ij}}$ or zero with probability $\theta_{ij} + (1-\theta_{ij})e^{-\lambda_{ij}}$. We assume there are two departments labeled RD and QC and employees are distributed through these two departments, so the relative similarity variable is categorical with three levels and is defined as follows:

$$s_{1ij} = \begin{cases} RDRD & \text{if both members are in department RD} \\ QCQC & \text{if both members are in department QC} \\ RDQC & \text{if one member is in department RD and the other is in department QC,} \end{cases} \tag{15}$$

These three levels of similarity variable require two separate variables and here defined as RDRD and QCQC; one level is excluded to avoid a linear dependency.

We choose θ_{ij} to be a function of work experience differences and a random similarity variable of uniform [1,12] distribution (i.e., $s_{2ij} \sim U[1,12]$).

The parameters of model are set to $\beta = (0.1, 0.4, 0.5)$ and $\gamma = (-0.1, 0.2)$, so λ_{ij} and θ_{ij} follow

$$\begin{aligned} \lambda_{ij} &= \exp(0.1 + 0.4 * RDRD_{ij} + 0.5 * QCQC_{ij}) \\ \theta_{ij} &= \operatorname{logit}^{-1}(-0.1 + 0.2 * s_{2ij}). \end{aligned} \tag{16}$$

We assume $m = 10$ networks are available, to compare the LRT control chart with combined R-SR and Hotelling's T^2 control charts, first we set the control charts to have the same false alarm probability equal to 0.05, then measuring their signal probability, the probability of an out-of-control signal, for various out-of-control conditions. All simulations are developed in R software and are available upon request.

Algorithm 1 indicates the process of setting control limit for the LRT method, with 2500 simulations, the control chart for the LRT method is set to 15.92 $(UCL_{LRT} = 15.92)$ corresponding to false (signal) probability 0.05.

```
Algorithm 1: Calculating the value of UCL using LRT method
Input: $Values\ \beta^{In}, \gamma^{In}$
Output: UCL
$LRT = \varnothing$
For i in 1:1000
        Generate a set of $m$ Networks with ZIP regression Model and $\beta^{In}, \gamma^{In}$ as true values
        For $\tau$ in 1:m-1
                Estimate the parameters of model under null and alternative hypotheses (Eqs. 5 and 6)
                Calculate Likelihood ratio test (LRT) according to Eq. 7.
        END FOR
        $LRT = LRT \cup \max_{\tau} LRT^{i}(\tau)$
END FOR
Set UCL so that $\frac{|\mathrm{LRT} > UCL|}{1000}$ is equal to a desired false positive rate
```

To achieve the false alarm probability of 0.05 for the R-SR control chart, we use 2500 simulation and follow Algorithm 2 so that each control chart is set to have a false alarm probability of about $1-\sqrt{1-0.05}=0.025$. We set $CL_{Residual} = \pm 4.59$, and $UCL_{SquarResidual} = 12819$.

```
Algorithm 2: Calculating the value of CL using Residual-Square Residual (R-SR) method
Input: $Values\ \beta^{In}, \gamma^{In}$
Output: CLs
$MPR = \varnothing, SPR = \varnothing$
For i in 1:1000
        Generate a set of $m$ Networks with ZIP regression Model and $\beta^{In}, \gamma^{In}$ as true values
        For $\tau$ in 1:m
                Estimate the mean and Standard deviation with the parameters $\hat{\beta}^{In}$ and $\hat{\gamma}^{In}$, Eq. 3
                Calculate Pearson Residual (PR), Eq. 8
                Calculate the mean of Pearson Residual ($MPR$), then normalize it using Eq. 9 and 10
                Calculate the sum of the square of Pearson Residuals ($SPR$), Eq. 11
        END FOR
$MPR = MPR \cup \max_{i} \overline{PR}^{i}(\tau)$
$SPR = SPR \cup \max_{i} SPR^{i}(\tau)$
END FOR
Set CLs so that $\frac{|LCL_{Residual} < MPR < UCL_{Residual}|}{1000} + \frac{|SPR < UCL_{SquarResidual}|}{1000}$ is equal to a desired false positive rate
```

Using the same number of simulations, the $UCL_{Hotelling' sT^2} = 8.97$ corresponds to the same value for false probability is achieved. Algorithm 3 shows the steps for setting Upper control limit for Hotelling's T^2.

Algorithm 3: Calculating the value of UCL using Hotelling's T^2 method

```
Input: Values $\beta^{In}, \gamma^{In}$
Output: UCL
$T = \varnothing$
For i in 1:1000
        Generate a set of $m$ Networks with ZIP regression Model and $\beta^{In}, \gamma^{In}$ as true values
        For $\tau$ in 1:m
                Estimate the $\bar{\phi} = (\hat{\beta}^{In}, \hat{\gamma}^{In})$ using MLE or EM method, and $S_R$, Eq. 14
                Calculate $T =$ Hotelling's $\mathrm{T}^2$ statistics, Eq. 13
          END FOR
$T = T \cup \max_i T$
END FOR
Set UCLs so that $\frac{|T < UCL|}{1000}$ +is equal to a desired false positive rate
```

Figs. 1, 2, 3, and 4 show the signal probabilities for the step shifts in the parameters of the model, occurring after the 5th and 2nd networks of 10. Signal probabilities from 2500 simulations were determined when a parameters of model changes from β^{In} to $\beta^{Out} = \beta^{In} + \delta\hat{\sigma}_{\beta}$ and γ^{In} to $\gamma^{Out} = \gamma^{In} + \delta\sigma_{\gamma}$. The standard deviation for parameters is estimated as $\sigma_{\beta_0} = 0.11$, $\sigma_{\beta_{RDRD}} = 0.13$, $\sigma_{\gamma_0} = 0.11$ and $\sigma_{\gamma_1} = 0.011$ with 1000 simulations.

From Figs. 1, 2, 3, and 4 we can see, the control chart based on residual-square residual (R-SR) performs better compared to Hotelling's T^2 to detect changes in β_0 and γ_0. Hotelling's T^2 are unable to detect small changes in β_0.

Interestingly, for changes in β_{RDRD} the Hotelling's T^2 method performs better than R-SR control chart. While, The Hotelling's T^2 method performs very poorly, almost unable to detect changes with various magnitude in γ_1. The LRT performs better compared to other methods to detect changes in all β and γ parameters.

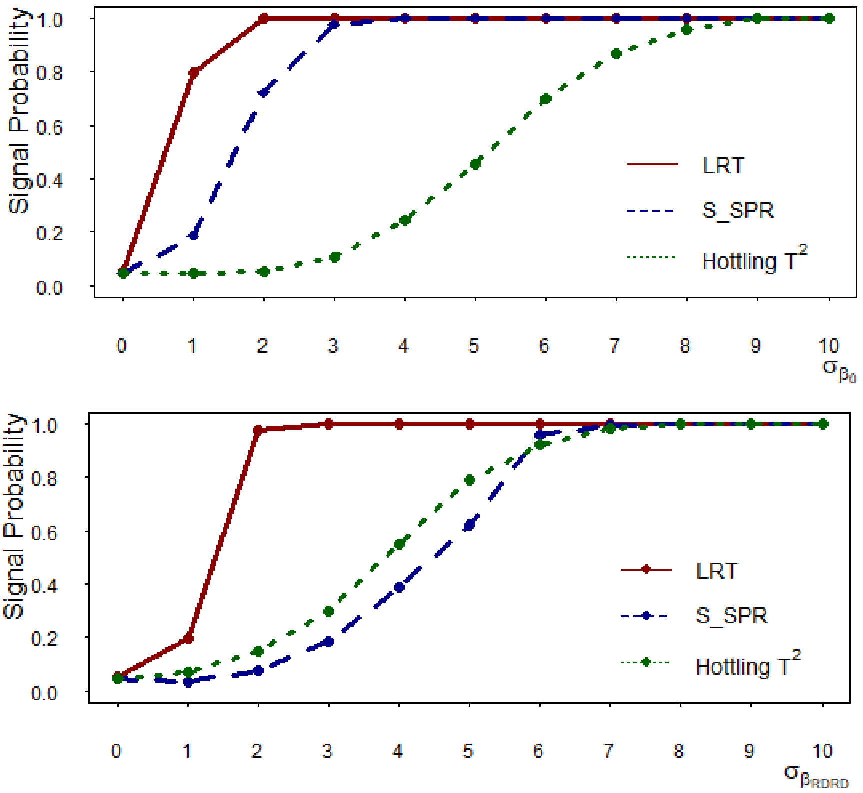


Fig. 1. The signal probabilities for a step shift in β occurring in the second half of the samples

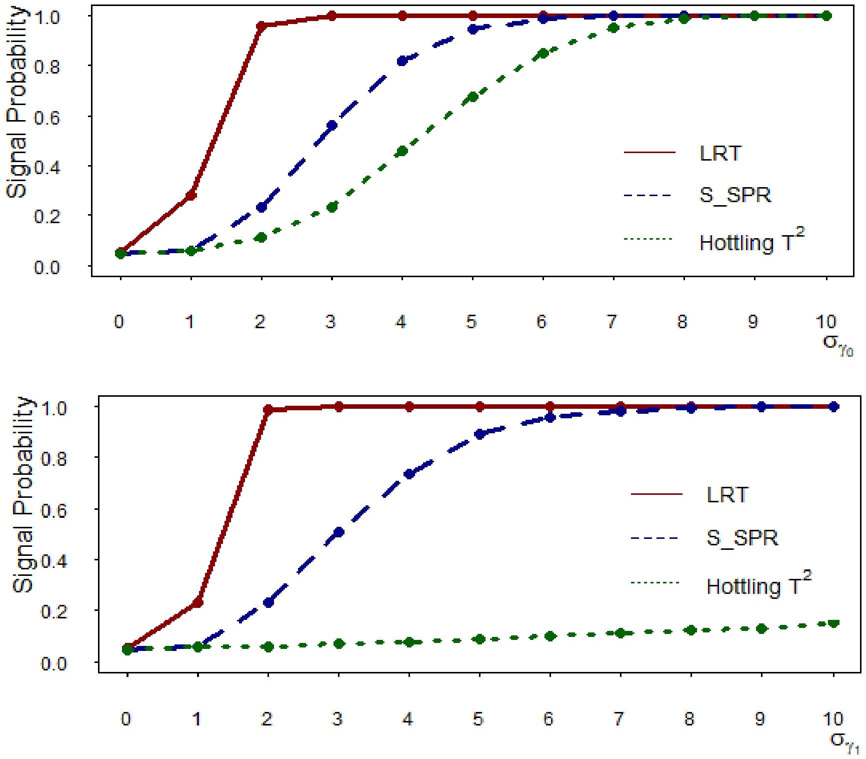


Fig. 2. The signal probabilities for a step shift in γ occurring in the second half of the samples

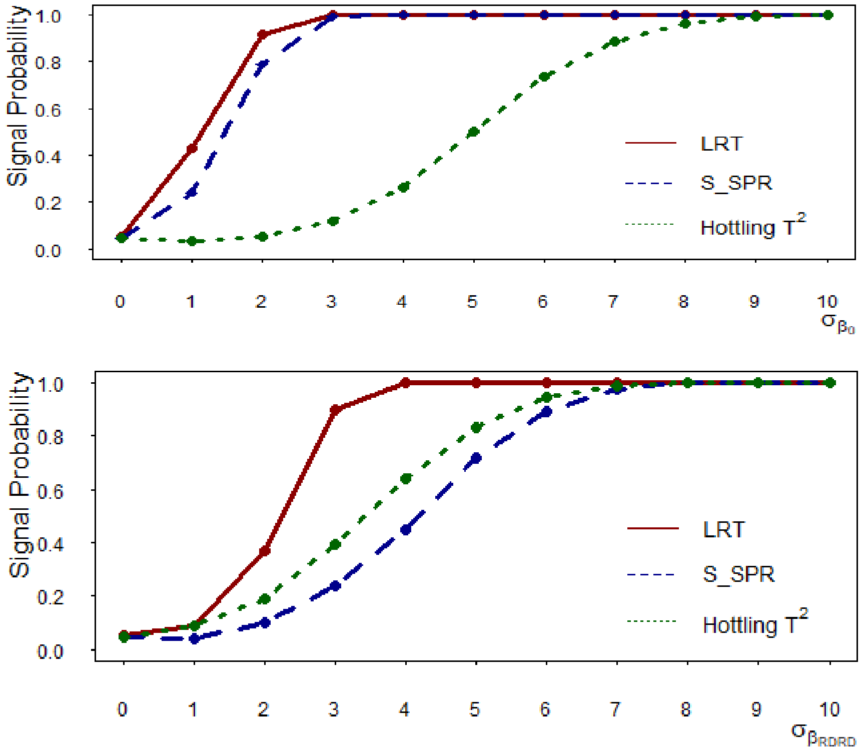


Fig.3. The signal probabilities for a step shift in β occurring after the 2nd of 10 networks

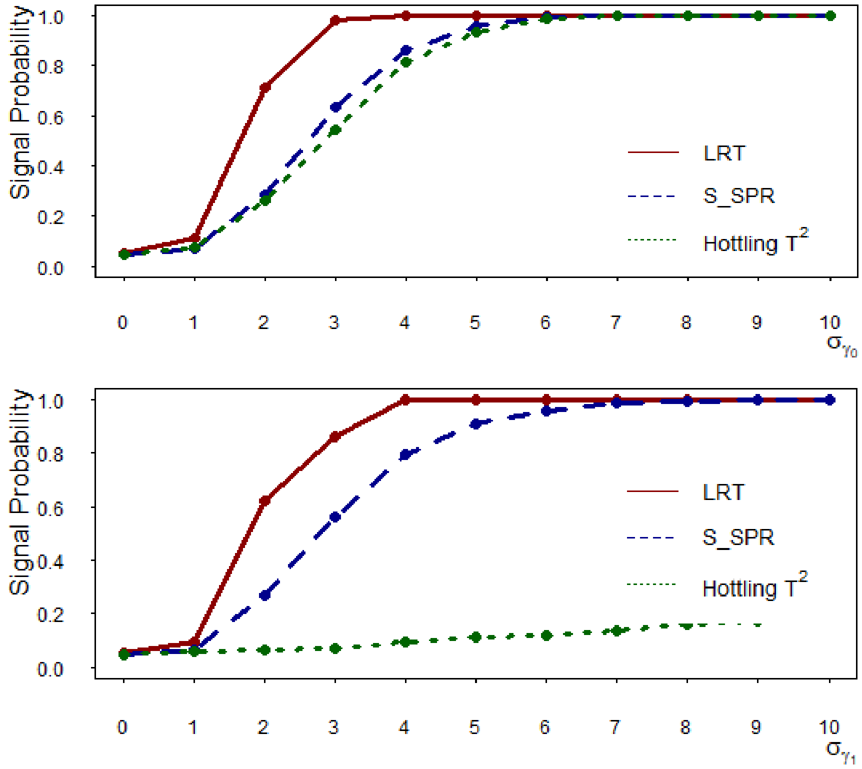


Fig. 4. The signal probabilities for a step shift in γ occurring after the 2nd of 10 networks

The results are shown in Table 1 in details. We find that the LRT chart performs better when the changes happened at the $\tau = 5$. In contrast, the R-SR and Hotelling's T^2 perform slightly better at $\tau = 2$.

Table 1. The signal probabilities occurring after the τ of 10 networks

Parameters	δ	$R-SR$		Hotelling's T^2		LRT	
		τ =5	τ =2	τ =5	τ =2	τ =5	τ =2
	0	0.0484		0.0488		0.0496	
	1	0.1868	0.2388	0.0488	0.0368	0.7924	0.4380
	2	0.7212	0.7872	0.0540	0.0508	1	0.9124
	3	0.9768	0.9896	0.1072	0.1196	1	1
	4	1	1	0.2436	0.2596	1	1
β_0	5	1	1	0.4532	0.5032	1	1
	6	1	1	0.6996	0.7352	1	1
	7	1	1	0.8684	0.8868	1	1
	8	1	1	0.9536	0.9648	1	1
	9	1	1	1	0.990	1	1
	10	1	1	1	0.998	1	1
	1	0.0344	0.0412	0.0716	0.0876	0.1984	0.0872
	2	0.0772	0.0984	0.1464	0.19	0.9746	0.3676
	3	0.1864	0.2356	0.2980	0.3916	1	0.8980
	4	0.3864	0.4464	0.5501	0.6408	1	0.9972
β_{RDRD}	5	0.6184	0.7152	0.7908	0.83	1	1
	6	0.8448	0.89	0.9176	0.9464	1	1
	7	0.9548	0.9768	0.9776	0.9856	1	1
	8	0.9952	0.9984	0.9956	0.9972	1	1
	9	1	1	1	1	1	1
	10	1	1	1	1	1	1
	1	0.0564	0.0692	0.0564	0.668	0.2812	0.11
	2	0.2332	0.2888	0.1148	0.1512	0.956	0.712
	3	0.5632	0.6324	0.2324	0.2860	1	0.9824
	4	0.8172	0.8600	0.4596	0.5384	1	1
γ_0	5	0.944	0.9592	0.6772	0.7360	1	1
	6	0.9848	0.9900	0.8504	0.8948	1	1
	7	0.9976	0.9984	0.9480	0.9616	1	1
	8	0.9992	1	0.9872	0.9908	1	1
	9	1	1	0.9980	0.9996	1	1
	10	1	1	1	1	1	1
	1	0.0568	0.0664	0.0564	0.0556	0.2316	0.094
	2	0.2296	0.2692	0.0576	0.0628	0.9870	0.622
	3	0.5064	0.5600	0.0704	0.0708	1	0.8584
	4	0.7372	0.7932	0.0752	0.0916	1	1
γ_1	5	0.888	0.9108	0.0892	0.1124	1	1
	6	0.9584	0.9592	0.0984	0.1200	1	1
	7	0.9796	0.9844	0.1100	0.1356	1	1
	8	0.9924	0.9940	0.1248	0.1604	1	1
	9	0.9972	0.9984	0.1328	0.1696	1	1
	10	0.9992	1	0.1568	0.1912	1	1

Case Study: ENRON'S EMAIL Network

In this section, we investigate the application of proposed methods in infamous Enron data [16] in Phase I. The data set consists of email communications between Enron employees which were released after the scandals in the company have been revealed. The time of the important event that resulted in the bankruptcy is known which makes it an appropriate data set for investigating the efficiency of various methods [17,18]. We have derived networks of 77 employees of this company for 81 weeks from July 2000 till February 2002, we fit the model in Eq. 17 to show the behavior of the data. The applicability of the model is investigated thoroughly in the extracted data set in Motalebi et al. [12]. y_{ij} is the number of communications between the CEOs and Presidents labeled as CP and directors and managers labeled as DM, which can be zero or positive with the probability below:

$$y_{ij} = \begin{cases} 0 & \theta_{ij} + (1-\theta_{ij})e^{-\lambda_{ij}} \\ C > 0 & (1-\theta_{ij})\dfrac{e^{-\lambda_{ij})}\lambda_{ij}^{y_{ij}}}{y_{ij}\,!}, \end{cases} \tag{17}$$

The similarity variable is categorical and has three levels. To avoid linear dependency we introduce two variables $CPDM$ and $DMDM$ so that the parameters of model follow:

$$\begin{aligned} &\lambda_{ij}(t) = \exp(\beta_0(t) + \beta_{CPDM}(t) * (CPDM) + \beta_{DMDM}(t) * (DMDM)) \\ &\theta_{ij}(t) = \mathrm{logit}^{-1}(\gamma_0(t)). \end{aligned} \tag{18}$$

To apply the proposed methods in Phase I we need to set control limits, to do that, the first step is to estimate the parameters of model for each week. Fig. 5 shows the estimated parameters of model in times. The averages of estimated parameters are $\hat{\beta}_0 = 2.38$, $\hat{\beta}_{CPDM} = 0.78$, $\hat{\beta}_{DMDM} = 0.9$ and $\hat{\gamma}_0 = 3.36$. We simulate 80 networks with these values and calculate the statistics of each method. With 1000 simulations the control limits for each method are set so that the Type I error is equal to 0.05. The control limits for LRT, R-SR, and Hotelling's T^2 are $UCL_{LRT} = 18.03$, $U_{SR} = 5483.83$, $CL_R = \pm 1.25$ and $UCL_{Hotelling\,T^2} = 153.34$, respectively. We can see the results in Figs. 6-8.

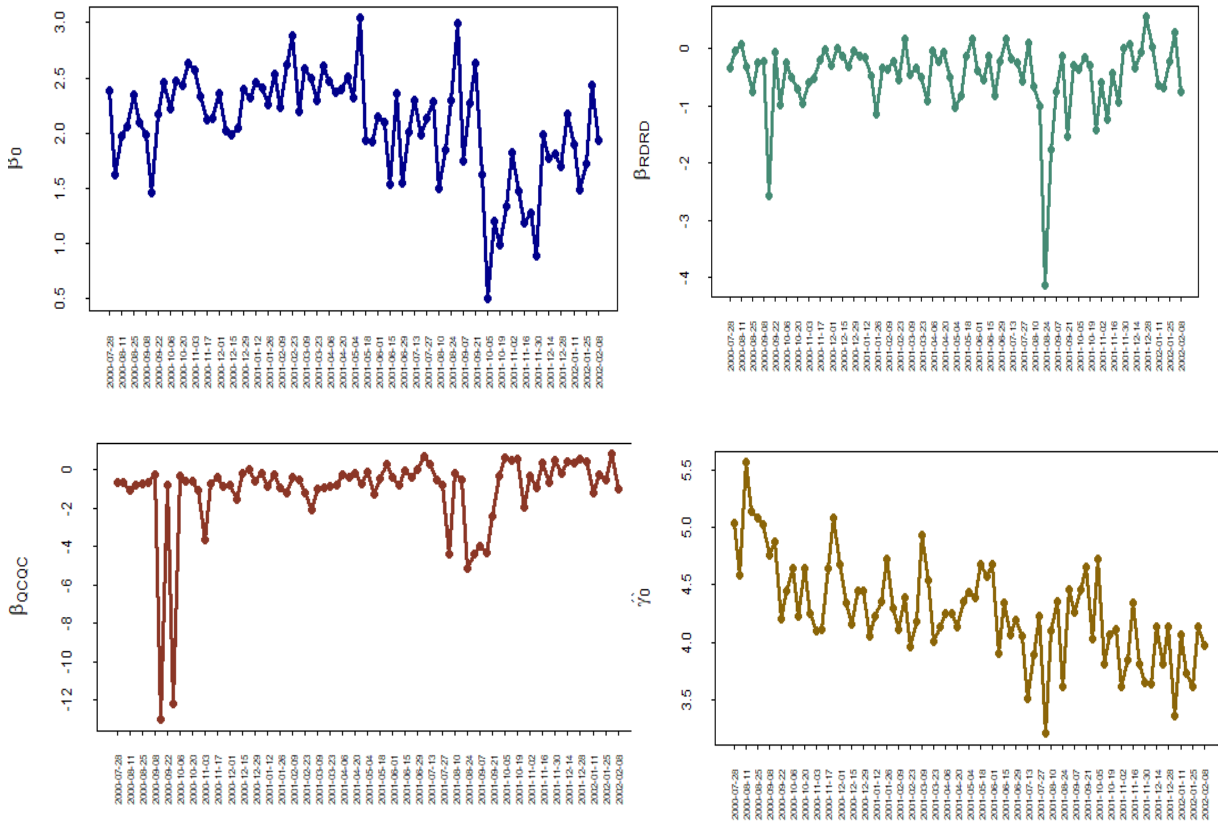


Fig. 5. The estimation values for β_0 , β_{CPDM} , β_{DMDM} and γ_0 in interval time from July 2000 till Feb 2002

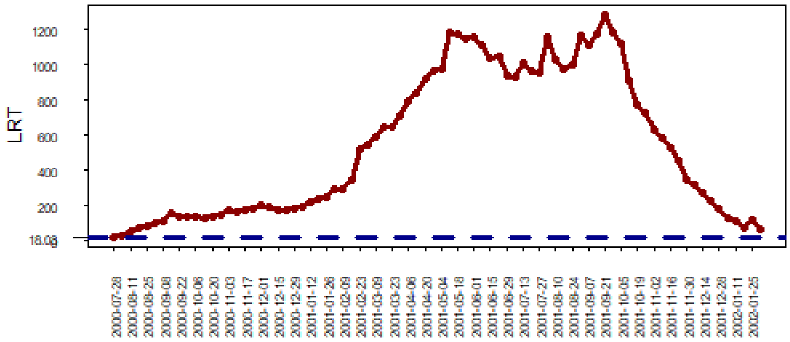


Fig. 6. The LRT method in Enron data from July 2000 till Feb 2002

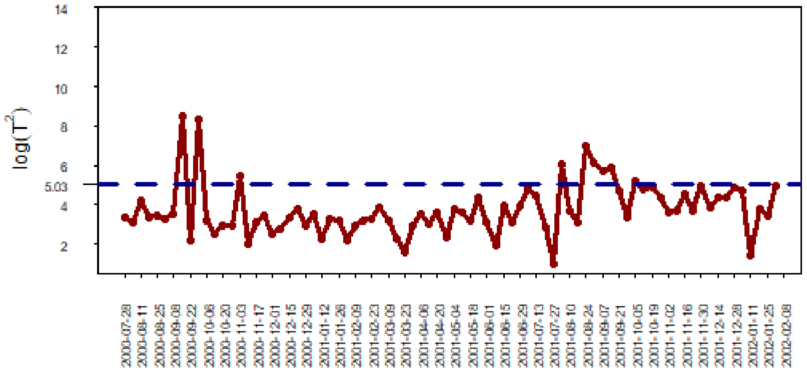
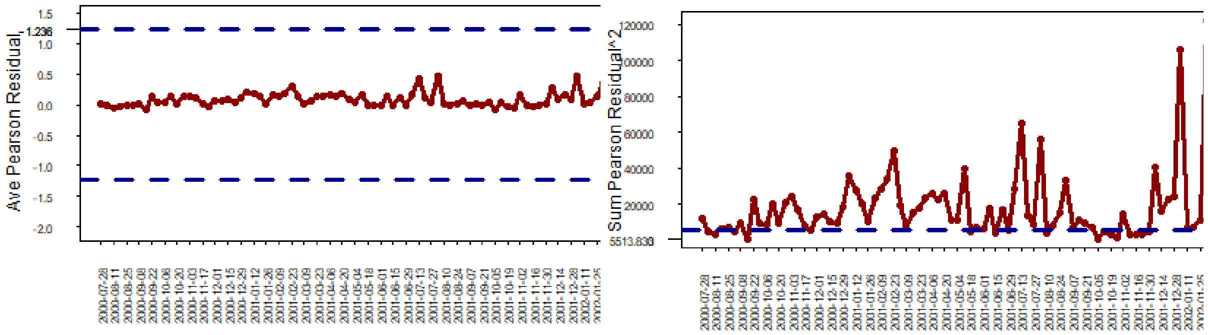


Fig. 7. The T^2 method in Enron data from July 2000 till Feb 2002



a. The mean of residuals

b. The sum square of residuals

Fig. 8. The R-SR method in Enron data from July 2000 till Feb 2002

Managerial insight

Although social networks data are not only formed through the Internet, the advancement of technology and the popularity of using various platforms such as Facebook, Telegram, etc. have expanded social networks, causing this form of data has become increasingly common. According to statistics, in November 2020 close to 50 percent (42.6%) of the population of Iran are users of various social media [19]. The increasing trend of cyberspace usage emphasizes the need to provide methods for monitoring events in this environment. Monitoring social data has many applications, as an example, a study conducted in Iran by Shariatpanahy et al. [20] in connection with cyberbullying in high school students indicates that about 82.29% of the participants in this study have been harassed online, 90.90% of people harass others, and 62.40% of participants have a friend who is harassed. Obviously, rapid identification and intervention can reduce the destructive effects of these behaviors on victims. In another example, at the time of the Queensland floods, messages on social networks about the situation helps first-aid officers to make the right response on time, reducing the disastrous effects [21].

Furthermore, the use of network concepts is evident in finding the terrorist activities and helping to uncover the role and capability of members in hidden networks; finding Saddam Hussein or those responsible for the March 11, 2004, Madrid train bombing are examples of using network tools [22].

While the effectiveness of quality tools especially those proposed in this article have been investigated in various areas of research such as manufacturing and health care, examining these tools to detect changes in such social networks is also recommended [7].

Overall, the two concepts integrated into this article, social network and quality tools have been proven to be effective in several managerial applications; due to the generality of the methods and model presented in this article, the results of this research can be useful in identifying different types of anomalies in different organizations and applications.

Conclusion

In this article, the power of three control charts based on the likelihood ratio test, residual-square residual and Hotelling's T^2 have been compared in Phase I. To do that, we simulated an environment of a company with the assumption that the interaction between employees follows a ZIP regression model.

The results show the LRT method outperforms the two methods significantly for changes in all parameters. Except the parameter of β_{RDRD}, the control chart based on residual-square residual performs better than T^2. We also investigated the efficiency of proposed methods over the infamous Enron data; The LRT and R-SR methods seem to perform better at detecting changes in the data set.

For future research, investigating the performance of methods with different values of m is suggested. Considering the dependency in times would also be interesting.
Adding other nodal statistics in the model, although making inference would be more complex also seems promising for future research.

References

[1] Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). “Latent space approaches to social network analysis”, Journal of the American Statistical association, 97(460), 1090–1098.

[2] Rastelli, R., Friel, N., and Raftery, A. E. (2016). “Properties of latent variable network models”, Network Science, 4(4), 407–432.

[3] Lee, C. and Wilkinson, D. J. (2019). “A review of stochastic block models and extensions for graph clustering”, Applied Network Science, 4(1), 1–50.

[4] Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). “An introduction to exponential random graph (p*) models for social networks”, Social networks, 29(2), 173–191.

[5] McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). “Birds of a feather: Homophily in social networks”, Annual review of sociology, 27(1), 415–444.

[6] Snijders, T. A. (2011). “Statistical models for social networks”, Annual review of sociology, 37.

[7] Woodall, W. H., Zhao, M. J., Paynabar, K., Sparks, R., and Wilson, J. D. (2017). “An overview and perspective on social network monitoring”, IISE Transactions, 49(3),354–365.

[8] L Allison Jones-Farmer, William H Woodall, Stefan H Steiner, and Charles W Champ (2014). “An overview of phase i analysis for process improvement and monitoring “. Journal of Quality Technology, 46(3):265–280.

[9] Fotuhi, H., Amiri, A., and Maleki, M. R. (2018). “Phase ı monitoring of social networks based on poisson regression profiles”, Quality and Reliability Engineering International, 34(4), 572–588.

[10] Mazrae Farahani, E. and Baradaran Kazemzadeh, R. (2019). “Phase i monitoring of social network with baseline periods using poisson regression”, Communications in Statistics-Theory and Methods, 48(2), 311–331.

[11] Ebrahimi, S., Reisi Gahrooei, M., Manakad, S., and Paynabar, K. (2020). “Monitoring sparse and attributed networks with online hurdle models”, IISE Transactions, 1–31.

[12] Motalebi, N., Owlia, M. S., Amiri, A., and Fallahnezhad, M. S. (2021). “Monitoring social networks based on zero-inflated poisson regression model”, Communications in Statistics-Theory and Methods, 1–17.

[13] Sullivan, J. H. and Woodall, W. H. (1996). “A control chart for preliminary analysis of individual observations”, Journal of Quality Technology, 28(3), 265–278.

[14] Motalebi, N., Owlia, M. S., Amiri, A., and Fallahnezhad, M. S. (2021). “Monitoring social networks based on zero-inflated poisson regression model in phase i”.

[15] Yeh, Longcheen Huwang, and Yu-Mei Li. ,(2009), Profile monitoring for a binary response. IIE Transactions, 41(11):931–941.

[16] https://www.cs.cmu.edu/~enron/

[17] Azarnoush, B., Paynabar, K., Bekki, J. and Runger, G (2016), “Monitoring temporal homogeneity in attributed network streams”, Journal of Quality Technology, 48(1), 28–43

[18] Yu, L., Woodall, and Tsui, K., (2018), “Detecting node propensity changes in the dynamic degree correctedstochastic block modelLisha”, Social Networks, 54, 209-227.

[19] https://datareportal.com/reports/digital-2021-iran

[20] Shariatpanahi, G., Tahouri, K., Asadabadi, M., Moienafshar, A., Nazari, M., and Sayarifard, Azadeh. (2021), “Cyberbullying and Its Contributing Factors Among Iranian Adolescents”, International Journal of High Risk Behaviors and Addiction, In Press.

[21] Zhou, X., anChen, Lei , “Event detection over twitter social media streams”, The VLDB journal, 23(3), 381—400.

[22] http://networksciencebook.com/