Zaki, Abeer A.; Saleh, Nesma A.; Mahmoud, Mahmoud A.

Article

Assessing the use of the moving window approaches when monitoring social networks using a degree corrected stochastic block model

Review of Economics and Political Science (REPS)

Provided in Cooperation with:
Cairo University, Cairo

Suggested Citation: Zaki, Abeer A.; Saleh, Nesma A.; Mahmoud, Mahmoud A. (2021) : Assessing the use of the moving window approaches when monitoring social networks using a degree corrected stochastic block model, Review of Economics and Political Science (REPS), ISSN 2631-3561, Emerald, Bingley, Vol. 6, Iss. 4, pp. 311-327, https://doi.org/10.1108/REPS-08-2020-0125

This Version is available at:
https://hdl.handle.net/10419/316047

Assessing the use of the moving window approaches when monitoring social networks using a degree corrected stochastic block model

Received 24 August 2020
Revised 26 July 2021
Accepted 26 July 2021

Abeer A. Zaki, Nesma A. Saleh and Mahmoud A. Mahmoud
Statistics Department, Faculty of Economics and Political Science, Cairo University, Cairo, Egypt

Abstract

Purpose – This study aims to assess the effect of updating the Phase I data – to enhance the parameters' estimates – on the control charts' detection power designed to monitor social networks.

Design/methodology/approach – A dynamic version of the degree corrected stochastic block model (DCSBM) is used to model the network. Both the Shewhart and exponentially weighted moving average (EWMA) control charts are used to monitor the model parameters. A performance comparison is conducted for each chart when designed using both fixed and moving windows of networks.

Findings – Our results show that continuously updating the parameters' estimates during the monitoring phase delays the Shewhart chart's detection of networks' anomalies; as compared to the fixed window approach. While the EWMA chart performance is either indifferent or worse, based on the updating technique, as compared to the fixed window approach. Generally, the EWMA chart performs uniformly better than the Shewhart chart for all shift sizes. We recommend the use of the EWMA chart when monitoring networks modeled with the DCSBM, with sufficiently small to moderate fixed window size to estimate the unknown model parameters.

Originality/value – This study shows that the excessive recommendations in literature regarding the continuous updating of Phase I data during the monitoring phase to enhance the control chart performance cannot generally be extended to social network monitoring; especially when using the DCSBM. That is to say, the effect of continuously updating the parameters' estimates highly depends on the nature of the process being monitored.

Keywords DCSBM, Fixed window, Moving window, Social network, SPC

Paper type Research paper

1. Introduction

Statistical process control (SPC) is a set of statistical techniques used to achieve process stability through reducing its variability to the extent possible. Control charts are one of the most important and common SPC tools. Their main objective is to monitor processes with the aim to detect changes in the process parameter(s) (e.g. mean, variance, number of defects, etc.) resulting from special causes of variation. Control charts are usually implemented through two phases; Phase I and Phase II. In the case of unknown process parameters, the Phase I analysis is implemented in order to reliably estimate the unknown control chart limits used to monitor the process in Phase II (the monitoring phase). SPC techniques are commonly applied

The authors would like to thank the editor and the anonymous reviewers for their insightful comments.

Review of Economics and Political Science
Vol. 6 No. 4, 2021
pp. 311-327
Emerald Publishing Limited
e-ISSN: 2631-3561
p-ISSN: 2356-9980
DOI 10.1108/REPS-08-2020-0125

in manufacturing and service industries. Recent applications of the SPC techniques have been discovered; among which is the social network analysis.

Social network analysis (SNA) implies monitoring a pattern of communications between a group of actors over time to identify and detect unusual levels of interactions among them. There exists a number of important applications of social networks; e.g. online social systems, security monitoring, advertisement targeting, financial transactions, terrorist networks and e-mail networks (Savage *et al.*, 2014). For example, SNA can be used to track, analyze and assess terrorist (dark) networks; by which it can shed the light on their structures of recruitment and evaluate the roles played by their individuals. Also, in the marketing field, SNA can be used to assess the extent by which a certain marketing campaign succeeded; through monitoring the clients' interactions on the promoted products on the company's website. In the literature, various statistical methods have been proposed for monitoring social networks with the aim of detecting significant changes in a networks' structure. Among these methods are the scan statistic methods (Priebe *et al.*, 2005; Marchette, 2012; Zhao *et al.*, 2018b) and the modeling method; e.g. the degree corrected stochastic block model (Karrer and Newman, 2011; Yan *et al.*, 2014; Wilson *et al.*, 2019), the log-linear model (Wasserman and Pattison, 1996; Miller *et al.*, 2013) and the logistic regression model (Wasserman and Pattison, 1996; Azarnoush *et al.*, 2016). Recently, SPC tools have been used as additional statistical methods for monitoring social networks. Zhao *et al.* (2018b) illustrated that a structural change in the network can be detected through using some baseline data (namely window of data) to identify the typical behavior of the network's individuals (nodes). If the network nodes' behavior deviated from this typical behavior, then it is considered "anomalous". Savage *et al.* (2014) defined social network anomalies to be the changes that happen in the network structure as a consequence of the deviations in the pattern of communications between some or all the nodes in the network. It is worth mentioning that social networks are non-stationary by nature; which means that the pattern of communications changes constantly over the time (Woodall *et al.*, 2017). Thus, the aim of social network monitoring is to detect significant and influential changes.

A variety of models have been proposed to describe social networks; ranging from the simplest models (such as Erdös-Rényi's (1959) model) that just represent the connection probability between the nodes to the most complex models depicting the properties of realistic networks such as degree heterogeneity and community structure. The degree corrected stochastic block model (DCSBM), introduced by Karrer and Newman (2011), is one of the models designed for network surveillance. It basically models the community structure in networks with degree heterogeneity. Wilson *et al.* (2019) integrated an SPC technique with a parametric random graph model to monitor dynamic networks in order to detect significant structural anomalies. They applied the Shewhart control chart to monitor the parameters of a dynamic version of the DCSBM. They depended on a fixed-time and fixed-size window of data (networks) to estimate reliable Shewhart control limits. Different studies in the literature have shown that continuously updating the control limits' estimates during the on-line monitoring of the process improves the chart's performance in detecting out-of-control conditions. To our knowledge, no comparative studies between fixing and updating the control limits have been conducted on social network monitoring.

Motivated by the recommendations of Wilson *et al.* (2019), this study basically aims to extend their work through the use of the moving window approaches in estimating the Shewhart limits rather than the fixed window approach. Two types of the moving window approach are considered; the fixed-sized moving window and the variable-sized moving window. A fixed-sized moving window implies that, with each newly drawn observation, the window is updated by excluding the oldest observation and including the newest one; under the condition that the latter is not an anomaly. A variable-sized moving window starts with a small-sized window and then gradually increases by the inclusion of in-control Phase II observations until it reaches a certain pre-determined size, then the oldest observation is

excluded and the most recent one is included. Also, it is of interest to extend Wilson *et al.*'s (2019) study by monitoring the DCSBM parameters using the exponentially weighted moving average (EWMA) control chart with both fixed and moving window types. The EWMA chart is well-known to be more efficient in detecting small magnitudes of shifts as compared to the Shewhart chart. For each control chart, Shewhart and EWMA, a performance comparison is conducted when using fixed window, fixed-sized moving window and variable-sized moving window in estimating the chart control limits.

The remainder of the paper is organized as follows. Section 2 provides a review on the DCSBM. Section 3 illustrates the procedure of using quality control charts in monitoring the DCSBM parameters. Section 4 presents a performance comparison between the fixed and the different moving window approaches when used in monitoring the DCSBM parameters. Finally, Section 5 highlights the concluding remarks and the study recommendations.

2. The degree corrected stochastic block model

Relationships within a social network, defined as $G = (V, E)$, can be modeled as a graph with a set of *nodes* (V) representing, e.g. people, e-mails, universities, etc., and a set of *edges* (E) representing the links or connections between these nodes. The number of the nodes in the network is referred to as the network *order*, while the number of edges between the network nodes is referred to as the network *size* (see, for example, Hoppe and Reinelt, 2010; Landherr *et al.*, 2010; Woodall *et al.*, 2017; Wilson *et al.*, 2019). Suppose having a network that consists of n nodes at time points $t = 1, 2, 3, \ldots$, then the information about the edges in this network at each time point can be presented in form of a matrix. This matrix is called *Adjacency Matrix* $\mathbf{A_t} = [a_{ij}]$ of order $n \times n$, where $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, n$. Each matrix entry (a_{ij}) represents the communication level between the ith node and jth node such that $i \neq j$. Hence, the main diagonal elements of $\mathbf{A_t}$ are all zeros. In literature, different statistical models have been used for monitoring social networks in order to detect significant changes in networks' structures. This study is interested in the DCSBM which is considered to be an extension for the stochastic block model (SBM) applied by Fienberg and Wasserman (1981), Holland *et al.* (1983), Snijders and Nowicki (1997) and Bickel and Chen (2009).

In practice, networks are often divided into groups or communities. A community is usually defined as a set of nodes with more connections inside the set than outside. This implies that the number of edges between the nodes in the same community is larger than the number of edges between the nodes in different communities. The SBM is one of the model-based methods introduced to represent the community structure in networks. However, the SBM has some restrictions reflected in its assumptions. For example, it assumes that all the nodes within the same community are stochastically equivalent. This means that the model ignores the variation in the nodes' degrees within a community. Thus, the probability of an edge between two nodes is a function in the nodes' membership only which is often unsuitable for most real networks (see, for example, Zhao *et al.*, 2012; Yan *et al.*, 2014). A node's *degree* is one of the most well-known centrality measures in social network analysis that reflects the role of the node in the network. It is defined as the number of communications (edges) in which the node of interest takes part with other nodes. It measures the tendency of the node to communicate with the remainder of the network by determining the number of nodes that could be reached by the node of interest directly. The ith node degree is calculated as follows;

$$d_{(i)} = \sum_{j=1}^{n} a_{ij}, \quad i = 1, 2, \ldots, n, \tag{1}$$

where a_{ij} is the ijth element of the adjacency matrix $\mathbf{A_t}$ defined previously.

Karrer and Newman (2011) presented an improved version of the SBM that is called the Degree Corrected SBM (DCSBM). The DCSBM takes into consideration the degree

heterogeneity within communities that allows variation in the nodes' degree within the same community. The DCSBM is a probability distribution $P(\cdot) = P(\cdot|\boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{P})$ that is characterized by three main parameters; which are: **(1)** the *degree parameters vector* $\boldsymbol{\theta}^{\mathbf{T}} = (\theta_1, \theta_2, \ldots, \theta_n)$ of order $1 \times n$ representing the propensity of the ith node to connect, $i = 1, 2, \ldots, n$; **(2)** the *containment probabilities vector* $\boldsymbol{\pi}^{\mathbf{T}} = (\pi_1, \pi_2, \ldots, \pi_k)$ of order $1 \times k$ representing the probability of a node to belong to community r, where $r = 1, 2, \ldots, k$, and $\sum_{r=1}^{k} \pi_r = 1$, and **(3)** the *symmetric connectivity matrix* $\mathbf{P} = [P_{r,s}]$ of order $k \times k$, where $r, s = 1, 2, \ldots, k$, representing the propensity of connection between nodes in the communities.

When generating a random graph with n nodes and k communities under the DCSBM, the nodes are usually assigned randomly to communities with labels $\mathbf{c}^{\mathbf{T}} = (c_1, c_2, .., c_n)$, where $c_i \sim$ Multinomial $(1, \boldsymbol{\pi})$, such that $i = 1, 2, \ldots, n$, and $\boldsymbol{\pi}^{\boldsymbol{T}} = (\pi_1, \pi_2, \ldots, \pi_k)$. However, the community labels are sometimes determined deterministically and not generated randomly as mentioned above. Wilson *et al.* (2019) assumed that the community labels are fixed, which is also assumed in this study. In addition, the number of the edges $(w_{i:j}|\boldsymbol{\theta}, \mathbf{c}, \mathbf{P})$, where $i, j = 1, 2, \ldots, n$, is generated from the Poisson distribution with mean $E(w_{i:j}|\boldsymbol{\theta}, \mathbf{c}, \mathbf{P}) = \theta_i \ \theta_j \ P_{c_i,c_j}$, where P_{c_i,c_j} presents the propensity of connection between nodes in the communities to which nodes i and j belong. Following Wilson *et al.* (2019), the maximum log-likelihood estimators (MLEs) for the DCSBM parameters are given by

$$\widehat{\theta}_i = \frac{d_{(i)}}{n_r^{-1} \sum\limits_{w:c_w=c_i} d_{(w)}} \qquad \widehat{p}_{r,s} = \frac{m_{r,s}}{n_r n_s} \tag{2}$$

where $\sum\limits_{w:c_w=c_i} d_{(w)}$ is the sum of the degree measures for all nodes that belong to the community to which the node of interest (i.e. ith node) belongs, n_r and n_s represent the number of nodes in community r and community s, respectively, and $m_{r,s}$ represents the total weight of edges between community r and s. As shown, the MLEs for all the DCSBM's parameters have a closed-form expression.

The identification of this model requires a constraint which is $\sum\limits_{i:c_i=r} \theta_i = n_r$ which means that the sum of θ_i for all nodes within the same community is equal to the total number of nodes in it. For further details, the reader is referred to Yan *et al.* (2014) and Wilson *et al.* (2019).

3. Monitoring the DCSBM using quality control charts

As previously illustrated, the main objective of network surveillance is to detect any influential change in the communication level between the network nodes. Identifying changes in the communication pattern requires inferring the normal structure of interactions across a sequence of networks, such that any shift from this norm at any time point is considered an anomaly.

One approach to detect anomalies is to monitor social networks using control charts. With control charts, a statistic S_t should be determined to represent some summary of the network structure. This statistic S_t could represent, for example, any of the common global centrality network metrics (e.g. closeness, betweenness, degree, links per node, ... etc.), or some likelihood ratio statistic associated with a parameter of a parametric model describing the network. Once S_t is specified, it is calculated for each network within a window of data which serves as a baseline data of size m. Then, using these m statistics (S_t), the Phase II control limits are estimated. Variations between these control limits define the typical behavior between the nodes in the network. For $t > m$, the on-line monitoring phase starts such that S_t is calculated for each new network and somehow compared with the estimated control limits. If it exceeds the limits, then a possible structural change in the network has occurred.

Wilson *et al.* (2019) studied the effectiveness of the DCSBM in monitoring the random graphs through a strategy that merges the probabilistic model for modeling the dynamic networks and control charts. They utilized this strategy to analyze and monitor the behavior of individuals in networks through time. They used the Shewhart control chart to monitor the estimated parameters of the DCSBM defined in Equation (2) to identify the anomalies that occur in the network structure. In this study, their steps are followed in monitoring the DCSBM parameters using both the Shewhart and EWMA charts. Their approach can be described as follows:

(1) Suppose having a group of networks G_t over the time $t = 1, 2, \ldots, m, \ldots$, such that $G_t = ([n], w_{i,j}; i, j = 1, 2, \ldots, n)$, where n is the number of nodes in each network, and w_{ij} presents the number of edges between node (i) and node (j). Each network consists of k communities.

(2) At each time point (t), the MLEs for $\boldsymbol{\theta}$ and $\mathbf{P}$ are calculated as defined in Equation (2); each representing the monitored statistic S_t. Note that each parameter is monitored separately.

(3) The first m networks are then used to design the desired control chart; representing the Phase I data set. That is, the control limits are estimated using the m statistics (S_t) calculated from the baseline networks $G_t \in \{G_1, \ldots, G_m\}$ representing different snapshots of the network, each of n nodes, through the time points $t = 1, 2, \ldots, m$. In this study, the control limits are designed using the mean and standard deviation of the m statistics (S_t); namely $\widehat{\mu}_S$ and $\widehat{\sigma}_S$, respectively.

(4) The estimated control limits are then used in monitoring the networks starting from time $t > m$. At each time point $t > m$, the statistic S_t (representing each one of the MLEs) is calculated for each new network G_t and plotted on its corresponding control chart. If the chart statistic exceeds one of the control limits, then there is a possible structural change in the network.

As previously mentioned, both the Shewhart and EWMA charts are used to monitor the maximum likelihood estimators of the DCSBM parameters defined in Equation (2). In the next two subsections, the two control charts designed to monitor the DCSBM parameters are presented.

3.1 The Shewhart control chart

Statistical control charts concept was first introduced by Shewhart (1924). For monitoring the DCSBM parameters, the plotted chart statistic at time t is S_t. The chart signals when S_t exceeds the estimated control limits given by;

$$\widehat{\mu}_S \pm L_h \widehat{\sigma}_S, \tag{3}$$

where L_h is a chart design parameter whose value is chosen to satisfy a specific value for the in-control average run length (ARL). The run length is defined as the number of plotted chart statistics until the chart gives a signal.

Shewhart charts are considered the simplest type of control charts in terms of its computations and interpretation. However, they are only effective in detecting large magnitudes of shifts in process parameters. This is because they depend only on the current statistic to sentence the process.

3.2 The EWMA control chart

The EWMA control chart was proposed by Roberts (1959). The EWMA chart assigns weights for both the recent and the previous observations. Accordingly, the chart effectively detects small and moderate magnitude shifts in process parameters. The EWMA chart statistic is defined as;

$$Z_t = \lambda S_t + (1-\lambda)Z_{t-1}, \quad 0 < \lambda \leq 1, \quad t = 1, 2, 3, \ldots, \tag{4}$$

where Z_t is the current EWMA statistic at time t, Z_{t-1} is the previous EWMA statistic at time $(t-1)$, S_t is the current sample statistic at time t and λ is a chart design parameter – referred to as a smoothing parameter – representing the weights assigned to the observations. The value Z_0 is often set to the target value of the process parameter of interest. The estimated asymptotic control limits of the EWMA chart are given by;

$$\widehat{\mu}_S \pm L_e \sqrt{\frac{\lambda}{2-\lambda}} \widehat{\sigma}_S, \tag{5}$$

where L_e is another chart design parameter that is chosen to produce a desired value for the in-control ARL. If one is interested in detecting small shifts in the process parameters, small weights (λ) are given to the recent observations.

4. Performance assessment

In this simulation study, it is of interest to monitor *undirected weighted* networks. An undirected network implies that each entry a_{ij} in the adjacency matrix $\mathbf{A}_t$ presents the communication level between nodes (i) and (j) regardless of who initiated the communication. In this case, the matrix $\mathbf{A}_t$ is symmetric. A weighted network implies that the entry a_{ij} presents weights for the level of communications between the nodes such as the number of contacts for each pair (i, j) of nodes; in which case a_{ij} is modeled by a Poisson distribution.

Furthermore, the networks' community structure is assumed to be previously determined, and the initial form of the matrix $\mathbf{P}$ is set the same as in Wilson *et al.*'s (2019) study; in which

$$\mathbf{P} = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{bmatrix}.$$

That is, the values on the main diagonal of $\mathbf{P}$ are twice those on the off-diagonal. This is because, it is expected that the nodes within the same community are more likely to communicate with each other than those from different communities. Also, each network is assumed to have $n = 100$ nodes and $k = 2$ equally-sized communities.

The uniform distribution is used to generate θ_i, $i = 1, 2, \ldots, n$, with parameters $(1-\delta_{c_i}, 1+\delta_{c_i})$; such that $\delta_{c_i=r} = 0.5; r = 1, 2, \ldots, k$ based on the ith node membership. The uniform distribution is chosen for simplicity. Wilson *et al.* (2019) pointed out that any non-negative random variable with finite mean and variance could be used. For example, Zhao *et al.* (2018a) used the Pareto distribution to generate θ_i to represent degree heterogeneity with skewed degree distributions. Then, these values are scaled to satisfy the constraint, $\sum_{i:c_i=r} \theta_i = n_r$, which is necessary for the model identification.

It is important to note that the statistic $\widehat{\theta}_i$ defined in Equation (2) is calculated for each node $i = 1, 2, \ldots, n$, but it is practically impossible to build a single control chart for each $\widehat{\theta}$. Accordingly, Wilson *et al.* (2019) suggested monitoring the pooled estimate of the standard deviation of the estimates $\widehat{\theta}_i$. Hence, when the chart is designed to monitor the connection variability between all nodes in the network, the pooled estimate of the standard deviation of $\widehat{\theta}$ is calculated based on sd_1 and sd_2 where;

$$sd_r = \sqrt{\frac{1}{n_r - 1} \sum_{i:c_i=r} (\widehat{\theta}_i - 1)^2}, \quad r = 1, 2. \tag{6}$$

In this study, both the fixed window and the fixed-sized moving window approaches are used with two possible window sizes; $m = 500$ or $m = 1{,}000$ networks. As for the variable-sized moving window, it starts with two possible initial window sizes; $m = 100$ or $m = 400$ networks, in which they gradually increase to reach $m = 500$ and $m = 1{,}000$ networks, respectively. Using 5,000 simulation runs, the steady-state average of average run length (AARL) metric is calculated. The steady-state ARL is defined to be the average number of samples until a signal is given after the process condition has reached a steady-state (i.e. process is running for a long period in an in-control condition). The values of the shift sizes (ε and τ) are same as that used in Wilson *et al.*'s (2019) study. Table 1 provides a description for the out-of-control scenarios considered in the simulation study. The first column enumerates the scenario's reference number. The second column illustrates the change introduced to the model parameter(s). The third column entitles the control chart expected to detect the corresponding parameter(s) change. It is worth to note that Wilson *et al.* (2019) ran all the control charts listed in the third column in each and every out-of-control scenario. In this simulation study, the same criteria are followed and the same conclusion is obtained. That is, as long as the control chart is not designed to monitor the parameter of interest, it does not signal prior to the chart designed for monitoring it. Accordingly, the results presentation is restricted on the performance of the chart designed for monitoring the specified parameter.

In the simulation settings, the Shewhart and the EWMA chart design parameters (L_h) and (λ, L_e), respectively, are determined such that they produce a nominal in-control ARL value of about 370. Tables 2 and 3 provide the values of these design parameters obtained using 5,000 simulation runs.

Tables 4–8 present the AARL values of the Shewhart and EWMA charts designed using the fixed window, fixed-sized moving window and variable-sized moving window approaches for each of the out-of-control scenarios (1–5) illustrated in Table 1.

Case	Description	Control chart
1	Local change in the mean interaction within the first community (i.e. $P_{11}^{*} = P_{11} + \varepsilon$)	P_{11}
2	Global changes in the mean interaction within and between the communities (i.e. $P_{ij}^{*} = P_{ij} + \varepsilon$)	P_{11}, P_{12}, P_{22}
3	Local change in the connection variability within the first community (i.e. $\delta_1^{*} = \delta_1 + \tau$)	"S" which is a pooled estimate of the standard deviation of $\widehat{\theta}$ based on sd_1 and sd_2 defined in Equation (6)
4	Global change in the connection variability between all nodes in the network (i.e. $\delta_r^{*} = \delta_r + \tau$)	"S" which is a pooled estimate of the standard deviation of $\widehat{\theta}$ based on sd_1 and sd_2 defined in Equation (6)
5	Merge of communities; i.e. all nodes are equally likely to communicate	P_{11}, P_{12}, P_{22}

Table 1. A description for the out-of-control scenarios considered in this simulation study

Window type	Fixed		Fixed-sized moving		Variable-sized moving	
Window size (m)	500	1,000	500	1,000	500	1,000
Control chart						
P_{11}	2.975	2.989	3.030	3.015	3.030	3.020
P_{12}	2.975	2.989	3.020	3.015	3.035	3.010
P_{22}	2.975	3.002	3.020	3.015	3.030	3.010
S	2.975	3.002	3.020	3.015	3.035	3.015

Table 2. The design parameter values of the Shewhart control chart (L_h) that produce an in-control ARL value of 370 when monitoring the DCSBM parameters

Generally, the simulation results show that global changes introduced either in the mean interaction or in the connection variability among the nodes (Tables 5 and 7) can be detected more quickly than local changes (Tables 4 and 6) using the control charts, as expected. In all cases, large changes can also be detected more quickly than small or moderate changes.

As shown, in all of the out-of-control scenarios considered, the Shewhart control chart performs significantly better when its control limits are estimated using a fixed window of networks than when estimated using any of the moving window approaches (fixed or variable). When a window of size $m = 500$ is used, the Shewhart chart designed with a fixed window of networks requires at least half the number of networks required if it is designed with any of the moving window approaches to detect an anomalous behavior in the network. When m increases to 1,000, the differences in the chart performance between the three approaches almost diminish. Moreover, the use of the fixed window approach incorporates a significantly lower variability in the ARL values among different practitioners than that of the moving window approaches. See, for example, Figures 1 and 2 where the ARL distribution of both the Shewhart P_{11}-chart and S-chart for simulation cases (2) and (3), respectively, is highly skewed to the right with a wide spread of values in the moving window approaches comparing with that of the fixed window. Additionally, in case of using the fixed window approach, it is not required to increase the window size to estimate the control limits, as the out-of-control AARL for both $m = 500$ and $m = 1{,}000$ are very close in values. On the other hand, if the fixed-sized or the variable-sized moving window approaches are used, it is recommended to enlarge the window size.

As for the EWMA chart, the simulation results indicate that its performance is indifferent whether the fixed window or the fixed-sized moving window of networks is used in estimating its control limits. This is because both approaches provide almost the same out-of-control AARL values. However, if the variable-sized moving window is used, the EWMA chart is delayed in detecting the out-of-control conditions comparing to the fixed window and fixed-sized moving window approaches, especially for small to moderate window sizes. As an example, Figures 3 and 4 present the out-of-control ARL distribution for both the EWMA P_{11}-chart and S-chart for simulation cases (2) and (3), respectively. As shown, the distribution of the ARL values and its spreading is almost identical whether the fixed window approach or the fixed-sized moving window approach is used. However, the ARL distribution that corresponds to the variable-sized moving window has higher values and more variability. Moreover, it is noticed that increasing the window size (m) has no remarkable effect as well on the performance of the EWMA chart; only if it was not designed using the variable-sized moving window. This is true for all cases except Case (3), in which at small shift size (e.g. $\varepsilon = 0.05$) and moderate λ (e.g. $\lambda = 0.2$), the fixed window approach is better than both of the moving window approaches. Generally, the EWMA chart performance significantly surpasses that of the Shewhart chart.

In literature, it is usually recommended and more preferable to continuously update the Phase I data during the monitoring process to have a better estimation for the control chart

Table 3.
The design parameters values of the EWMA control chart (L_e, λ) that produce an in-control ARL value of 370 when monitoring the DCSBM parameters

Window type	Fixed				Fixed-sized moving				Variable-sized moving			
Window size (m)	500		1,000		500		1,000		500		1,000	
Smoothing parameter (λ)	0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2
Control chart												
P_{11}	2.52	2.85	2.51	2.85	2.408	2.85	2.449	2.85	2.35	2.84	2.42	2.84
P_{12}	2.52	2.85	2.51	2.85	2.408	2.85	2.45	2.85	2.35	2.85	2.43	2.85
P_{22}	2.52	2.85	2.51	2.85	2.408	2.85	2.45	2.85	2.35	2.85	2.44	2.85
S	2.52	2.85	2.51	2.85	2.408	2.85	2.45	2.85	2.35	2.84	2.42	2.85

Control chart type	Shewhart						EWMA											
			Fixed-sized moving		Variable-sized moving		Fixed				Fixed-sized moving				Variable-sized moving			
Window type	Fixed																	
Window size (m)	500	1,000	500	1,000	500	1,000	500		1,000		500		1,000		500		1,000	
Smoothing parameter (λ)			–				0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2
P_{11} – control chart																		
$\varepsilon = 0.01$	58.6	60.2	99.2	75.2	171.3	91.9	12.7	13.8	12.8	13.8	13.6	16.5	13.1	15.1	20.6	43.2	14.0	16.6
$\varepsilon = 0.05$	1.2	1.3	1.3	1.3	1.3	1.3	2.6	1.8	2.6	1.8	2.7	1.8	2.6	1.8	3.0	2.0	2.7	1.8
$\varepsilon = 0.10$	1.0	1.0	1.0	1.0	1.0	1.0	1.5	1.1	1.5	1.1	1.5	1.1	1.6	1.1	1.6	1.1	1.6	1.1

Table 4.
Out-of-control AARL values for the Shewhart and EWMA control charts when $m = 500$ and 1,000 networks assuming fixed, fixed-sized moving and variable-sized moving windows for simulation case (1)

Table 5.
Out-of-control AARL values for the Shewhart and EWMA control charts when $m = 500$ and 1,000 networks assuming fixed, fixed-sized moving and variable-sized moving windows for simulation case (2)

Control chart type	Shewhart					EWMA											
Window type	Fixed		Fixed-sized moving		Variable-sized moving	Fixed				Fixed-sized moving				Variable-sized moving			
Window size (m)	500	1,000	500	1,000	500	1,000	500		1,000		500		1,000		500		1,000	
Smoothing parameter (λ)			–				0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2
Control chart																		
P_{11}																		
$\varepsilon = 0.01$	59.0	60.5	100.9	73.9	167.1	91.6	12.9	14.1	12.6	13.7	13.8	16.8	13.2	15.0	20.7	42.8	14.3	16.7
$\varepsilon = 0.05$	1.3	1.3	1.3	1.3	1.3	1.3	2.6	1.8	2.5	1.8	2.6	1.8	2.6	1.8	3.0	2.0	2.7	1.9
$\varepsilon = 0.10$	1.0	1.0	1.0	1.0	1.0	1.0	1.6	1.1	1.6	1.1	1.5	1.1	1.5	1.1	1.6	1.2	1.5	1.1
P_{12}																		
$\varepsilon = 0.01$	11.1	11.1	12.5	11.8	25.9	12.7	6.3	4.7	6.3	4.7	6.4	4.9	6.2	4.8	7.5	5.7	6.5	5.0
$\varepsilon = 0.05$	1.0	1.0	1.0	1.0	1.0	1.0	1.6	1.1	1.5	1.1	1.5	1.1	1.6	1.1	1.6	1.1	1.5	1.1
$\varepsilon = 0.10$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
P_{22}																		
$\varepsilon = 0.01$	59.5	61.6	98.6	71.2	173.5	91.2	13.2	13.9	12.8	13.7	13.7	16.4	13.0	14.9	20.7	42.0	14.5	16.9
$\varepsilon = 0.05$	1.3	1.3	1.3	1.3	1.3	1.3	2.6	1.8	2.6	1.8	2.6	1.8	2.6	1.8	3.0	2.0	2.7	1.8
$\varepsilon = 0.10$	1.0	1.0	1.0	1.0	1.0	1.0	1.6	1.1	1.5	1.1	1.5	1.1	1.6	1.1	1.7	1.2	1.6	1.1

Control chart type			Shewhart			EWMA														
Window type	Fixed		Fixed-sized moving		Variable-sized moving			Fixed				Fixed-sized moving				Variable-sized moving				
Window size (m)	500	1,000	500	1,000	500	1,000	500		1,000		500		1,000		500		1,000			
Smoothing parameter (λ)			–				0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2		
S – control chart																				
$\tau = 0.05$	140.3	142.3	207.9	177.2	257.0	199.1	32.3	44.9	31.2	43.4	34.6	60.6	32.5	48.4	53.1	108.0	34.4	55.6		
$\tau = 0.10$	41.0	42.1	60.2	47.0	125.0	58.1	12.1	11.3	11.8	11.3	12.2	11.9	12.1	11.6	14.4	14.2	12.5	11.9		
$\tau = 0.25$	3.2	3.2	3.3	3.3	3.7	3.4	4.1	2.8	4.0	2.8	4.1	2.9	4.1	2.9	4.4	2.9	4.1	2.9		

Table 6.
Out-of-control AARL values for the Shewhart and EWMA control charts when $m = 500$ and 1,000 networks assuming fixed, fixed-sized moving and variable-sized moving windows for simulation case (3)

Table 7.
Out-of-control AARL values for the Shewhart and EWMA control charts when $m = 500$ and 1,000 networks assuming fixed, fixed-sized moving and variable-sized moving windows for simulation case (4)

Control chart type			Shewhart			EWMA											
			Fixed-sized moving		Variable-sized moving			Fixed				Fixed-sized moving				Variable-sized moving	
Window type	Fixed																
Window size (m)	500	1,000	500	1,000	500	1,000	500		1,000		500		1,000		500		1,000	
Smoothing parameter (λ)			–				0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2
S – control chart																		
$\tau = 0.05$	45.9	48.0	70.8	55.0	143.3	65.6	12.7	12.3	12.6	12.2	13.0	13.0	12.7	12.7	15.3	16.2	13.0	13.1
$\tau = 0.10$	7.8	8.1	8.7	8.5	13.8	8.7	5.7	4.2	5.6	4.2	5.7	4.3	5.7	4.2	6.3	4.4	5.7	4.2
$\tau = 0.25$	1.1	1.1	1.1	1.1	1.1	1.1	2.3	1.6	2.3	1.6	2.3	1.6	2.3	1.6	2.4	1.6	2.3	1.6

Control chart type	Shewhart					EWMA													
Window type	Fixed		Fixed-sized moving		Variable-sized moving	Fixed				Fixed-sized moving				Variable-sized moving					
Window size (m)	500	1,000	500	1,000	500	1,000	500		1,000		500		1,000		500		1,000		
Smoothing parameter (λ)				–			0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2	0.05	0.2	
Control chart																			
P_{11}	1.2	1.2	1.2	1.2	1.2	1.2	2.7	1.8	2.7	1.8	2.7	1.8	2.7	1.8	3.2	2.0	2.8	1.9	
P_{12}	1.0	1.0	1.0	1.0	1.0	1.0	1.6	1.1	1.6	1.1	1.5	1.1	1.5	1.1	1.6	1.1	1.5	1.1	
P_{22}	1.2	1.2	1.2	1.2	1.2	1.2	2.7	1.8	2.7	1.8	2.8	1.9	2.7	1.8	3.2	2.0	2.8	1.9	

Table 8.
Out-of-control AARL values for the Shewhart and EWMA control charts when $m = 500$ and 1,000 networks assuming fixed, fixed-sized moving and variable-sized moving windows for simulation case (5)

limits, and consequently an efficient Phase II chart performance (see, for example, Xia *et al.*, 2013; De Ketelaere *et al.*, 2015; Azarnoush *et al.*, 2016). Yet, the results on monitoring the parameters of the DCSBM that models social networks show the complete opposite. One of the explanations for this is that the social networks are non-stationary by nature, and hence the criteria of the moving window would let the nodes' behavior close to the current behavior. Accordingly, the control chart would delay in differentiating between the anomalous situation and the typical situation. In addition, this delay in detecting anomalous cases results in including some contaminated samples (undetected out-of-control samples) in the continuously updated Phase I data set, and hence affecting the estimates reliability and accordingly the chart performance, especially for small shifts. A further explanation for the deteriorated performance of the variable moving window case in comparison with that of the fixed window approaches might be due to the use of smaller initial window sizes.

To summarize, the findings show multiple advantages for the use of the EWMA chart over the Shewhart chart when monitoring the DCSBM parameters. Accordingly, it is highly recommended the use of the EWMA chart with small smoothing parameter to detect any shift

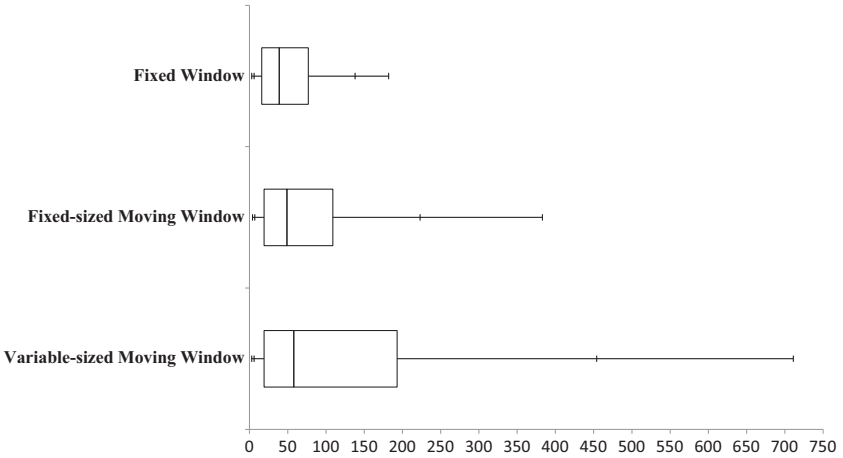

Figure 1.
Out-of-control distribution of the conditional ARL of the Shewhart P_{11}-chart when $m = 500$, and $\epsilon = 0.01$ – simulation case (2). The boxplots show the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the conditional ARL distribution

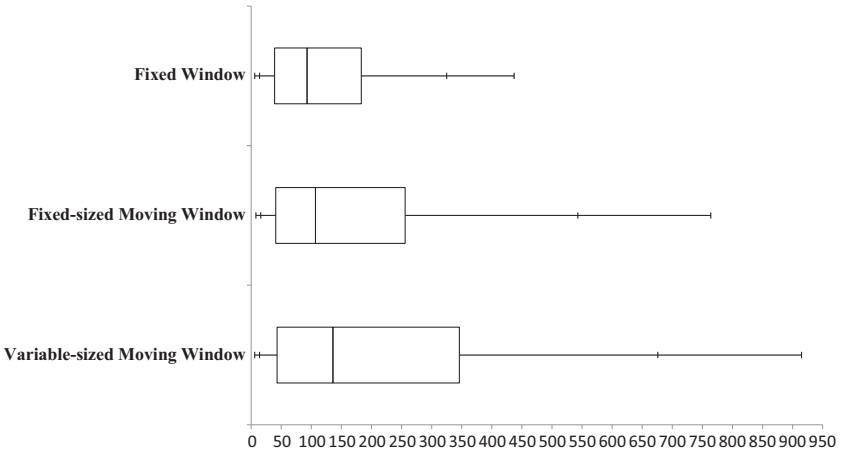

Figure 2.
Out-of-control distribution of the conditional ARL of the Shewhart *S*-chart when $m = 500$, and $\epsilon = 0.05$ – simulation case (3). The boxplots show the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the conditional ARL distribution

size in the DCSBM parameters. To overcome complexity in application, it is also recommended to design the EWMA chart using the fixed window approach. This is because it is simpler and provides almost the same or even better EWMA out-of-control performance than the moving window approaches. Additionally, only small- to moderate-sized window of networks would be sufficient to estimate the EWMA limits. Thus, the practitioner would not wait long till he/she starts monitoring a network.

It is worth to mention that if it is of interest to enlarge the window size in order to assess the performance of the methodology under convenient conditions as recommended by Wilson *et al.* (2019), then it is recommended to design the EWMA chart using a variable-sized moving window. This is because it provides almost the same out-of-control performance as the fixed window approach in such a case. In addition, the practitioner would not wait long before starting monitoring a network as it starts with smaller initial window size of networks (e.g. $m = 400$) and then gradually increase to reach the pre-determined size (e.g. $m = 1{,}000$) networks instead of using a large fixed window size (e.g. $m = 1{,}000$) networks. By that, the practitioner would not be under the risk of including undetected contaminated observations in the Phase I data.

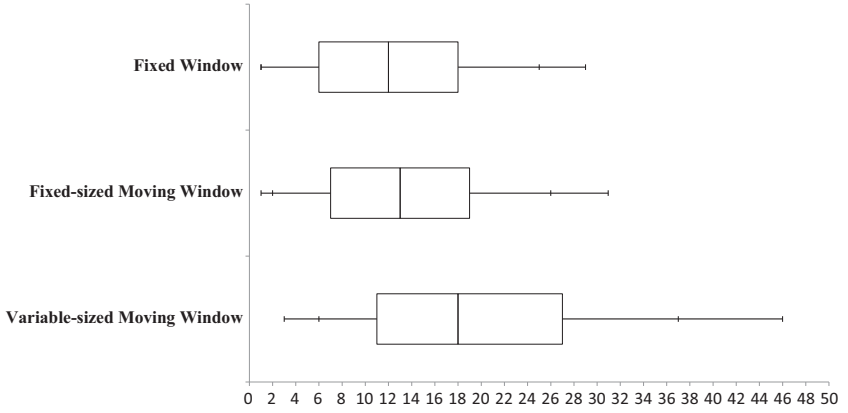

Figure 3. Out-of-control distribution of the conditional ARL of the EWMA P_{11}-chart when $m = 500$, $\lambda = 0.05$, and $\epsilon = 0.01$ – simulation case (2). The boxplots show the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the conditional ARL distribution

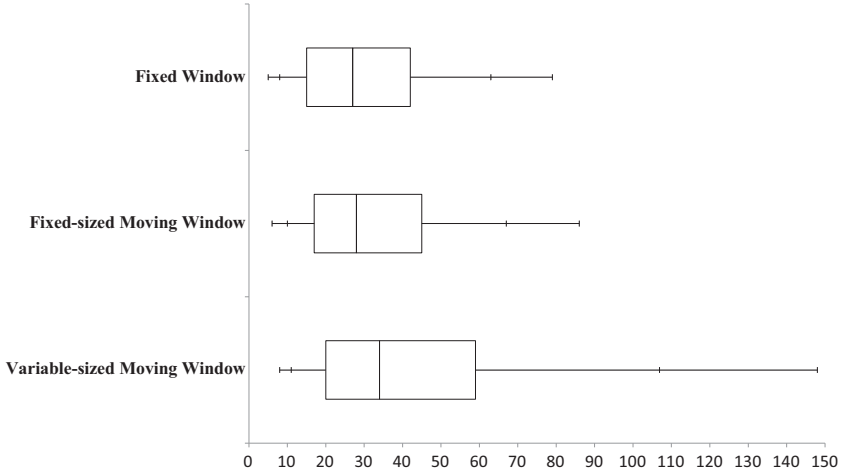

Figure 4. Out-of-control distribution of the conditional ARL of the EWMA S-chart when $m = 500$, $\lambda = 0.05$, and $\epsilon = 0.05$ – simulation case (3). The boxplots show the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the conditional ARL distribution

5. Conclusion

Wilson *et al.* (2019) evaluated the DCSBM using a strategy that combines a parametric model for modeling the random graphs and one of the statistical process monitoring tools. In their study, they monitored the DCSBM parameters using the Shewhart control chart while relying on fixed-sized windows of networks to estimate its control limits. In literature, a usual recommendation is to continuously update the process parameters' estimates with the recent behavior (observations) of the monitored process. This would guarantee a more efficient control chart performance.

Motivated by these recommendations, it is of interest to extend the work of Wilson *et al.* (2019) by conducting a performance comparison on the DCSBM strategy when evaluated using a fixed window, fixed-sized moving window and variable-sized moving window approaches; under different scenarios of changes in the network structure. Furthermore, the Shewhart and EWMA charts are also selected as the SPC tools used in monitoring the DCSBM parameters.

The simulation results show that enhancing the chart detection power for out-of-control conditions by continuously updating the parameters' estimates cannot be generally extended to social network monitoring. Monitoring the DCSBM parameters showed that the fixed window approach provides better out-of-control performance to the Shewhart chart than the moving window approaches, with sufficiently small to moderate window size. On the other hand, the EWMA chart performance is almost indifferent whether the fixed window or fixed-sized moving window approach is used, but highly deteriorates when the variable-sized moving window approach is used. One of the explanations is that social networks behavior is different from other industrial/manufactory processes; as the former is non-stationary by nature.

We recommend using the EWMA chart over the Shewhart chart while monitoring networks modeled with DCSBM for detecting any expected shift size. Also, we recommend the use of the fixed window approach, for simplicity reasons, with a small to moderate window size or the variable-sized moving window approach, for statistical reasons, with a large window size. In future work, it would be useful to evaluate this framework that combines a parametric model for modeling the random graphs and one of the statistical process monitoring tools under more different realistic models and different types of control charts.

References

Azarnoush, B., Paynabar, K., Bekki, J. and Runger, G. (2016), "Monitoring temporal homogeneity in attributed network streams", *Journal of Quality Technology*, Vol. 48 No. 1, pp. 28-43.

Bickel, P.J. and Chen, A. (2009), "A nonparametric view of network models and Newman–Girvan and other modularities", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 106 No. 50, pp. 21068-21073.

De Ketelaere, B., Hubert, M. and Schmitt, E. (2015), "Overview of PCA-based statistical process-monitoring methods for time-dependent, high-dimensional data", *Journal of Quality Technology*, Vol. 47 No. 4, pp. 318-335.

Erdös, P. and Rényi, A. (1959), "On random graphs", *Publicationes Mathematicae*, Vol. 6, pp. 290-297.

Fienberg, S.E. and Wasserman, S.S. (1981), "Categorical data analysis of single sociometric relations", *Sociological Methodology*, Vol. 12, pp. 156-192.

Holland, P.W., Laskey, K.B. and Leinhardt, S. (1983), "Stochastic blockmodels: first steps", *Social Networks*, Vol. 5 No. 2, pp. 109-137.

Hoppe, B. and Reinelt, C. (2010), "Social network analysis and the evaluation of leadership networks", *The Leadership Quarterly*, Vol. 21 No. 4, pp. 600-619.

Karrer, B. and Newman, M.E.J. (2011), "Stochastic block models and community structure in networks", *Physical Review E*, Vol. 83 No. 1, p. 016107.

Landherr, A., Friedl, B. and Heidemann, J. (2010), "A critical review of centrality measures in social networks", *Business and Information Systems Engineering*, Vol. 2 No. 6, pp. 371-385.

Marchette, D. (2012), "Scan statistics on graphs", *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 4 No. 5, pp. 466-473.

Miller, B.A., Arcolano, N. and Bliss, N.T. (2013), "Efficient anomaly detection in dynamic, attributed graphs: emerging phenomena and big data", *Proceedings of 2013 IEEE International Conference on Intelligence and Security Informatics*, IEEE, Seattle, WA, USA, pp. 179-184.

Priebe, C.E., Conroy, J.M., Marchette, D.J. and Park, Y. (2005), "Scan statistics on Enron graphs", *Computational and Mathematical Organization Theory*, Vol. 11 No. 3, pp. 229-247.

Roberts, S.W. (1959), "Control chart tests based on geometric moving averages", *Technometrics*, Vol. 1 No. 3, pp. 239-250.

Savage, D., Zhang, X., Yu, X., Chou, P. and Wang, Q. (2014), "Anomaly detection in online social networks", *Social Networks*, Vol. 39 No. 1, pp. 62-70.

Shewhart, W.A. (1924), "Some applications of statistical methods to the analysis of physical and engineering data", *Bell System Technical Journal*, Vol. 3 No. 1, pp. 43-87.

Snijders, T. and Nowicki, K. (1997), "Estimation and prediction for stochastic blockmodels for graphs with latent block structure", *Journal of Classification*, Vol. 14 No. 1, pp. 75-100.

Wasserman, S. and Pattison, P. (1996), "Logit models and logistic regressions for social networks: i. an introduction to Markov graphs and p*", *Psychometrika*, Vol. 61 No. 3, pp. 401-425.

Wilson, J.D., Stevens, N.T. and Woodall, W.H. (2019), "Modeling and detecting change in temporal networks via the degree corrected stochastic block model", *Quality and Reliability Engineering International*, Vol. 35 No. 5, pp. 1363-1378.

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*, Vol. 49 No. 3, pp. 354-365.

Xia, L., Chu, J. and Geng, Z. (2013), "Process monitoring based on improved recursive PCA methods by adaptive extracting principal components", *Transactions of the Institute of Measurement and Control*, Vol. 35 No. 8, pp. 1024-1045.

Yan, X., Shalizi, C., Jensen, J.E., Krzakala, F., Moore, C., Zdeborová, L., Zhang, P. and Zhu, Y. (2014), "Model selection for degree-corrected block models", *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2014 No. 5, p. 05007.

Zhao, Y., Levina, E. and Zhu, J. (2012), "Consistency of community detection in networks under degree-corrected stochastic block models", *The Annals of Statistics*, Vol. 40 No. 4, pp. 2266-2292.

Zhao, M.J., Driscoll, A.R., Sengupta, S., Stevens, N.T., Fricker, R.D., Jr and Woodall, W.H. (2018a), "The effect of temporal aggregation level in social network monitoring", *PLoS ONE*, Vol. 13 No. 12, p. e0209075.

Zhao, M.J., Driscoll, A.R., Sengupta, S., Fricker, R.D., Jr, Spitzner, D.J. and Woodall, W.H. (2018b), "Performance evaluation of social network anomaly detection using a moving window-based scan method", *Quality and Reliability Engineering International*, Vol. 34 No. 8, pp. 1699-1716.

Corresponding author
Mahmoud A. Mahmoud can be contacted at: mamahmou@feps.edu.eg