*The Korean Journal of Applied Statistics*
2021, Vol. 34, No. 1, 25–38

DOI: http://dx.doi.org/10.5351/KJAS.2021.34.1.025

# Self-starting monitoring procedure for the dynamic degree corrected stochastic block model

Joo Weon Lee[a], Jaeheon Lee[1,a]

[a]Department of Applied Statistics, Chung-Ang University

**Abstract**

Recently the need for network surveillance to detect abnormal behavior within dynamic social networks has increased. We consider a dynamic version of the degree corrected stochastic block model (DCSBM) to simulate dynamic social networks and to monitor for a significant structural change in these networks. To apply a control charting procedure to network surveillance, in-control model parameters must be estimated from the Phase I data, that is from historical data. In network surveillance, however, there are many situations where sufficient relevant historical data are unavailable. In this paper we propose a self-starting Shewhart control charting procedure for detecting change in the dynamic networks. This procedure can be a very useful option when we have only a few initial samples for parameter estimation. Simulation results show that the proposed procedure has good in-control performance even when the number of initial samples is very small.

Keywords: control charts, network surveillance, self-starting, social network, statistical process control

## 1. 서론

최근 컴퓨터와 인터넷의 발전에 따라 다수의 연결망(network) 데이터가 생성 되어지고 있다. 예를 들어, 사회 연결망(social network)인 트위터(twitter), 인스타그램(instagram), 페이스북(facebook), 그리고 개인 및 회사에서 보내는 이메일(e-mail) 등은 각기 다양한 분야에서 의사소통을 위해 매일 다량으로 쏟아지고 있다. 이뿐만 아니라 연결망은 뇌과학 분야에서의 신경 연결망과 개인 컴퓨터와 서버 시스템의 연결인 컴퓨터 연결망 등으로 사용되고 있다.

기존의 연결망 분석에는 시간에 영향을 받지 않는 과거 데이터(historical data) 및 스냅샷 데이터(snapshot data)가 주로 사용되어져 왔다. 그러나 연결망은 시간에 따라 계속 변화하기 때문에, 최근에는 시간에 따라 변화하는 동적 연결망(dynamic network)에 대한 연구가 많아지고 있다.

이러한 동적 연결망에서 비정상적 변화를 탐지하는 것을 연결망 감시(network surveillance)라고 한다. 연결망 감시의 주된 목적은 비정상적 변화가 발생할 경우 가능한 한 빨리 이를 탐지하는 것으로, 이를 위해 통계적 공정 모니터링(statistical process monitoring) 절차를 사용할 수 있다. 사회 연결망 감시에서 많이 사용되어진 연결망으로는 2001년 말 파산 신청을 한 미국의 에너지회사 엔론(Enron) 구성원들의 이메일 연결망이 있다 (Priebe 등, 2005; Shetty와 Adibi, 2005). 이 연결망에서는 회사가 파업 신청을 하기 직전 회사 간부들 간에 비정상적으로 많은 이메일이 교환되어진 것을 알 수 있다. 또한 2010년 말 튀니지에서 시작되어 아랍의

This research was supported by the Chung-Ang University Research Scholarship Grants in 2019.
[1] Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: jaeheon@cau.ac.kr

중동 국가와 북아프리카로 확산된 반정부 시위인 아랍의 봄(Arab Spring)의 주요 인물들이 Facebook에서 서로 연락한 연결망과 테러 집단인 알 카에다(Al Qaeda)의 테러 사건 이전과 이후의 연결망 등에서 비정상적 변화가 발생한 것을 발견할 수 있다. 따라서 사회의 주요한 사건의 발생 직전에 사회 연결망에서 평소와 다른 비정상적 변화가 있었음을 알 수 있으며, 이를 빠르게 탐지할 수 있는 절차에 대한 연구는 그 중요성이 점점 더 커지고 있다.

최근 통계적 공정 모니터링 절차를 사용하여 동적 연결망을 감시하는 연구가 활발히 진행되고 있다. 예를 들어 Wilson 등 (2019)은 동적 연결망 모형인0 degree corrected stochastic block model (DCSBM)에서 비정상적인 변화를 탐지하기 위해 Shewhart 관리도를 사용하는 절차를 제안하였다. 이때 개인이 속한 커뮤니티에 부분적 변화 또는 전체적인 변화가 발생한 경우 이를 탐지하는 성능에 대해 연구하였다. Yu 등 (2018)은 Hotelling의 $T^2$ 관리도를 사용하여 개인의 경향성(propensity)을 모니터링하는 다변량 절차를 제안하고, 제안한 절차의 효율을 이전 연구의 결과와 비교하였다. 그리고 Yu 등 (2020)은 연결망의 요약 통계량에 기초한 방법, Priebe 등 (2005)이 제안한 스캔 방법(scan method), 그리고 모형에 기초한 방법들에 대해 그 성능을 비교하는 연구를 수행하였다.

연결망 감시를 수행하는 통계적 공정 모니터링의 대표적인 도구에는 관리도(control chart)가 있다. 시간에 따라 관측되는 연결망을 감시하는 제2국면(Phase II)을 진행하기 위해서는 관리한계를 설정해야 하는데, 이를 위해서 제1국면(Phase I)에서 일정 수의 표본을 추출하여 관리상태일 때의 모형 모수를 추정하는 절차가 필요하다. 이때 우리가 원하는 관리상태의 성능을 만족하기 위해서는 일반적으로 많은 수의 제1국면 표본을 사용해야 한다. 그러나, 생산 공정과 달리 연결망의 감시에서는 많은 수의 제1국면 연결망 표본을 얻기 어려운 경우가 대부분이다.

생산 공정을 모니터링할 때, 충분한 제1국면의 표본을 얻기 어려운 경우 사용하는 절차로 자기출발(self-starting) 관리도 절차가 있다. 이 절차는 최소한의 제1국면 표본을 사용하여 공정 모수를 추정하고 관리한계를 설정한 후, 관측되는 표본을 사용하여 공정 모수와 관리한계를 업데이트 시키는 방법이다. 따라서 자기출발 절차를 사용하면 초기에 많은 수의 표본이 없이 빠르게 관리도를 사용하여 공정을 모니터링 할 수 있다는 장점이 있다. 생산 공정에 적용하는 자기출발 절차에 대해 다음과 같은 연구들이 있다. 초기에 Hawkins (1987), Quesenberry (1991a, 1991b, 1995), 그리고 Hawkins와 Olwell (1998) 등이 단일변량 관리도에 대한 절차를 제안하였고, Sullivan과 Jones (2002), Hawkins와 Maboudou-Tchao (2007), 그리고 Capizzi과 Masarotto (2010)는 다변량 관리도에 대한 절차를 제안하였다.

관리도의 성능을 평가하기 위해 일반적으로 평균런길이(average run length; ARL)를 사용하는데, 이때 런길이는 공정에서 이상신호가 발생할 때까지 추출한 표본의 수를 나타낸다. ARL은 공정 모수에 의존하는 값인데, 제1국면에서 이를 추정할 경우 실무자에 따라 추출하는 표본이 달라지기 때문에 추정된 모수, 관리한계, 그리고 ARL값도 서로 달라지게 된다. 따라서 최근 관리도에 대한 제1국면 성능을 평가할 때, 관리상태에서의 평균런길이의 평균(average of the ARL; AARL)과 표준편차(standard deviation of the ARL; SDARL)를 사용하는 연구가 많다. 즉, 관리상태에서의 AARL은 만족하기를 희망하는 ARL값($ARL_0$)과 근접하면서 SDARL이 너무 크지 않도록 제1국면의 표본 수를 결정하는 것이 바람직하다는 것이며, Zhang 등 (2013)은 그 기준으로 SDARL값이 주어진 $ARL_0$의 10% 이내가 되도록 표본의 수를 결정하는 것이 좋다고 제안하였다.

생산 공정에 대한 자기출발 절차의 연구에서 AARL과 SDARL을 측도로 사용하여 그 성능을 평가한 연구는 다음과 같다. Keefe 등 (2015)은 정규분포를 따르는 데이터에 대해 Shewhart의 $X$ 관리도와 CUSUM(cumulative sum) 관리도를 사용한 자기출발 절차의 성능을 평가하였고, Shen 등 (2016)은 다양한 모집단 크기를 갖고 포아송 분포를 따르는 데이터에 대한 자기출발 관리도 절차를 제안하고 그 성능을 평가하였다. 또한 Lee 등 (2018)은 Keefe 등 (2015)과 유사한 상황에서 자기출발 절차를 exponentially weighted moving average (EWMA) 관리도와 $\bar{X}$ 관리도에 적용하여 그 성능을 비교하였다.

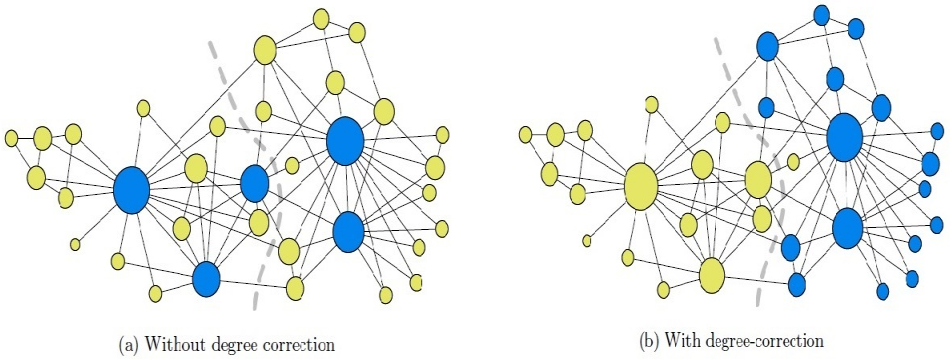


Figure 1: *Divisions of the karate club network found using the (a) uncorrected and (b) corrected blockmodels.*

이 논문은 동적 DCSBM을 가정한 사회 연결망의 감시에서 많은 수의 제1국면의 연결망 표본을 확보하기 어려운 경우 사용할 수 있는 자기출발 관리도 절차를 제안하고, 제안된 절차의 관리상태 성능을 평가하고자 한다. 이때 성능 평가를 위해 관리상태에서의 AARL과 SDARL을 그 측도로 사용할 것이다.

이 논문의 구성은 다음과 같다. 2장에서는 이 논문에서 가정하는 사회 연결망 모형인 DCSBM을 소개한다. 3장에서는 DCSBM에 적용하는 자기출발 관리도 절차를 제안하고, 4장에서는 모의실험의 절차, 설계, 그리고 수행한 결과를 설명한다. 마지막 5장에서는 이 논문에 대한 결론을 제시한다.

## 2. DCSBM

이번 장에서는 최근 사회 연결망의 모형으로 자주 사용되는 Karrer과 Newman (2011)이 제안한 DCSBM을 소개하고자 한다. $G_t = (V, E)$를 시점 $t$에서의 사회 연결망이라 할 때, $V = \{1, \ldots, n\}$은 개인을 나타내는 노드(node)이고 $E$는 가중엣지(weighted edge)이다. 여기서 가중엣지(weighted edge)란 시점 $t$에서 노드 $i$와 $j$의 의사소통(communication)의 수를 나타낸다. 한편 비가중엣지(unweighted edge)는 시점 $t$에서 노드 $i$와 $j$의 의사소통의 유무를 1과 0으로만 나타낸다. 또한 연결망은 의사소통의 방향성을 고려하는가 안하는가에 따라 방향성 연결망(directed network)과 무방향성 연결망(undirected network)으로 구분하는데, 이 논문에서는 무방향성 연결망을 가정한다. $c$의 주요 특징은 커뮤니티 구조(community structure)와 연결정도의 이질성(degree heterogeneity)을 고려하는 것이다. 연결망 $G$의 노드들은 $k$ ($\geq 1$)개의 상호 배타적인 꼭짓점(disjoint vertex) 집합인 커뮤니티로 나누어지며, 이를 식으로 표현하면 $V = V_1 \cup V_2 \cup \cdots \cup V_k$가 된다. 동일한 커뮤니티 안에 속한 노드들 간의 엣지 크기는 서로 다른 커뮤니티에 속한 노드들과의 엣지 크기보다 큰 경우가 일반적이다. 벡터 $\mathbf{c} = (c_1, c_2, \ldots, c_n)$는 각 노드에 대한 커뮤니티 레이블(community label)이고, 여기서 $c_i$는 노드 $i$가 속한 커뮤니티를 나타낸다. 또한 각 노드들은 소속된 커뮤니티 이외에 각 노드만의 특성인 연결정도(degree)에 따라 특징지울 수 있다. 즉, DCSBM에서는 노드가 속한 커뮤니티와 각 노드의 연결정도의 특성으로 연결망을 설명하고 있다.

DCSBM에 대한 이해를 돕기 위해 Karrer과 Newman (2011)에 있는 가라데(karate) 클럽 연결망을 모형에 적합시킨 결과를 Figure 1에 도시하였다. 이 연결망은 미국의 어떤 대학 내 34명의 가라데 클럽 회원의 교우관계 패턴을 나타낸 것이다. 그런데 이 클럽은 내부의 의견 차이 등의 이유로 2개의 집단으로 나누어져 있는 상태이고, 회원들이 속한 집단을 Figure 1에서는 점선으로 구분하였다. 이 연결망에 대해 커뮤니티의 수는 2개로 설정하여 모형에 적합시켰는데, (a)는 커뮤니티의 특성만을 고려한 결과이고, (b)는 커뮤니티의 특성 이외에 각 노드의 연결정도를 함께 고려한 DCSBM에 적합시킨 결과이다. 각 노드에 적합된 커뮤니티는 노란

색과 파란색으로 구분되어 있다. (a)와 (b)를 비교할 때, DCSBM으로 적합한 (b)의 결과가 실제 연결망을 더 잘 설명한다는 것을 알 수 있다.

연결망은 인접행렬(adjacency matrix)을 사용하여 수리적으로 표현할 수 있다. 노드 $i$와 $j$의 엣지를 나타내는 $W_{ij}$를 원소로 하는 인접행렬을 $W$라 할 때, $W$는 $n \times n$ 행렬이고 이 논문에서는 무방향성 연결망을 고려하기 때문에 $W_{ij} = W_{ji}$ $(i \neq j)$인 대칭행렬이 된다. 또한 일반적으로 동일한 노드간의 의사소통, 즉 자가루프(self-loop)는 허용하지 않기 때문에, $W$의 대각원소들은 모두 0이 된다. Karrer과 Newman (2011)는 DCSBM에서 노드 $i$와 $j$의 가중엣지 $W_{ij}$는 서로 독립이고 평균이 $\lambda_{ij}$인 포아송 분포(Poisson distribution)를 따른다고 가정하였다. 이때 $\lambda_{ij}$는

$$\lambda_{ij} = \theta_i \theta_j P_{c_i c_j}$$

이다. 여기서 $\theta_i$ $(i = 1, \ldots, n)$는 노드 $i$의 연결정도 모수이고 각 노드의 연결정도의 특성을 나타내며, $P_{c_i c_j}$ $(i, j = 1, \ldots, n)$는 노드 $i$와 $j$가 속하는 커뮤니티의 의사소통에 대한 경향성을 나타낸다. 이 값이 클수록 소속된 커뮤니티 간의 의사소통의 경향이 큰 것을 나타내며, 일반적으로 $c_i = c_j$인 경우가 그렇지 않은 경우에 비해 큰 값을 갖는다.

이때 각 노드의 연결정도의 모수 벡터 $\theta = (\theta_1, \ldots, \theta_n)$는 연결망의 구별성(identifiability)을 위하여 다음과 같은 제약조건을 만족하도록 설정한다.

$$\sum_{i:C_i=r} \theta_i = n_r, \quad r = 1, \ldots, k, \tag{2.1}$$

여기서 $n_r$은 커뮤니티 $r$에 속한 노드의 수를 나타내기 때문에, 이 제약조건은 어떤 커뮤니티에 속한 노드들의 모수 $\theta_i$의 합을 그 커뮤니티의 크기와 같게 한다는 제약인 것이다. Wilson 등 (2019)과 Yu 등 (2020)도 이와 동일한 제약조건을 사용하였고, Karrer과 Newman 등 (2011)과 Yu 등 (2018)은 이와 유사하게 $\theta_i$의 합을 1로 지정하는 제약조건을 사용하였다.

이제 비정상적 변화가 없는 관리상태와 변화가 발생한 이상상태에서의 DCSBM에 대해 설명하고자 한다. 먼저 $P(G \mid \mathbf{c}, \theta, \mathbf{P})$는 커뮤니티 레이블 벡터 $\mathbf{c}$, 노드의 연결정도의 모수 벡터 $\theta$, 그리고 커뮤니티의 경향성 행렬 $\mathbf{P}$를 갖는 연결망 $G$의 분포를 나타낸다고 하자. 여기서 $\mathbf{P}$는 원소가 $P_{r,s}$ $(r, s = 1, \ldots, k)$인 $k \times k$ 행렬이다. 시점 $t^*$에서 비정상적 변화가 발생한다고 가정할 때, 시점 $t$에서의 연결망 $G_t$는 다음과 같이 모델링할 수 있다.

$$G_t = \begin{cases} P(G \mid \mathbf{c}, \theta, \mathbf{P}), & t < t^*, \\ P(G \mid \mathbf{c}, \theta^*, \mathbf{P}^*), & t \geq t^*, \end{cases} \tag{2.2}$$

여기서 $\theta^*$와 $\mathbf{P}^*$는 비정상적 변화가 발생한 후의 모수 벡터 $\theta$와 행렬 $\mathbf{P}$를 나타낸다.

이 논문에서 가정한 식 (2.2)의 모형에서 노드의 커뮤니티 레이블 $\mathbf{c}$가 알려져 있는 않은 경우 커뮤니티 구조를 추정하는 절차를 먼저 수행해야 한다. 커뮤니티 구조의 추정에 대해서는 Qin과 Rohe (2013)와 Sengupta와 Chen (2015) 등을 비롯한 많은 연구들이 있으며, 이에 대해서는 이 논문에서 초점을 맞추고 있는 내용의 범위 밖이기 때문에 $\mathbf{c}$는 알려져 있으며 시간에 따라 변화가 없는 경우를 가정한다.

## 3. 자기출발 관리도 절차

이제 시간에 따라 변화하는 동적 DCSBM을 가정한 사회 연결망의 감시에 사용할 수 있는 자기출발 관리도 절차를 제안하고자 한다.

먼저 시점 $t$에서의 인접행렬은 $W_t$이고 커뮤니티 $r$에 속하는 노드의 수는 $n_r$ $(r = 1, 2, \ldots, k)$이라고 하자. 먼저 절차를 적용하기 전에 $W_t$를 동일한 커뮤니티에 속하는 노드들끼지 묶어서 $W_t$를 재배열한다. 재배열을

하고 나면 $W_t$는 각 커뮤니티에 따른 소행렬로 분할된다. 즉, 처음 $n_1 \times n_1$ 소행렬 $R_t(1,1)$의 원소는 첫 번째 커뮤니티에 속하는 노드들 간의 가중엣지를 나타내며, 다음의 $n_1 \times n_2$ 소행렬 $R_t(1,2)$의 원소는 첫 번째 커뮤니티에 속하는 노드와 두 번째 커뮤니티에 속하는 노드 간의 가중엣지를 나타낸다. 이를 일반화할 경우 소행렬 $R_t(r,s)$ $(r,s = 1,2,\ldots,k)$는 차원이 $n_r \times n_s$이고, 그 원소는 $r$번째 커뮤니티에 속하는 노드와 $s$번째 커뮤니티에 속하는 노드 간의 가중엣지를 나타낸다. 이 논문에서 가정하는 무방향성 연결망에서는 $R_t(s,r) = R_t(r,s)^T$가 되기 때문에, $r \le s$인 경우만 고려하기로 한다. 따라서 고려하는 커뮤니티 소행렬의 수는 모두 $k(k+1)/2$개가 된다.

소행렬 $R_t(r,s)$에서 의사소통의 비정상적 변화를 탐지하기 위해, 다음과 같이 정의되는 통계량 $Y_t(r,s)$를 고려하자.

$$Y_t(r,s) = \sum \sum_{\{(i,j):W_t(i,j)\in R_t(r,s)\}} W_t(i,j)$$

즉, $Y_t(r,s)$는 커뮤니티 $r$과 $s$에 속하는 노드들의 가중엣지의 합을 나타내며, 서로 독립이고 평균이 $\lambda_t(r,s)$인 포아송 분포를 따름을 쉽게 알 수 있다. 여기서 $\lambda_t(r,s)$는

$$\lambda_t(r,s) = \sum \sum_{\{(i,j):W_t(i,j)\in R_t(r,s)\}} \lambda_{ij}.$$

포아송 분포를 따르는 생산 공정의 데이터를 모니터링하기 위해 Quesenberry (1991b)는 포아송 $Q$ 관리도 (Poisson $Q$ chart)라는 자기출발 관리도 절차를 제안하였다. 이 논문에서 각 커뮤니티 소행렬 $R_t(r,s)$를 모니터링하기 위해서, 포아송 $Q$ 관리도 절차의 아이디어를 통계량 $Y_t(r,s)$에 적용하였다. $W_1, W_2, \ldots W_m$ $(m \ge 1)$을 시간에 따라 관측된 초기 연결망의 인접행렬이라 할 때, 시점 $t \ge m+1$에서 관리도의 통계량 $Q_t(r,s)$를 다음과 같이 정의한다.

$$U_t(r,s) = B\left(Y_t(r,s); T_t(r,s), \frac{n_t(r,s)}{N_t(r,s)}\right) = \sum_{y=0}^{Y_t(r,s)} \binom{T_t(r,s)}{y} \left(\frac{1}{t}\right)^y \left(1-\frac{1}{t}\right)^{T_t(r,s)-y} \tag{3.1}$$

$$Q_t(r,s) = \Phi^{-1}(U_t(r,s)) \tag{3.2}$$

여기서 $B(\cdot\,; n, p)$는 모수 $n$과 $p$를 가지는 이항분포의 누적분포함수(cumulative distribution function)이고 $\Phi(\cdot)$는 표준정규분포의 누적분포함수를 나타내며, $T_t(r,s) = \sum_{i=1}^{t} Y_t(r,s)$이다. 또한 $n_t(r,s)$는 소행렬 $R_t(r,s)$의 원소의 총 개수이고 $N_t(r,s)$는 $t$시점까지의 이를 누적한 개수라고 할 때, 이 논문에서 커뮤니티에 속한 노드는 시간에 따라 변화가 없음을 가정하기 때문에 다음과 같은 관계가 성립한다.

$$\frac{n_t(r,s)}{N_t(r,s)} = \frac{1}{t}$$

식 (3.1)과 (3.2)와 같이 통계량을 정의할 경우, Quesenberry (1991b)는 통계량 $Q_t(r,s)$ $(t = m+1, m+2, \ldots)$가 서로 독립이고 근사적으로 표준정규분포를 따르는 확률변수가 된다는 사실을 보였다. 좀 더 상세하게 설명하면, 포아송 분포를 따르는 통계량 $Y_t(r,s)$의 누적확률을 이항분포로 근사시켜 계산하고, 이 확률과 동일한 값을 갖는 표준정규분포의 백분위수(percentile)인 $Q_t(r,s)$로 변환시키는 것이다.

이 논문에서는 시점 $t$에서 식 (3.2)로 계산되는 $k(k+1)/2$개의 통계량 $Q_t(r,s)$의 최댓값과 최솟값만을 사용하여 사회 연결망에 비정상적 변화가 발생했는지를 판단하는 절차를 제안한다. 즉, 다음과 같은 경우에 비정상적 변화가 발생했다는 신호를 주는 것이다.

$$\max_{r\le s} Q_t(r,s) > h \quad \text{또는} \quad \min_{r\le s} Q_t(r,s) < -h.$$

위의 식은

$$\max_{r\le s} |Q_t(r,s)| > h$$

로 간단하게 표현할 수 있다. 여러 개의 통계량을 사용하여 공정을 모니터링하는 관리도에서 주어진 전체 오경보율(false alarm rate) $\alpha$를 만족하는 관리한계를 설정하는 문제는 간단하지 않다. 그러나 이 논문에서 제안된 통계량 $Q_t(r,s)$는 서로 독립이고 근사적으로 표준정규분포를 따르기 때문에, $k(k+1)/2$개 통계량의 최댓값(또는 최솟값)에 대한 관리한계 $h$를 주어진 $\alpha$에 대해

$$h = \Phi^{-1}\left(\left(1-\frac{\alpha}{2}\right)^{\frac{2}{k(k+1)}}\right) \tag{3.3}$$

로 설정할 수 있다.

또한 이 논문에서 제안된 자기출발 절차는 $W_{ij}$가 비가중엣지인 경우에도 적용할 수 있음을 참고하기 바란다. 이 경우 통계량 $Y_t(r,s)$는 이항분포를 따르기 때문에 Quesenberry (1991a)가 제안한 이항 Q 관리도(binomial Q chart)를 사용하여 절차를 적용할 수 있다.

## 4. 모의실험

동적 DCSBM 연결망에 대해 제안된 자기출발 관리도 절차의 관리상태 성능을 알아보기 위해 모의실험을 실시하였다. 이 논문에서는 관측된 초기 연결망의 수 $m$에 따라 절차의 성능이 어떻게 변화하는지에 촛점을 맞추어 모의실험을 진행하였는데, 모의실험의 절차, 설계, 수행한 결과의 순서로 그 내용을 설명하고자 한다.

### 4.1. 모의실험 절차

서론에서 언급한 바와 같이 제1국면의 초기 연결망은 관측하는 실무자에 따라 달라지기 때문에, 모의실험에서 서로 다른 초기 연결망을 생성하여 ARL의 평균인 AARL과 표준편차인 SDARL을 계산하고 이 측도들을 사용하여 관리상태의 성능을 판단하려 한다. 모의실험을 진행한 절차를 요약하면 다음과 같다.

단계 1. 먼저 분포 $P(G\,|\,\mathbf{c},\theta,\mathbf{P})$에서 $m$개의 서로 독립인 초기 연결망 $G_1, G_2, \ldots, G_m$을 생성하고, 시점 $t \le m$에 대해 통계량 $Y_t(r,s)$와 $T_t(r,s)$를 계산한다 $(r,s = 1,\ldots,k,\ r \le s)$.

단계 2. $P(G\,|\,\mathbf{c},\theta,\mathbf{P})$로부터 새로운 연결망 $G_t$를 하나 생성하여 통계량 $Y_t(r,s)$와 $T_t(r,s)$ $(t \ge m+1)$를 계산하고, 자기출발 관리도의 통계량인 $\max_{r\le s}|Q_t(r,s)|$를 계산한다.

단계 3. 주어진 $\alpha$에 대해 식 (3.3)에 의해 관리한계 $h$를 계산하고, $\max_{r\le s}|Q_t(r,s)| \le h$인 경우 런길이를 1씩 증가시키고 단계 2를 통계량이 관리한계를 벗어나서 이상신호를 줄 때까지 반복한다.

단계 4. 만일 이상신호가 발생하면 런길이를 기록하고, 단계 1에서 생성한 동일한 초기 연결망을 사용하여 단계 2와 3을 10,000번 반복한다. 이를 통하여 런길이의 평균인 ARL을 계산한다.

단계 5. $m$개씩 1,000개의 초기 연결망을 생성하여 각 초기 연결망에 대해 단계 1-4를 반복한 후 ARL을 계산하고, ARL들의 평균인 AARL과 표준편차인 SDARL을 계산한다.

위의 절차를 요약하면, $m$개의 초기 연결망으로부터 10,000개의 런길이를 구하여 하나의 ARL을 계산하고, 이런 초기 연결망을 1,000개 생성하여 1,000개의 ARL을 계산하였다. 그리고 이를 이용하여 평균인 AARL과 표준편차인 SDARL을 계산하고, 초기 연결망의 수 $m$에 따른 자기출발 관리도 절차의 관리상태 성능을 평가하였다.

Table 1: Simulation settings for the DCSBM parameters

| Case | $n$ | | **P** | $\delta$ | |
|---|---|---|---|---|---|
| 1 | 40 | | 0.2 | 0.5 | |
| 2 | 100 | | 0.1 | 0.5 | |
| Case | $n_1$ | $n_2$ | **P** | $\delta_1$ | $\delta_2$ |
| 3 | 20 | 20 | $\begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$ | 0.5 | 0.5 |
| 4 | 10 | 30 | $\begin{pmatrix} 0.3 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$ | 0.5 | 0.75 |
| 5 | 50 | 50 | $\begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$ | 0.5 | 0.5 |
| 6 | 25 | 75 | $\begin{pmatrix} 0.3 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$ | 0.5 | 0.75 |

## 4.2. 모의실험 설계

모의실험의 설계에서 사용한 모수는 다음과 같다. 먼저 초기 연결망의 수 $m$은 1, 2, 3, 5, 10, 20, 50, 100, 200, 500을 사용하였고, 오경보율 $\alpha$는 0.0027로 두어 희망하는 관리상태의 ARL값을 $\text{ARL}_0 = 370.4$로 설정하였다.

커뮤니티의 수는 하나인 경우와 두 개인 경우를 고려하였고 (즉, $k = 1, 2$), 전체 노드의 수는 $n = 40$과 $n = 100$으로 설정하여 연결망의 크기에 따른 성능을 평가하였다. $k = 2$인 경우 각 커뮤니티에 속하는 노드의 수는 $n_2 = n_1$인 경우와 $n_2 = 3n_1$인 경우를 고려하였다.

커뮤니티의 경향성 모수인 **P**에 대해서는, 연결망의 크기가 커지면 연결망의 밀도(density)가 높아지기 때문에, 비교 가능한 결과를 위해 연결망의 크기에 따라 **P**를 적절히 조정하였다. 특히 $k = 2$이고 $n_1 < n_2$인 경우, 규모가 작은 커뮤니티의 경향성을 더 큰 값으로 사용하였다 (즉, $P_{11} > P_{22}$).

마지막으로 각 노드의 연결정도 모수 $\theta_i$는 먼저 균등분포 $U(1 - \delta_{c_i}, 1 + \delta_{c_i})$에서 무작위로 $\theta_i'$를 생성한 후 제약조건인 식 (2.1)을 만족하도록 다음과 같이 변환하였다.

$$\theta_i = \frac{n_{c_i}\theta_i'}{\sum_{j:c_j=c_i}\theta_j'},$$

여기서 $\theta_i'$은 동적 연결망의 특성을 만족시키기 위해서 매시점마다 동일한 값을 사용하지 않고 일정한 범위 $(1 - \delta_{c_i}, 1 + \delta_{c_i})$ 내에서 변화를 시켰으며, $k = 2$이고 $n_1 < n_2$인 경우 규모가 큰 커뮤니티에 속한 노드의 모수 $\theta_i$의 변동을 더 큰 값으로 사용하였다 (즉, $\delta_1 < \delta_2$).

위에서 언급한 동적 DCSBM 모수의 설정을 Table 1에 6가지 경우로 요약하여 정리하였다.

## 4.3. $k = 1$인 경우 모의실험 결과

먼저 전체 연결망이 하나의 커뮤니티를 형성하는 DCSBM을 고려해 보자. 단계 1–5의 모의실험을 수행하여 초기 연결망의 수 $m$에 대한 AARL과 SDARL값을 Table 2에 제시하였다. 또한 $m$개의 초기 연결망에서 계산된 $Y_t(1, 1)$과 실제 참값인 $\lambda_t(1, 1)$의 평균 차이를 $e$라고 할 때, 반복하여 생성했던 연결망 1,000개 (모의실험 절차에서 단계 5를 참조)에서 계산된 $e$들의 평균 $\bar{e}$와 표준편차 sd($e$)를 산출하여 함께 제시하였다.

Case 1의 경우를 살펴보면, $m$이 아주 작은 경우에도 AARL은 주어진 값인 $\text{ARL}_0$=370.4를 근사적으로 잘 만족한다는 것을 알 수 있다. 이것은 $m$이 작은 경우에도 각 연결망의 크기는 40으로 어느 정도 크기 때문에 모수 추정이 비교적 정확하게 되었기 때문이라고 판단된다. 이것은 $m = 3$이 되면 $\bar{e}$값(=-0.07)이 아주 작아지는 것으로도 유추할 수 있다. SDARL의 경우 $m = 1$과 2와 같이 아주 작은 때 9.31과 9.84로 큰 값을 가졌지만,

Table 2: Simulation results for Caes 1 and 2

| Case | $m$ | AARL | SDARL | $\bar{e}$ | sd($e$) |
|---|---|---|---|---|---|
| 1 | 1 | 369.48 | 9.31 | −0.28 | 12.57 |
| | 2 | 369.69 | 9.84 | −0.19 | 8.86 |
| | 3 | 370.38 | 8.01 | −0.07 | 7.49 |
| | 5 | 370.35 | 8.61 | 0.05 | 5.70 |
| | 10 | 369.75 | 9.03 | 0.06 | 3.84 |
| | 20 | 369.71 | 9.65 | −0.08 | 2.90 |
| | 50 | 370.52 | 8.60 | −0.02 | 1.81 |
| | 100 | 370.36 | 8.31 | 0.03 | 1.25 |
| | 200 | 370.10 | 7.78 | −0.01 | 0.88 |
| | 500 | 369.70 | 7.18 | 0.02 | 0.58 |
| 2 | 1 | 370.44 | 7.35 | −1.14 | 21.49 |
| | 2 | 370.20 | 8.29 | 0.50 | 15.53 |
| | 3 | 370.72 | 7.58 | −0.03 | 12.66 |
| | 5 | 370.61 | 8.14 | 0.06 | 9.73 |
| | 10 | 370.00 | 9.15 | 0.03 | 7.00 |
| | 20 | 370.13 | 8.42 | 0.04 | 4.74 |
| | 50 | 370.24 | 7.62 | 0.03 | 3.05 |
| | 100 | 370.11 | 7.15 | 0.03 | 2.27 |
| | 200 | 370.28 | 6.32 | −0.01 | 1.53 |
| | 500 | 370.62 | 5.30 | −0.04 | 1.00 |

AARL = average of the ARL(average run length); SDARL = standard deviation of the ARL.

$m = 3$이 되면 8.01로 감소하는 것으로 나타났다. 그리고 SDARL값은 $m = 20$이 될 때까지 증가하다가 그 이후로 다시 감소하는 것을 알 수 있다. 그리고 $m = 3$인 경우 SDARL값(= 8.01)은 $m = 100$과 $m = 200$ 사이일 때의 SDARL값과 유사한 것으로 판단된다. 따라서 Case 1의 경우 초기 연결망의 수를 $m = 3$으로 설정해도 관리상태의 성능에는 큰 문제가 없음을 알 수 있다.

Case 2의 경우 전반적으로 Case 1과 유사한 경향의 결과를 얻었다. 먼저 AARL은 $m$이 작은 경우에도 근사적으로 주어진 값인 $ARL_0$=370.4를 잘 만족하는 것으로 나타났다. 그리고 SDARL은 $m = 3$부터 점차 증가하여 $m = 10$ 이후 다시 감소는 경향을 나타냈으며, $m = 3$인 경우 SDARL값(= 7.58)은 $m = 50$인 경우의 값(= 7.62)과 거의 유사함을 알 수 있다. $\bar{e}$와 sd($e$)값으로 판단할 때에도 $m = 3$인 경우의 성능이 나쁘지 않음을 알 수 있다.

따라서 위의 두 경우를 종합해서 판단할 때, AARL은 초기 연결망의 수 $m$과 상관 없이 근사적으로 주어진 값과 유사하게 나왔으나 SDARL은 $m$이 커지면서 점차 증가하다 어느 시점 부터 감소하는 경향을 나타냈다. 또한 두 경우 모두 $m = 3$일 때 초기 연결망의 수가 작음에도 불구하고 성능이 좋은 것을 알 수 있다. 또한 연결망 모수의 추정이 어느 정도 안정적일 때 $\bar{e}$는 $|\bar{e}| < 0.1$의 값을 가지며, sd($e$)값은 $m$이 커질수록 계속 감소하였다.

## 4.4. $k = 2$인 경우 모의실험 결과

이제 연결망이 두 개의 커뮤니티로 형성되어 있는 DCSBM을 고려해 보자. 이때 $m$개의 초기 연결망에서 계산된 $Y_t(1,1)$과 실제 참값인 $\lambda_t(1,1)$의 평균 차이를 $e_{11}$이라 하고 이에 대한 평균과 표준편차를 각각 $\bar{e}_{11}$과 sd($e_{11}$)으로 나타내었다. 이와 유사하게 $Y_t(1,2)$와 $\lambda_t(1,2)$의 평균 차이는 $e_{12}$라 하고 평균과 표준편차는 각각 $\bar{e}_{12}$와 sd($e_{12}$), 그리고 $Y_t(2,2)$와 $\lambda_t(2,2)$의 평균 차이는 $e_{22}$라 하고 평균과 표준편차는 각각 $\bar{e}_{22}$와 sd($e_{22}$)로 나타내었

Table 3: Simulation results for Case 3

| Case | $m$ | AARL | SDARL | $\bar{e}_{11}$ | sd($e_{11}$) | $\bar{e}_{12}$ | sd($e_{12}$) | $\bar{e}_{22}$ | sd($e_{22}$) | $\bar{e}$ | sd($e$) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 1 | 375.42 | 9.01 | −0.23 | 5.98 | 0.29 | 6.26 | −0.30 | 5.96 | −0.08 | 6.07 |
| | 2 | 375.90 | 8.42 | 0.04 | 4.25 | 0.19 | 4.42 | −0.03 | 4.45 | 0.07 | 4.37 |
| | 3 | 375.67 | 8.60 | 0.05 | 3.64 | −0.13 | 3.59 | −0.04 | 3.52 | −0.04 | 3.58 |
| | 5 | 375.37 | 9.07 | −0.01 | 2.80 | 0.02 | 2.95 | −0.15 | 2.75 | −0.05 | 2.83 |
| | 10 | 376.44 | 9.90 | 0.05 | 1.98 | 0.00 | 1.97 | 0.00 | 1.97 | 0.02 | 2.34 |
| | 20 | 376.09 | 10.70 | −0.06 | 1.39 | −0.01 | 1.35 | 0.02 | 1.38 | −0.02 | 1.37 |
| | 50 | 376.58 | 12.26 | 0.05 | 0.88 | 0.00 | 0.94 | −0.04 | 0.88 | 0.00 | 0.90 |
| | 100 | 377.79 | 12.58 | 0.02 | 0.61 | −0.01 | 0.65 | −0.01 | 0.61 | 0.00 | 0.62 |
| | 200 | 379.29 | 14.78 | −0.02 | 0.44 | 0.00 | 0.44 | 0.01 | 0.44 | 0.00 | 0.44 |
| | 500 | 384.70 | 16.56 | 0.00 | 0.28 | 0.00 | 0.28 | 0.01 | 0.27 | 0.00 | 0.28 |

AARL = average of the ARL(average run length); SDARL = standard deviation of the ARL.

Table 4: Simulation results for Case 4

| Case | $m$ | AARL | SDARL | $\bar{e}_{11}$ | sd($e_{11}$) | $\bar{e}_{12}$ | sd($e_{12}$) | $\bar{e}_{22}$ | sd($e_{22}$) | $\bar{e}$ | sd($e$) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 1 | 375.08 | 11.20 | 0.10 | 3.54 | 0.01 | 5.58 | 0.03 | 9.06 | 0.05 | 6.47 |
| | 2 | 375.14 | 9.60 | −0.01 | 2.53 | −0.04 | 3.94 | 0.03 | 6.64 | −0.01 | 4.69 |
| | 3 | 375.39 | 9.40 | −0.04 | 2.10 | 0.13 | 3.13 | −0.32 | 5.20 | −0.08 | 3.71 |
| | 5 | 375.29 | 10.07 | −0.05 | 1.65 | −0.09 | 2.41 | −0.05 | 4.06 | −0.07 | 2.88 |
| | 10 | 375.95 | 10.85 | 0.00 | 1.16 | 0.03 | 1.71 | 0.12 | 2.94 | 0.05 | 2.07 |
| | 20 | 376.46 | 11.37 | 0.02 | 0.80 | 0.07 | 1.21 | −0.02 | 2.02 | 0.03 | 1.44 |
| | 50 | 375.77 | 13.47 | −0.02 | 0.51 | −0.03 | 0.77 | −0.02 | 1.31 | −0.03 | 0.93 |
| | 100 | 376.15 | 12.21 | 0.00 | 0.36 | −0.01 | 0.54 | 0.03 | 0.98 | 0.01 | 0.68 |
| | 200 | 375.53 | 11.31 | 0.00 | 0.26 | −0.01 | 0.39 | −0.04 | 0.66 | −0.02 | 0.47 |
| | 500 | 375.66 | 8.66 | 0.00 | 0.17 | 0.00 | 0.24 | 0.01 | 0.42 | 0.00 | 0.29 |

AARL = average of the ARL(average run length); SDARL = standard deviation of the ARL.

다. 또한 모든 차이들에 대한 평균과 표준편차를 각각 $\bar{e}$와 sd($e$)로 나타내었다.

Case 3과 Case 4는 전체 노드의 수가 40인 경우이다. Case 3에 대한 모의실험의 결과를 Table 3에 제시하였다. Case 3은 커뮤니티의 규모를 동일하게 설정한 경우로서, AARL은 $m$이 커짐에 따라 다소 증가하는 경향을 나타냈으며, SDARL은 $m = 2$부터 $m$이 커짐에 따라 계속 증가하였다. 여기에는 제시하지 않았지만 $m = 2,000$ 이후부터 SDARL값은 감소하는 것으로 타나났다. 이와 같은 결과는 각 커뮤니티의 크기가 20으로 다소 작기 때문에 각 커뮤니티의 모수가 정확하게 추정되지 못한 것에서 기인한 것으로 판단된다.

Case 4에 대한 결과는 Table 4에 제시되어 있다. 이때 AARL은 대략적으로 375 정도의 값을 가지기 때문에, 주어진 $ARL_0 = 370.4$보다 조금 큰 값을 갖는 것으로 나타났다. 그리고 SDARL은 $m = 3$일 때 9.40으로 작은 값을 갖으며, 그 이후 $m = 50$까지 증가하다 다시 감소하는 경향을 보였다. $m = 3$일 때 SDARL값(= 9.40)은 $m = 200$일 때의 값(= 11.31)보다 더 작은 값을 갖는다. Case 4는 Case 3일 때와 달리 $n_1 < n_2$이지만 커뮤니티의 경향성 모수는 $P_{11} > P_{22}$로 설정했고 각 노드의 연결모수의 변동성은 $\delta_1 < \delta_2$로 설정했기 때문에, 커뮤니티 1에서 차이에 대한 표준편차인 sd($e_{11}$)이 커뮤니티 2에서의 sd($e_{22}$)에 비하여 상대적으로 작은 값을 가짐을 알 수 있다. Case 3에 대한 Table 3의 결과와 비교해도, 동일한 $m$에 대하여 sd($e_{11}$)은 더 작지만 sd($e_{22}$)는 더 큰 값을 갖는 것으로 나타났다.

결론적으로 Case 1과 Case 3, 4의 결과를 비교하면, 전체 노드의 수가 40과 같이 크지 않은 때에는 커뮤니티가 분리된 경우의 효율이 상대적으로 낮음을 알 수 있다. 또한 두 개의 커뮤니티로 형성되는 경우에도 $m = 3$

Table 5: Simulation results for Case 5

| Case | $m$ | AARL | SDARL | $\bar{e}_{11}$ | sd($e_{11}$) | $\bar{e}_{12}$ | sd($e_{12}$) | $\bar{e}_{22}$ | sd($e_{22}$) | $\bar{e}$ | sd($e$) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 1 | 370.57 | 7.68 | −0.73 | 15.76 | −0.36 | 15.54 | 0.13 | 16.09 | −0.32 | 15.80 |
| | 2 | 370.56 | 6.62 | −0.08 | 11.07 | −0.58 | 11.11 | −0.61 | 10.81 | −0.42 | 11.00 |
| | 3 | 370.59 | 6.30 | −0.08 | 9.03 | 0.08 | 8.88 | 0.30 | 9.04 | 0.10 | 8.98 |
| | 5 | 370.55 | 6.16 | −0.03 | 7.07 | −0.11 | 7.22 | 0.26 | 6.86 | 0.04 | 7.05 |
| | 10 | 370.52 | 6.25 | 0.15 | 4.90 | −0.09 | 4.68 | 0.06 | 5.01 | 0.04 | 4.86 |
| | 20 | 370.62 | 6.72 | −0.10 | 3.34 | 0.01 | 3.48 | 0.04 | 3.51 | −0.02 | 3.44 |
| | 50 | 370.08 | 6.54 | −0.10 | 2.17 | −0.10 | 2.29 | −0.07 | 2.27 | −0.09 | 2.24 |
| | 100 | 370.37 | 6.23 | −0.02 | 1.54 | −0.06 | 1.59 | −0.02 | 1.59 | −0.03 | 1.57 |
| | 200 | 370.49 | 5.67 | −0.02 | 1.11 | 0.01 | 1.16 | −0.03 | 1.11 | −0.01 | 1.13 |
| | 500 | 370.87 | 5.26 | 0.05 | 0.69 | −0.03 | 0.72 | −0.01 | 0.70 | 0.00 | 0.70 |

AARL = average of the ARL (average run length); SDARL = standard deviation of the ARL.

Table 6: Simulation results for Case 6

| Case | $m$ | AARL | SDARL | $\bar{e}_{11}$ | sd($e_{11}$) | $\bar{e}_{12}$ | sd($e_{12}$) | $\bar{e}_{22}$ | sd($e_{22}$) | $\bar{e}$ | sd($e$) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 1 | 370.37 | 9.01 | 0.07 | 9.71 | 0.04 | 13.12 | −1.34 | 23.66 | −0.41 | 16.60 |
| | 2 | 370.52 | 6.69 | 0.23 | 6.71 | 0.50 | 9.63 | 0.25 | 16.11 | 0.33 | 11.50 |
| | 3 | 370.52 | 6.28 | 0.18 | 5.38 | −0.32 | 7.94 | −0.35 | 13.41 | −0.16 | 9.52 |
| | 5 | 370.73 | 6.71 | 0.18 | 4.24 | 0.11 | 6.03 | 0.07 | 10.84 | 0.12 | 7.57 |
| | 10 | 370.89 | 6.73 | 0.12 | 2.97 | 0.05 | 4.28 | 0.08 | 7.48 | 0.08 | 5.26 |
| | 20 | 370.52 | 6.91 | −0.13 | 2.07 | −0.12 | 3.05 | 0.13 | 5.19 | −0.04 | 3.67 |
| | 50 | 370.66 | 7.23 | 0.00 | 1.39 | 0.02 | 1.95 | −0.12 | 3.20 | −0.03 | 2.31 |
| | 100 | 370.32 | 7.16 | 0.00 | 0.94 | −0.08 | 1.37 | −0.08 | 2.43 | −0.05 | 1.70 |
| | 200 | 370.85 | 6.85 | 0.02 | 0.65 | 0.02 | 0.96 | −0.04 | 1.72 | 0.00 | 1.20 |
| | 500 | 371.22 | 7.63 | 0.02 | 0.44 | 0.03 | 0.60 | 0.01 | 1.04 | 0.02 | 0.74 |

AARL = average of the AR L(average run length); SDARL = standard deviation of the ARL.

일 때의 성능이 우수한 것을 알 수 있었다.

Case 5와 Case 6은 전체 노드의 수가 100인 경우이다. Case 5에 대한 모의실험의 결과를 Table 5에 제시하였다. 이 경우 AARL은 $m$이 작은 경우에도 주어진 370.4과 유사했으며, SDARL은 $m = 5$일 때까지 감소하다가 $m = 20$일 때까지 증가하고 그 이후에 다시 감소하는 경향을 보였다. SDARL 측면에서는 $m = 5$일 때의 성능이 가장 좋았는데 (SDARL = 6.16), 이 경우 $m = 100$일 때와 유사한 SDARL값(= 6.23)을 갖는 것으로 나타났다. 또한 Case 5를 커뮤니티 크기의 비율이 동일한 Case 3과 비교했을 때, 예상한 결과이지만 각 노드의 수인 $n_i$ ($i = 1, 2$)가 커짐에 따라 전반적인 성능이 좋아진 것을 알 수 있다. 즉, 동일한 $m$에 대해 Case 5의 AARL은 주어진 $\text{ARL}_0 = 370.4$와 더 유사한 값을 가지면서 SDARL은 더 작은 값을 갖는 것으로 나타났다.

Case 6에 대한 결과는 Table 6에 제시되어 있는데, 전반적으로 Case 5와 유사한 경향을 나타내었다. 즉, AARL은 근사적으로 주어진 370.4과 유사한 값을 가지며, SDARL은 $m = 3$일 때까지 감소하다가 $m = 50$일 때까지 증가한 후 다시 감소하는 경향을 보였다. 또한 Case 6을 Case 5와 비교했을 때, sd($e_{11}$)값은 더 작지만 sd($e_{22}$)값은 더 큰 것을 알 수 있다. 이것은 커뮤니티의 크기, $P_{rr}$과 $\delta_r$ ($r = 1, 2$) 값의 조정이 표준편차에 영향을 준 것을 판단된다. 또한 Case 6을 커뮤니티 크기의 비율이 동일한 Case 4와 비교했을 때, Case 6의 성능이 Case 4에 비하여 더 좋은 것으로 나타났다. 이는 앞에서 언급한 바와 같이 커뮤니티의 규모가 더 커졌기 때문에 나타난 결과라고 생각된다.

통계량 $Y_t(r, s)$ $(r, s = 1, \ldots, k,\ r \le s)$는 모형의 모수인 $\lambda_t(r, s)$의 추정량의 역할을 한다. 따라서 제1국면의

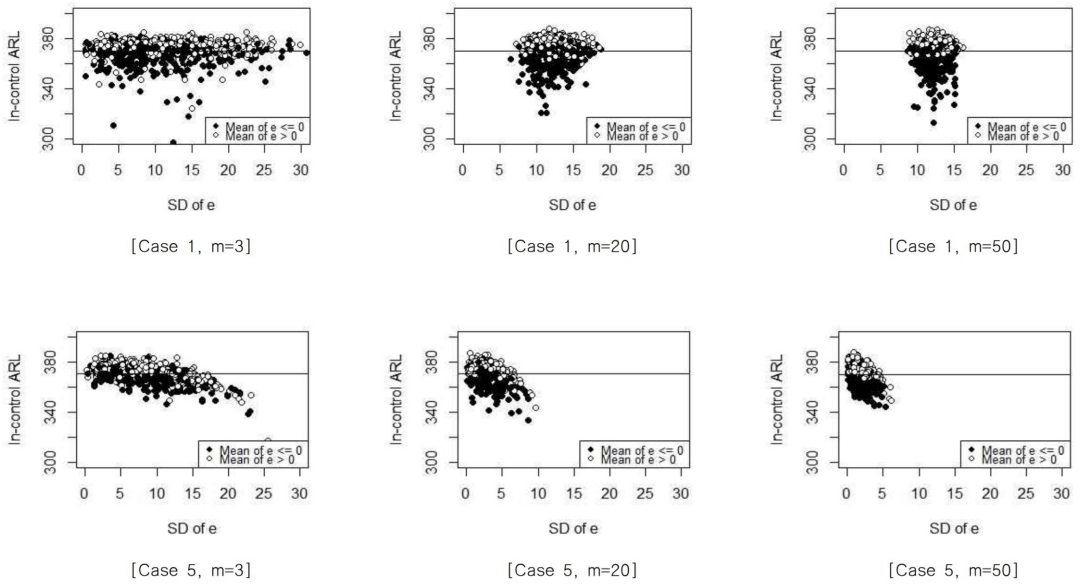


Figure 2: *Scatterplot of in-control arl versus standard deviation of e.*

$m$개의 초기 연결망에서 계산된 $Y_t(r, s)$과 $\lambda_t(r, s)$의 차이를 $e$라고 할 때, 이 값들이 관리상태의 ARL에 어떠한 영향을 주는지 파악하는 것은 의미가 있다고 판단된다.

Case 1과 Case 5에서 $m$ = 3, 20, 50일 때 $e$들의 평균, 표준편차, 그리고 각 경우에 계산된 ARL의 관계를 Figure 2에 도시하였다. 각 경우에 대해 $x$축은 $m$개의 초기 연결망 1,000개에서 계산된 $e$들의 표준편차이고, $y$축은 각 경우에 대한 관리상태의 ARL을 나타낸다. $y$축에서 주어진 값인 $ARL_0$ = 370.4에는 선을 그어서 비교를 용이하게 하였다. 그리고 그래프에 타점된 점은 초기 연결망 1,000개에서 계산된 $e$들이 0보다 큰 경우와 작은 경우로 구분하여 표시하였다.

눈에 띄는 사실은 초기 모수가 크게 추정된 경우 (오차의 평균이 양수인 경우) 관리상태의 ARL이 주어진 370.4보다 더 크게 나오고, 작게 추정된 경우 (오차의 평균이 음수인 경우) ARL이 370.4보다 더 작게 나오는 경향이 있다는 것이다. 이러한 경향은 초기 연결망의 수 $m$이 증가할수록 더 두드러지게 나타났다. 예상할 수 있는 결과이지만, 초기 연결망을 사용하여 수행한 모수 추정의 정확성이 관리상태의 ARL에 작지 않은 영향을 준다는 사실을 알 수 있다.

## 5. 결론

최근 사회 연결망에 어떤 비정상적 변화가 발생했는가를 감시하는 연구가 활발하게 진행되고 있는데, 이를 위해 통계적 공정 모니터링 절차인 관리도가 사용되고 있다. 물론 생산 공정에서 사용되고 있는 관리도 절차를 사회 연결망에 적용하기 위해서는 연결망 모형의 특징을 파악하고 이에 맞는 통계량과 관리한계를 설정해야 한다.

연결망 감시를 위한 관리도 절차를 적용하기 위해서는 적절한 모형을 선택하고, 모형의 모수를 추정하는 제1국면을 진행해야 한다. 그러나 일반적으로 사회 연결망은 제1국면의 진행을 위해 충분한 수의 연결망을 확보하기가 어려운 경우가 대부분이다. 따라서 이러한 경우 사용할 수 있는 절차에 대한 연구가 필요한데,

이에 대한 연구는 아직 진행되고 있지 않는 것으로 판단된다. 이 논문에서는 최근에 많이 사용되는 모형인 동적 DCSBM을 가정한 경우, 최소한의 초기 연결망으로 관리도를 적용할 수 있는 자기출발 관리도 절차를 제안하고 그 성능을 살펴보았다.

절차의 성능을 판단하기 위해, DCSBM의 특징에 따라 커뮤니티의 경향성과 연결정도의 이질성을 고려하여 모의실험을 설계하고 모의실험을 진행하였다. 커뮤니티의 수가 하나와 두 개인 경우에 대해 AARL과 SDARL 측도를 사용하여 절차의 성능을 살펴본 결과, 초기 연결망이 $m = 3$개 정도 확보가 된다면 제안된 절차의 효율성이 어느 정도 확보된다는 사실을 알 수 있었다. 즉, 제안된 절차는 적은 수의 초기 연결망을 사용하여 사회 연결망을 감시하는 효율적인 방법이라고 할 수 있다.

이 논문에서는 비정상적 변화가 없는 관리상태에서, 각 노드의 연결정도 모수만 시간에 따라 변화하는 가장 간단한 동적 모형을 가정하였다. 향후 각 노드의 커뮤니티 레이블과 커뮤니티의 의사소통 경향성도 시간에 따라 변화할 수 있는 동적 모형에 대한 절차를 개발하는 것은 중요한 연구 과제가 될 것이다. 또한 이 논문에서는 제안된 자기출발 절차에 대해 관리상태의 성능만을 평가했는데, 비정상적 변화가 발생한 이상상태에서의 성능을 평가하는 것도 중요한 연구 과제가 될 것이라 생각한다.

## References

Capizzi, G. and Masarotto, G. (2010). Self-starting CUSCORE control charts for individual multivariate observations, *Journal of Quality Technology*, **42**, 136–151.

Hawkins, D. M. (1987). Self-starting CUSUM charts for location and scale, *The Statistician*, **36**, 299–315.

Hawkins, D. M. and Maboudou-Tchao, E. M. (2007). Self-starting multivariate exponentially weighted moving average control charting, *Technometrics*, **49**, 199–209.

Hawkins, D. M. and Olwell, D. H. (1998). Cumulative sum charts and charting for quality improvement, *New York*, Springer.

Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks, *Physical Review E*, **83**, 016107.

Keefe, M. J., Woodall, W. H., and Jones-Farmer, L. A. (2015). The conditional in-control performance of self-starting control charts, *Quality Engineering*, **27**, 488–499.

Lee, J. W., Kim, M., and Lee, J. (2018). The in-control performance of self-starting EWMA and $\bar{X}$ charts, *Journal of the Korean Data & Information Science Society*, **29**, 851–860.

Priebe, C. E., Conroy, J. M., Marchette, D. J., and Park, Y. (2005). Scan statistics on Enron graphs, *Computational and Mathematical Organization Theory*, **11**, 229–247.

Qin, T. and Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic block model, *Advances in Neural Information Processing Systems*, 3120–3128.

Quesenberry, C. P. (1991a). SPC Q charts for a binomial parameter p: Short or long runs, *Journal of Quality Technology*, **23**, 239–246.

Quesenberry, C. P. (1991b). SPC Q charts for a Poisson parameter $\lambda$: Short or long runs, *Journal of Quality Technology*, **23**, 296–303.

Quesenberry, C. P. (1995). Geometric Q charts for high quality processes, *Journal of Quality Technology*, **27**, 304–315.

Sengupta, S. and Chen, Y. (2015). Spectral clustering in heterogeneous networks, *Statistica Sinica*, **25**, 1081–1106.

Shen, X., Tsui, K. L., Zou, C., and Woodall, W. H. (2016). Self-starting monitoring scheme for Poisson count data with varying population sizes, *Technometrics*, **58**, 460–471.

Shetty, J. and Adibi, J. (2005). Discovering important nodes through graph entropy the case of Enron email database, *Proceeding of the 3rd international workshop on Link discovery*, 74–81.

Sullivan, J. H. and Jones, L. A. (2002). A self-starting control chart for multivariate individual observations, *Technometrics*, **44**, 24–33.

Wilson, J. D., Stevens, N. T., and Woodall, W. H. (2019). Modeling and detecting change in temporal networks via a dynamic degree corrected stochastic block model, *Quality and Reliability Engineering International*, **35**, 1363–1378.

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

Yu, L., Zwetsloot, I. M., Stevens, N. T., Wilson, J. D., and Tsui, K. L. (2020). Monitoring dynamic networks: a simulation-based strategy for comparing monitoring methods and a comparative study, submitted to a journal for publication.

Zhang, M., Peng, Y., Schuh, A., Megahed, F. M., and Woodall, W. H. (2013). Geometric charts with estimated control limits. *Quality and Reliability Engineering International*, **29**, 209–223.

*Received October 5, 2020; Revised October 31, 2020; Accepted November 4, 2020*

# 동적 DCSBM을 모니터링하는 자기출발 절차

이주원[a], 이재헌[1,a]

[a]중앙대학교 응용통계학과

요 약

최근 동적 연결망의 비정상적 변화를 감시하기 위한 연결망 모니터링의 필요성이 높아지고 있다. 이 논문에서는 연결망의 구조적 변화를 감시하기 위한 동적 연결망의 모형으로 DCSBM(degree corrected stochastic block model)을 고려하였다. 관리도 절차를 사용하여 동적 연결망을 감시하려면 제1국면을 통해 초기 연결망을 확보한 후 모형의 모수를 추정하는 단계를 거쳐야 한다. 그러나 연결망의 감시에서는 충분한 수의 초기 연결망을 확보하기 어려운 경우가 대부분이다. 이 논문에서는 동적 DCSBM을 감시하기 위한 자기출발 관리도 절차를 제안한다. 이 절차는 모형의 모수 추정을 위해 확보한 연결망의 수가 아주 적은 경우에 유용하게 사용할 수 있는 절차이다. 모의실험을 통해 절차의 성능을 평가한 결과, 제안된 절차는 초기 연결망의 수가 아주 적은 경우에도 좋은 관리상태의 성능을 나타내는 것을 알 수 있었다.

주요용어: 관리도, 사회 연결망, 연결망 감시, 자기출발 절차, 통계적 공정관리

**이 논문은 2019년도 중앙대학교 연구 장학기금 지원에 의한 것임.**
교신저자: (06974) 서울특별시 동작구 흑석로 84, 중앙대학교 응용통계학과. E-mail: jaeheon@cau.ac.kr