碩 士 學 位 論 文

소셜네트워크자료를 위한 잠재적 위치 군집 분석 개선 연구

高麗大學校 大學院

應用統計學科

金 志 容

2017年 12月

全 秀 榮 敎 授 指 導

碩 士 學 位 論 文

소셜네트워크자료를 위한 잠재적 위치 군집 분석 개선 연구

이 論文을 統計學 碩士學位 論文으로
提出함.

2017年 12月

高 麗 大 學 校 大 學 院

應 用 統 計 學 科

金 志 容 (印)

金　志　容의 統計學 碩士學位論文

審査를 完了함.

2017年 12月

委員長　　전수영　　(印)

委　員　　홍승만　　(印)

委　員　　진서훈　　(印)

요 약 문

소셜네트워크(Social Network)자료는 객체(actor) 또는 노드(node)들 사이의 관계를 나타내기 위한 자료로 사회학, 경제학, 경영학을 비롯한 사회과학 분야에서뿐만 아니라 물리학, 의학, 생물학 등의 자연과학과 같이 다양한 분야에서 사용되고 있다. 소셜네트워크에서 노드는 한 사람 또는 가족, 조직과 같이 더 큰 사회적 그룹을 나타내고 연결선(edge, link, tie, arc)은 두 노드 사이의 연결 여부 혹은 연결 강도를 나타낸다. 소셜네트워크분석은 이러한 연결선을 통해 노드들 간의 규칙적인 관계 패턴을 파악하여 네트워크 전체적인 구조를 이해하고 분석하는 것으로 분석 목적에 따라 적절한 통계적 모형을 이용해 다양한 의미를 도출해 낼 수 있다. 하지만 소셜네트워크분석에 이용되는 통계적 모형의 우수성은 정확한 모수 추정에 의존한다.

본 논문은 잠재적 위치 군집모형을 이용한 분석에서 개선된 모수 추정을 위해 새로운 표본생성 방법을 제안한다. 소셜네트워크분석 시 지수족 랜덤 그래프모형과 잠재변수를 고려한 모형은 고전적 모형으로 매우 복잡하고 모수 추정에 어려움이 많다. 따라서 이를 개선하기 위해 본 연구는 각 군집별 정보를 이용하는 새로운 표본생성 방법을 제시한다. 제안된 방법은 모의실험과 실 자료 분석을 통해 다른 방법보다 더 효율적인 추정치를 제공하는 우수성을 보여 주었다.

핵심어 : 소셜네트워크, Exponential-family Random Graph Model, Latent Position Cluster Model, Markov Chain Monte Carlo

목 차

표 목 차

제 1 장 서 론

소셜네트워크(social networks)는 사람은 사회적 관계를 기반으로 상호작용을 통해 서로 엮여있다는 개념에서부터 시작한다. 또한, 다양한 사회적 현상을 설명할 때 소셜네트워크의 개념을 적용할 수 있다. 소셜네트워크는 노드(node) 또는 객체(actor)들 사이의 관계를 나타내기 위한 자료로 사회학, 경제학, 경영학을 비롯한 사회과학 분야에서 뿐만 아니라 물리학, 의학, 생물학 등의 자연과학과 같이 다양한 분야에서 사용된다(Gormley, Murphy, 2010). 소셜네트워크에서 노드는 한 사람 또는 가족, 조직과 같이 더 큰 사회적 그룹을 나타내거나 개념, 문자, 변수 등의 추상적인 존재를 표현하는데 사용할 수 있다(Ryan et al., 2017). 연결선(edge, link, tie, arc)은 두 노드 사이의 연결 여부 혹은 연결 강도를 나타낸다(Handcock et al., 2007).

소셜네트워크분석은 연결되거나 연결되지 않은 많은 노드들 간의 규칙적인 관계패턴을 파악하여 네트워크의 전체적인 구조를 추정하고 분석하는 것으로 분석 목적에 따라 다양한 의미를 도출할 수 있다. 이와 관련된 연구는 오랫동안 연구되어 왔으며 대표적인 책으로 Wasserman, Faust(1994)가 있다(Hunter et al., 2012).

일반적으로 소셜네트워크자료는 n개의 노드와 연결선 $y_{i,j}$로 구성된다. $y_{i,j}$가 관계의 유무를 나타내는 이항 변수라면 친구들 사이의 관계 유무, 회사들 간의 협력관계 유무 등이 될 수 있다. 자료는 $n \times n$형태의 행렬 Y로 정의하며 소시오매트릭스(Sociomatrix)라고 부른다. 또한, 네트워크에서의 관계는 방향성(directed)이거나 비방향성(undirected)인 경우로 나눠지며 자기 자신과의 관계 (i,i)는 가질 수 없게 제한한다. 소셜네트워크의 노드를 효율적이고 유연하게 군집화(clustering)하여 네트워크의 구조를 파악하는 것이 소셜네트워크분석의 목표이며 이 때 다양한 통계적 방법이 적용되고 있다.

소셜네트워크분석에 관한 선행연구는 먼저 연결선들의 조건부 독립을 가정한

p_1모형(Holland, Leinhardt, 1981)이 제안되었다. 하지만 모형 절약성 부족과 공변량을 고려하지 않은 문제가 존재해 확률변수 개념을 추가한 p_2모형(van Duijn et al., 1995)이 제안되었다. p_2모형은 모형이 복잡하고 해석하기 어려운 문제점이 있어 p_1모형과 p_2모형의 한계를 보완한 지수족 랜덤 그래프 모형(Exponential-family Random Graph Model, ERGM)이 제안되었다. 지수족 랜덤 그래프 모형은 Frank, Strauss(1986)에 의해 일반화 되었으며 지속적으로 연구되고 있지만 계산상의 어려움과 모형 퇴화 문제가 존재한다. 이러한 문제를 해결하기 위해 잠재변수를 고려한 모형이 제안되었으며 그 중 Hoff et al.(2002)는 각 노드는 알려지지 않은 잠재적 위치를 가지고 있다고 가정한 잠재 공간 모형을 제안했다. 잠재 공간 모형은 모수 추정 시 비식별 문제가 존재한다. Handcock et al.(2007)은 잠재 공간 모형을 바탕으로 모형 기반 군집화 개념을 적용한 잠재적 위치 군집모형을 제안했으나 대용량 네트워크에 적용 시 계산적 문제가 존재한다.

본 논문에서는 소셜네트워크분석을 위해 사용되는 여러 가지 통계적 모형 중 고전적 모형인 지수족 랜덤 그래프 모형과 잠재변수 모형 중 대표적인 모형 세 가지를 2장에서 소개한다. 3장에서는 잠재적 위치 군집모형(Latent Position Cluster Model, LPCM)에 대한 소개와 개선된 모수 추정을 위한 새로운 표본생성 방법을 제안한다. 4장에서는 모의실험 자료를 바탕으로 새로운 표본생성 방법을 이용한 모수 추정 결과를 기존의 방법들과 비교한다. 5장에서는 네트워크 분석에서 대표적으로 많이 사용되는 실제 예제를 적용해 보았다.

제 2 장 고전적 모형

소셜네트워크분석에 사용되는 대표적인 고전적 통계 모형으로는 지수족 랜덤 그래프 모형과 잠재변수 모형이 있다.

2.1 지수족 랜덤 그래프 모형

지수족 랜덤 그래프 모형은 네트워크의 구조를 네트워크 내에 존재하는 모든 연결선들의 결합 분포로 표현한 대표적인 모형이다. $p1$모형과 $p2$모형의 한계를 보완하여 만들어 졌기 때문에 p^*모형으로 불렸으나 현재는 ERGM으로 더 많이 불린다.

$p1$모형은 Holland, Leinhardt(1981)에 의해 제안되었으며 연결선들의 조건부 독립을 가정하고 연결선들의 상호성(reciprocity)을 모형화 하는 것이 목적이다. $p1$모형에서 연결 관계는 다음과 같이 표현할 수 있다(Wasserman, Faust, 1994).

a) 연결이 없는 상태($y_{ij} = y_{ji} = 0$ 또는 $Y_{ij00} = 1$)

b) 상호 연결이 있는 상태($y_{ij} = y_{ji} = 1$ 또는 $Y_{ij11} = 1$)

c) 비대칭 상태($y_{ij} = 1,\ y_{ji} = 0$또는 $Y_{ij10} = 1$, $y_{ij} = 0,\ y_{ji} = 1$또는 $Y_{ij01} = 1$)

이를 바탕으로 로그 선형 모형식으로 표현하면 아래와 같다.

$$\log P(Y_{ij00} = 1) = \lambda_{ij}$$
$$\log P(Y_{ij10} = 1) = \lambda_{ij} + \theta + \alpha_i + \beta_j$$
$$\log P(Y_{ij01} = 1) = \lambda_{ij} + \theta + \alpha_j + \beta_i$$
$$\log P(Y_{ij11} = 1) = \lambda_{ij} + 2\theta + \alpha_i + \alpha_j + \beta_i + \beta_j + \alpha\beta$$

α는 연결을 보내는 경향(발신자), β는 연결을 받는 경향(수신자), $\alpha\beta$는 상호 연결된 경향을 나타내며 λ는 로그 선형 모형식에서의 절편을 의미한다. θ는 연결을 주고받는 전체 선택 효과의 수를 의미한다. 비대칭 상태로 연결된 경우 θ, 상호 연결 상태인 경우 2θ이다.

$p2$모형은 $p1$모형에서 확장됐으며 van Duijn et al.(1995)이 제안했다. $p2$모형은 α, β, $\alpha\beta$들을 확률변수로 보았다. 이로 인해 노드들 간의 이질성을 고려할 수 있게 돼 더 정교한 모형이 되었다.

ERGM은 Frank, Strauss(1986)에 의해 일반화 되었으며 Lusher et al.(2013)이 ERGM의 기본 이론적 가정을 다음과 같이 정리하였다.

Lusher's ERGM 기본 가정:

a) 사회 연결망은 국소적(local)으로 나타난다.
b) 네트워크의 연결은 스스로 구성될 뿐만 아니라 노드의 속성과 다른 외생적 요인들에 의해 영향 받는다.
c) 네트워크의 유형은 구조적 과정의 증거로 볼 수 있다.
d) 여러 과정들이 동시에 일어날 수 있다.
e) 사회 연결망은 구조적이며 확률적이다.

위와 같은 가정을 전제로 구성된 식은 (1)과 같다.

$$P_\theta(Y=y)=\frac{\exp\{\eta(\theta)^\top g(y)\}}{\kappa(\theta)},\quad y\in S \tag{1}$$

θ는 모수이고 $g(\bullet)$는 네트워크의 특성을 나타내는 충분통계량이다. 식 (1)과 같이 노드들의 연결 확률을 지수족의 형태로 표현할 수 있다면 모수에 의존하지 않는 충분통계량과 네트워크의 특성을 쉽게 찾을 수 있다. S는 y의 표본공간이고 $\eta(\bullet)$는 자연모수이다. 식 (1)을 모두 더하면 1이고 $\kappa(\theta)$는 정규화 상수로 식 (2)와 같다.

$$\kappa(\theta) = \sum_{y} \exp\{\eta(\theta)^{\top} g(y)\} \tag{2}$$

Frank, Strauss(1986)는 나머지 네트워크가 조건으로 주어지고 두 노드가 적어도 하나의 노드를 공유하고 있을 경우 확률적으로 종속이라는 마코브(Markov)가정 하에서 ERGM의 통계량을 계산했다. 하지만 분모의 정규화 상수 $\kappa(\theta)$의 계산이 어려워 추론 시 문제점이 발생한다. 이를 해결하기 위해 유사 우도 추정(Pseudo-likelihood estimation), 확률적 근사에 의한 최우 추정(MLE by stochastic approximation), 몬테카를로 최대화에 의한 최우 추정(MLE by Monte Carlo maximization), 베이지안 방법과 같이 여러 방법들이 고안되었고 본 연구에서는 추론 방법의 설명은 생략한다.

ERGM은 네트워크의 전체적인 특성을 모형화 하는데 적절하지만 몇 가지 문제점이 있다. 첫째, ERGM은 이해하기 어렵고 모형 퇴화(model degeneracy)와 같은 불필요한 특성을 가진다. 둘째, ERGM의 우도함수는 통계적으로 계산하기 복잡하고 어렵다. 셋째, 노드들 간의 관찰되지 않은 이질성(heterogeneity)이나 구조가 있을 수 있다.

이 중 ERGM의 가장 큰 문제점은 네트워크의 크기가 증가함에 따라 네트워크의 특정 부분에 확률이 집중되는 퇴화현상이다. 퇴화현상은 모형 부적합 문제를 야기할 수 있으며 개선하기 위한 방법들이 지속적으로 연구되고 있다. 이러한 ERGM의 여러 문제 때문에 잠재변수를 고려한 새로운 모형들이 제안되었다.

2.2 잠재변수 모형

잠재변수 모형은 크게 네 가지로 구분된다. 랜덤효과 모형(random effect model)과 혼합효과 모형(mixed effect model), 확률적 블록 모형(stochastic block model), 잠재 공간 모형(latent space model), 변형 방법으로 구분 된다. 잠재변수 모형 중 Wyatt et al.(2008), Koskinen(2009)을 제외하고 모든 잠재변수 모형은 연결 쌍(dyad)들의 조건부 독립을 가정한다. 만약 Z가 연속형 또는 이산형인 잠재변수를 나타내는 경우에 모형은 잠재변수가 주어졌을 때 연결 쌍(노드들의 고유한 비 순서쌍, $(Y_{i,j}, Y_{j,i})$)들의 조건부 독립을 가정한다.

$$P_\theta(Y=y|Z=z)=\prod_{(i,j)\in Y} P_\theta(Y_{i,j}=y_{i,j}, Y_{j,i}=y_{j,i}|Z=z)$$

방향을 고려한 네트워크는 연결 쌍들이 조건부 독립이라고 해서 연결선 $Y_{i,j}$와 $Y_{j,i}$가 항상 독립임을 의미하지 않는다. 상호성으로 인해 잠재변수가 조건으로 주어져도 $Y_{i,j}$와 $Y_{j,i}$가 종속이 될 수 있다. 그러나 대부분의 잠재변수 모형에서 연결선 $Y_{i,j}$들은 조건부 독립을 강하게 가정한다.

$$P_\theta(Y=y|Z=z)=\prod_{(i,j)\in Y} P_\theta(Y_{i,j}=y_{i,j}|Z=z) \tag{3}$$

잠재변수 모형의 장점으로는 첫째, 연결 쌍들의 조건부 독립을 가정하더라도 네트워크의 종속성을 파악할 수 있다. 잠재 공간 모형은 잠재변수 모형에서 개선된 모형으로 잠재변수를 효율적으로 사용해 상호성과 전이성(transitivity)같은 종속적 구조를 파악할 수 있다. 둘째, 연결선들의 조건부 독립 가정으로 모형 퇴화의 문제를 해결하여 모형 구축이 쉽다. 셋째, 마코브체인 몬테카를로(Markov Chain Monte Carlo, MCMC) 방법을 사용할 수 있어 계산상의 이점이 있다.

2.2.1 랜덤효과 모형과 혼합효과 모형

초기 잠재변수 모형은 van Duijn(1995)의 랜덤효과 모형이다. van Duijn의 랜덤효과 모형은 Holland, Leinhardt(1981)의 p_1모형에서 모형 절약성(parsimony) 부족으로 인해 제안되었다. p_1모형은 공변량들을 무시했지만 van Duijn은 공변량들을 확률변수로 고려하여 p_2모형을 제안했다. 모수 추정을 위해 van Duijn et al.(2004)은 p_1모형이 일반화 선형 모형(generalized linear model)으로 표현될 수 있는 점과 p_2모형이 일반화 선형 혼합 모형(generalized linear mixed model)으로 표현될 수 있는 사실을 활용했다. Zijlstra et al.(2009)은 p_2모형을 위한 베이지안 MCMC 방법을 개발했다. 계산에 필요한 방법은 R 패키지 eigenmodel을 통해 구현할 수 있다.

2.2.2 확률적 블록 모형

확률적 블록 모형은 Snijders, Nowicki(1997)에 의해 연구되었고 Nowicki, Snijders(2001)에 의해 확장되었다. 확률적 블록 모형은 노드들을 블록(block)이라 불리는 하위 집합으로 나눈다. 블록 구성원(block membership)이 조건부로 주어진 경우 연결선들은 독립이다. 블록 간의 연결 확률은 노드가 속한 블록에 의존한다. 또한 같은 블록에 속한 노드들의 연결 확률은 다른 노드와의 연결 확률보다 큰 값을 가진다. Tallberg(2005)는 블록 구성원을 예측하기 위해 공변량을 포함하였다. Airoldi et al.(2008)은 노드의 블록 구성원이 노드의 쌍에 의존하는 혼합 구성원 모형(mixed membership model)이라는 더 발전된 확률적 블록 모형을 제안했다. 즉, 노드가 속한 블록은 노드들의 연결선에 따라 달라진다는 개념이다.

확률적 블록 모형은 Nowicki, Snijders(2001)가 제안한 두 가지 가정에 기초한다.

첫째, 노드의 집합은 K개 블록으로 나눠지며 K는 알려진 고정 값이다.

$$Z_i|\pi_i,\dots,\pi_K \sim Multinomial(1;\pi_i,\dots,\pi_K) \tag{4}$$

둘째, $P_\theta(Y_{i,j}=1|Z_i=z_i,Z_j=z_j)=\theta_{z_i,z_j}$에서 $\theta_{k,l}$은 군집 k에 있는 노드가 군집 l의 노드와 연결될 확률을 의미한다.

확률적 블록 모형은 원래 이항 네트워크자료를 주로 분석하기 위해 개발되었지만 값이 있는 연결선(Mariadassou et al.,(2010)), 범주형 연결선(Jernite et al.,(2014))과 같은 자료의 분석도 가능하게 되었다(Bouveyron et al., 2016).

그러나 확률적 블록 모형은 블록의 레이블(label)에 따라 우도함수가 변하지 않아 모수를 식별할 수 없는 레이블 전환(label-switching) 문제점이 있다. 즉, 각각 다른 블록의 레이블이 같은 우도함수의 값을 가지는 것이다. 문제 개선을 위해 모수에 대한 순서를 제한하거나 두 노드가 같은 블록에 속해 있는지 알려주는 지시함수를 사용하는 방법 등 지속적으로 레이블 전환 문제에 대한 연구가 이뤄지고 있다.

2.2.3 잠재 공간 모형

잠재 공간 모형은 Hoff et al.(2002)에 의해 제안되었다. 네트워크자료의 확률적 모형이며 잠재적 위치는 일반적인 통계 이론들로 추정된다. 각 노드는 유클리디언(Euclidean) 공간에서 잠재적 위치(latent position)를 갖는다(Hoff et al., 2002). 즉, 각 노드는 알려지지 않은 잠재적 위치 Z를 가진다고 가정한다. 또한 이 모형을 이용해 전이성과 관측된 자료로부터 동질성을 찾을 수 있다. 잠재 공간 모형은 거리 공간을 유클리디언 공간으로 가정하는 경우(Hoff et al., 2002)와 울트라메트릭(ultrametric)으로 가정하는 경우(Schweinberger, Snijders, 2003)가 있다. 두 경우 모두 식 (5)의 형태로 표현된다.

$$\mathrm{logit}(P_\beta(Y_{i,j}=1|Z=z_i,Z=z_j))=\beta_0+x_{i,j}^\top\beta+d(z_i,z_j) \quad (5)$$

식 (5)에서 $x_{i,j}$는 연결 쌍(i,j)의 공변량을 나타내고 β_0는 네트워크의 밀도(density)를 조절하는 모수이며 $d(\bullet,\bullet)$는 i와 j가 가지는 잠재적 위치의 거리 함수이다. 잠재 공간 모형은 노드의 특성을 나타내는 발신자와 수신자 효과를 추가하여 Hoff(2005)에 의해 아래와 같은 식으로 확장되었다.

$$logit(P_\beta(Y_{i,j}=1|Z=z_i,Z=z_j))=\beta_0+x_{i,j}^\top\beta+d(z_i,z_j)+\delta_i+\gamma_j$$

δ은 발신자 효과를 의미하고 γ은 수신자 효과를 의미하며 두 효과는 정규분포를 가정한다. Handcock et al.(2007)은 잠재 공간 모형을 바탕으로 확장된 잠재적 위치 군집모형을 제안하였다. 잠재적 위치 군집모형이 잠재 공간 모형과 다른 점은 잠재적 위치를 K개의 원형 가우시안 군집의 혼합(mixture of K spherical gaussian clusters)으로 모형화 한 것이다.

$$Z_i \sim \sum_{k=1}^{K}\pi_k MVN(\mu,\sigma_Z^2 I)$$

본 논문에서는 Handcock et al.(2007)이 제안한 잠재적 위치 군집모형을 바탕으로 개선된 모수 추정을 위해 새로운 표본생성 방법을 제안한다.

2.2.4 변형 방법

베이지안 MCMC 방법은 잠재변수 모형에서 대부분 유용하게 사용하지만 속도 측면에서 느리다. 또한, 수 백 개의 노드를 가진 대용량 네트워크의 경우엔 적용할 수 없는 문제가 있다. 변형 방법은 베이지안 MCMC 방법의 근사적인 대안으로 빠르게 실현 가능하다. 이를 바탕으로 확률적 블록 모형의 근사 최우 추정은 Daudin et al.(2008)에 의해 도입되었고 변형된 EM 알고리즘을 제시하여 600개 이상의 노드에 적용했다. 근사 최우 추정의 일치성은 Nowicki, Snijders(2001)의 확률적 블록 모형을 고려한 Celisse et al.(2011)에 의해 확립되었다.

다양한 변형 방법 중 Salter-Townshend, Murphy(2013)는 잠재 공간 모형의 근사 베이지안 추정 방법을 제안하였고 80개 이상의 노드에 적용했다. 이 방법은 MCMC 방법을 통한 표본추출 대신 최적화(optimization) 방법을 사용하여 대용량 네트워크에도 적용가능하게 했다(Salter-Townshend, Murphy, 2013). 본 논문에서 이 방법을 새로운 표본생성 방법과의 모수 추정 결과 비교를 위해 사용했다.

제 3 장 잠재적 위치 군집모형 분석을 위한 추론 개선

본 장에서는 잠재적 위치 군집모형에 대한 소개와 베이지안 모형을 통한 MCMC 샘플링의 정확한 모수 추정을 위해 잠재적 위치 군집모형에서 기존의 표본생성 분포 대신 새로운 표본생성 분포를 제시한다. 잠재적 위치 군집모형을 기반으로 한 이유는 고전적 모형에 비해 이해하기 쉽고 고전적 모형에서 발생한 문제점들을 보완했기 때문이다. 또한, 노드들의 군집화까지 가능해 잠재적 위치 군집모형을 바탕으로 연구를 진행했다.

3.1 잠재적 위치 군집모형

잠재적 위치 군집모형(Latent Position Cluster Model, LPCM)은 Handcock et al.(2007)에 의해 제안되었다. 잠재적 위치 군집모형은 Hoff et al.(2002)의 잠재 공간 모형을 바탕으로 모형 기반 군집화(Fraley, Raftery, 2002) 개념을 적용한 모형이다. 이 모형은 전이성과 관측된 자료로부터 동질성을 찾을 수 있고 동시에 노드들의 군집화까지 가능하다.

잠재적 위치 군집모형은 d차원의 유클리디언 잠재 공간에서 노드들이 관찰되지 않은 랜덤 잠재 위치 z_i를 가지고 있다고 가정한다. 두 노드 사이의 연결 확률은 노드들의 잠재 위치가 주어진 경우 다른 연결선들과 독립이라고 가정하고 식 (6)과 같이 표현할 수 있다.

$$P(Y|Z,\beta)=\prod_{i\neq j}P(y_{i,j}|z_i,z_j,\beta) \tag{6}$$

Z는 모든 노드들의 잠재 위치를 나타내는 $n\times d$행렬로 $Z=(z_1,z_2,...,z_n)^\top$이다.

z_i는 $z_i = (z_{i1}, z_{i2}, ..., z_{id})$인 $1 \times d$벡터로 각 노드가 가지는 잠재 위치를 나타낸다. β는 추정해야 할 모수 중 하나로 스칼라 값을 가진다. 로지스틱 회귀 모형을 이용하여 식 (6)을 (7)과 같이 모형화 할 수 있다.

$$log-odds(y_{i,j} = 1 | z_i, z_j, \beta) = \beta - |z_i - z_j| \tag{7}$$

여기서 어떤 사건 A의 로그 오즈는 $log-odds(A) = log[P(A)/\{1-P(A)\}]$이며 식 (7)에서 β는 절편을 나타내고 $|z_i - z_j|$는 잠재 공간에서 노드 i와 j사이의 거리차이다. 모든 노드들의 확률은 다음과 같다.

$$P(Y|\beta, Z) = \prod_{i=1}^{n} \prod_{\substack{j=1 \\ j \neq i}}^{n} \left[\frac{\exp(\beta - |z_i - z_j|)}{1 + \exp(\beta - |z_i - z_j|)} \right]^{y_{i,j}} \left[\frac{1}{1 + \exp(\beta - |z_i - z_j|)} \right]^{(1 - y_{i,j})}$$

연결 확률은 β와 z_i에 의존하므로 정확한 β와 z_i의 추정이 본 연구의 목표이다. 잠재적 위치 군집모형이 잠재 공간 모형과 다른 점은 첫째, 노드들의 군집화를 위해 잠재 위치 z_i를 G개의 다변량 정규분포의 유한한 혼합으로부터 뽑는다. 각각의 다변량 정규분포는 군집별로 다른 평균과 공분산 행렬을 가진다.

$$z_i \sim \sum_{g=1}^{G} \lambda_g MVN_d(\mu_g, \sigma_g^2 I_d) \tag{8}$$

식 (8)에서 λ_g는 노드가 g번째 군집에 속할 확률로 $\lambda_g \geq 0 \, (g = 1, ..., G)$이고 $\sum_{g=1}^{G} \lambda_g = 1$이다. I_d은 $d \times d$ 단위행렬이다. 식 (8)는 Banfield, Raftery(1993)에 의해 제안되었으며 변수들의 군집화를 위한 모형이다.

둘째, 잠재 위치가 너무 큰 값을 가지지 않게 다음과 같은 제약조건을 설정한다.

$$\sqrt{\left(\frac{1}{n}\sum_i |z_i|^2\right)} = 1 \tag{9}$$

3.2 추론 개선

잠재적 위치 군집모형의 모수 추정에는 두 가지 방법이 있다. 첫 번째는 두 단계에 걸쳐 최우 추정치를 계산하는 방법이다. 먼저 군집화 되지 않은 잠재 공간 모형의 최우 추정치를 구한 뒤 혼합 모형에 대한 최우 추정치를 계산한다. 이 방법은 상대적으로 빠르고 간단하지만 잠재 위치를 추정할 때 군집의 정보를 이용하지 않는다. 두 번째는 MCMC 샘플링을 이용한 완전한 베이지안(fully Bayesian) 방법으로 잠재 공간과 군집화 모형을 동시에 추정한다. 첫 번째 방법보다 계산측면이나 수학적으로 더 까다롭지만 군집화에 관한 정보와 잠재 위치의 불확실성의 정보 손실을 막을 수 있다. Handcock et al.(2007)은 두 번째 방법을 사용하여 모수 추정을 실시하였다.

식 (6)-(9)와 MCMC 샘플링을 이용해 잠재적 위치 군집모형의 베이지안 추정을 실시한다. 여기서 i번째 노드가 어느 군집에 속해있는지 알려주는 새로운 변수 K_i를 이용한다. K_i는 군집 구성(group membership) 변수를 의미한다. 모수에 대한 사전 분포는 다음과 같다.

$$\begin{aligned}
\beta &\sim Normal(\xi, \psi^2) \\
(\lambda_1,...,\lambda_G) &\sim Dirichlet(\nu_1,...,\nu_G) \\
\mu_g &\sim MVN_d(0, \omega^2 I_d), \quad g = 1,...,G \\
\sigma_g^2 &\sim \sigma_0^2 Inv\chi_\alpha^2, \quad g = 1,...,G
\end{aligned}$$

ξ, ψ^2, $(\nu_1,...,\nu_G)$, σ_0^2, α, ω^2은 하이퍼 파라미터(hyperparameters)들로 Handcock et al.(2007)과 같이 $\xi=0$, $\psi^2=2I$, $\nu_g=3$, $\sigma_0^2=0.103$, $\alpha=2$, $\omega^2=2$로 설정했다. MCMC 알고리즘은 앞에서 정의한 사전 분포, 잠재 위치 z_i 그리고 군집 구성 변수 K_i를 이용하여 모수 추정을 반복한다. 모수 추정에는 깁스 샘플링(Gibbs sampling)과 메트로폴리스 헤스팅스(Metropolis-Hastings, MH)를 이용한다. z_i와 β의 경우 완전 조건부 사후분포를 구할 수 없어 MH알고리즘을 사용하고 남은 모수들은 깁스샘플링을 사용해 추정한다. 잠재 위치와 군집 구성 변수를 비롯한 수식에 명시되지 않는 모수들은 'others'로 표시한다. 완전 조건부 사후 분포는 아래와 같다. $\phi_d(\bullet\,;\mu,\Sigma)$는 d차원 다변량 정규분포의 확률밀도를 나타낸다.

$$z_i \mid K_i=g,\ others \propto \phi_d(z_i;\mu_g,\sigma_g^2 I_d)\,P(Y\mid Z,\beta), \qquad i=1,...,n \tag{10}$$

$$\beta \mid Z,\ others \propto \phi(\beta;\xi,\psi^2)\,P(Y\mid Z,\beta), \tag{11}$$

$$\lambda \mid others\ \sim\ Dirichlet(m+\nu), \tag{12}$$

$$\mu_g \mid others\ \sim\ MVN_d\left(\frac{m_g\bar{z}_g}{m_g+\sigma_g^2/\omega^2},\ \frac{\sigma_g^2}{m_g+\sigma_g^2/\omega^2}I\right), \qquad g=1,...,G \tag{13}$$

$$\sigma_g^2 \mid others\ \sim\ (\sigma_0^2+d\,s_g^2)\,Inv\,\chi^2_{\alpha+m_g d}, \qquad g=1,...,G \tag{14}$$

$$P(K_i=g\mid others)=\frac{\lambda_g\phi_d(z_i;\mu_g,\sigma_g^2 I_d)}{\sum_{r=1}^{G}\lambda_r\phi_d(z_i;\mu_r,\sigma_r^2 I_d)}, \qquad i=1,...,n,\ g=1,...,G \tag{15}$$

여기서 m_g, s_g^2과 $\bar{z}_g$는 다음과 같이 정의한다.

$$m_g=\sum_{i=1}^{n} I_{[K_i=g]},$$

$$s_g^2=\frac{1}{d}\sum_{i=1}^{n}(z_i-\mu_g)^{\top}(z_i-\mu_g)\,I_{[K_i=g]},$$

$$\bar{z}_g=\frac{1}{m}\sum_{i=1}^{n} z_i I_{[K_i=g]},$$

모수 추정을 위해 Handcock et al.(2007)에 의해 제안된 알고리즘은 다음과 같다.

Handcock's MH 알고리즘(2007):

t번째 반복에서 얻은 표본을 $Z^t=\left(z_1^t, z_2^t, ..., z_n^t\right)^\top$, $z_i^t=\left(z_{i1}^t, z_{i2}^t, ..., z_{id}^t\right)$, β^t, K_i^t, μ_g^t, $\sigma_g^{2(t)}$, λ_g^t라고 하자.

(단계 1) MH 알고리즘을 이용해 새로운 $Z^{t+1}=\left(z_1^{t+1}, ..., z_n^{t+1}\right)^\top$을 생성하고 각 노드는 임의의 순서로 업데이트 한다. $i=1,...,n$에 대하여,

(a) 표본생성분포 $MVN_d(z_i^t, \delta_Z^2 I_d)$로부터 새로운 z_i*를 생성한다.

(b) 채택확률을 다음과 같이 구한다.

$$\alpha_{z_i^t, z_i^*}=\min\left\{1, \frac{P(Y|Z^*, \beta^t)\phi_d(z_i^*; \mu_{K_i}, \sigma_{K_i}^2 I_d)}{P(Y|Z^t, \beta^t)\phi_d(z_i^t; \mu_{K_i}, \sigma_{K_i}^2 I_d)}\right\}$$

(c) 채택확률 $\alpha_{z_i^t, z_i^*}$의 확률로 $z_i^{t+1}=z_i^*$을 채택하고, $1-\alpha_{z_i^t, z_i^*}$의 확률로 $z_i^{t+1}=z_i^t$을 채택한다.

(단계 2) MH 알고리즘을 이용하여 새로운 β^{t+1}을 생성한다.

(a) 표본생성분포 $Normal(\beta^t, \delta_\beta^2 I_d)$로부터 새로운 β*를 생성한다.

(b) 채택확률을 다음과 같이 구한다.

$$\alpha_{\beta^t, \beta^*}=\min\left\{1, \frac{P(Y|Z^{t+1}, \beta^*)\phi_d(\beta^*; \xi, \psi^2)}{P(Y|Z^{t+1}, \beta^t)\phi_d(\beta^t; \xi, \psi^2)}\right\}$$

(c) 채택확률 $\alpha_{\beta^t, \beta^*}$의 확률로 $\beta^{t+1}=\beta^*$을 채택하고, $1-\alpha_{\beta^t, \beta^*}$의 확률로 $\beta^{t+1}=\beta^t$을 채택한다.

(단계 3) 식 (10)-(15)를 이용하여 남은 모수 K_i^t, μ_g^t, $\sigma_g^{2(t)}$, λ_g^t를 깁스샘플링(Gibbs sampling)을 이용하여 K_i^{t+1}, μ_g^{t+1}, $\sigma_g^{2(t+1)}$, λ_g^{t+1}로 업데이트 한다.

잠재적 위치 군집모형은 z_i와 β를 추정할 때 MH 알고리즘을 사용한다. (단계 1)에서 일정한 분산을 가지는 표본생성 분포를 이용했는데 이는 노드들이 각각 다른 잠재 위치와 군집을 가지고 있다는 정보를 고려하지 않은 것이다. 따라서 본 논문에서는 이 부분을 개선하기 위해 새로운 표본생성 방법을 제안한 뒤 기존의 방법들과 비교한다. β역시 MH 알고리즘을 사용해 추정하고 남은 모수들은 깁스 샘플링을 이용해 추정한다.

R에서의 잠재적 위치 군집모형의 모수 추정을 위한 알고리즘은 latentnet 패키지에 내장 된 ergmm함수를 사용하며 기본 형식은 다음과 같다.

```
ergmm(formula, response = NULL, family = "Bernoulli", fam.par = NULL,
  control = control.ergmm(), user.start = list(), prior = ergmm.prior(),
  tofit = c("mcmc", "mkl", "mkl.mbc", "procrustes", "klswitch"),
  Z.ref = NULL, Z.K.ref = NULL, seed = NULL, verbose = FALSE)
```

여기서 꼭 필요한 부분은 formula으로 모형 식을 써야 한다. 나머지는 옵션으로 상황에 따라 사용 여부가 달라진다. 옵션들은 사용하지 않을 경우 위에서 명시된 값이 기본으로 지정된다. ergmm의 기본적인 사용을 위한 코드 설명은 4장에서 설명한다.

본 논문에서는 개선된 모수 추정을 위해 Handcock et al.(2007)에서 사용한 표본생성 분포 대신 새로운 표본생성 분포를 제시한다. 새로운 표본생성 분포를 위해 본 연구는 일정한 분산이 아닌 각 군집별 분산을 이용하였다. 또한 식 (9)와 같이 잠재 위치 z_i에 관한 제약을 설정하지 않음으로 인해 더 유연하게 모수 추정을 하도록 하였다. 이를 바탕으로 본 연구는 개선된 모수 추정을 위한 새로운 MH 알고리즘을 다음과 같이 제안한다.

수정된 Handcock's MH 알고리즘:

t번째 반복에서 얻은 표본을 $Z^t = \left(z_1^t, z_2^t, ..., z_n^t\right)^\top$, $z_i^t = \left(z_{i1}^t, z_{i2}^t, ..., z_{id}^t\right)$, β^t, K_i^t, μ_g^t, $\sigma_g^{2(t)}$, λ_g^t라고 하자.

(단계 1) MH 알고리즘을 이용해 새로운 $Z^{t+1} = \left(z_1^{t+1}, ..., z_n^{t+1}\right)^\top$을 생성하고 각 노드는 임의의 순서로 업데이트 한다. $i = 1, ..., n$에 대하여,

(a) (단계 3)으로 부터 얻어진 $\sigma_{K_i}^2$를 이용한 표본생성분포 $MVN_d(z_i^t, \sigma_{K_i}^2 I_d)$로부터 새로운 z_i*를 생성한다.

(b) 채택확률을 다음과 같이 구한다.

$$\alpha_{z_i^t,\, z_i*} = \min\left\{1, \frac{P(Y|Z*, \beta^t)\phi_d(z_i*; \mu_{K_i}, \sigma_{K_i}^2 I_d)}{P(Y|Z^t, \beta^t)\phi_d(z_i^t; \mu_{K_i}, \sigma_{K_i}^2 I_d)}\right\}$$

(c) 채택확률 $\alpha_{z_i^t,\, z_i*}$의 확률로 $z_i^{t+1} = z_i*$을 채택하고, $1 - \alpha_{z_i^t,\, z_i*}$의 확률로 $z_i^{t+1} = z_i^t$을 채택한다.

(단계 2-3) Handcock's MH 알고리즘(2007)과 동일하다.

잠재 위치 z_i는 노드별로 다르고 노드가 속한 군집 역시 다르기 때문에 군집별 분산을 이용한 새로운 표본생성 분포를 이용하면 더 개선된 모수 추정이 기대된다.

제 4 장 모의실험

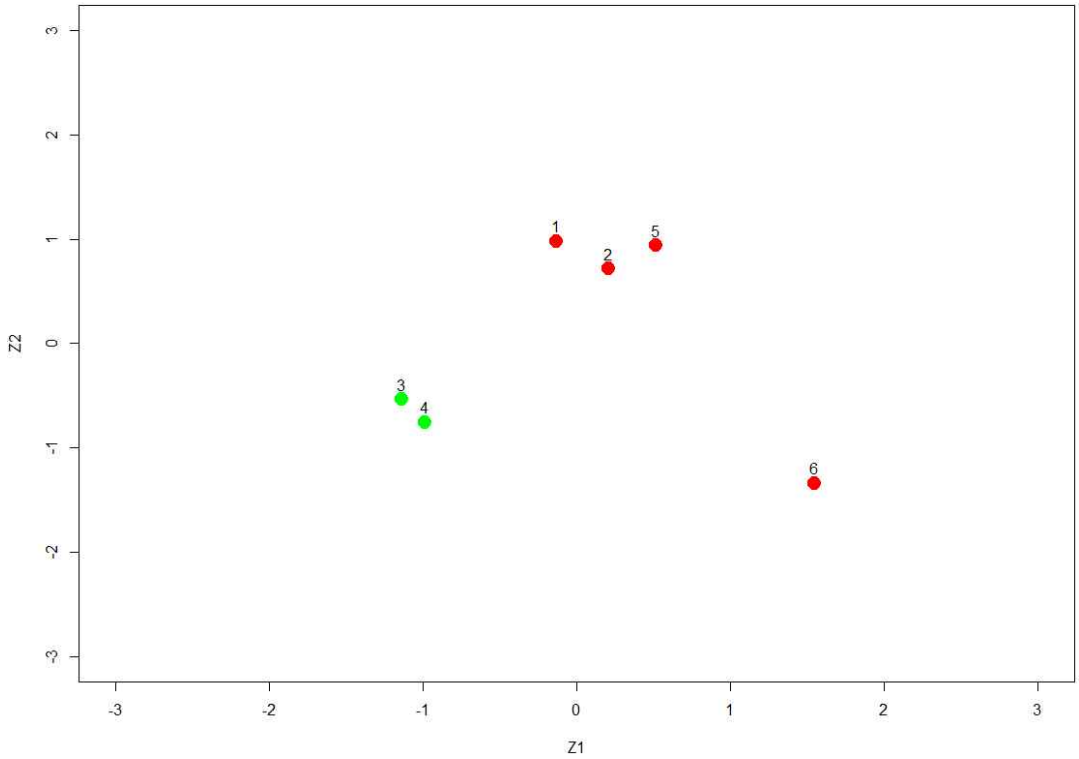
본 장에선 모의실험을 통해 세 가지 모수 추정 알고리즘을 비교하고자 한다. 비교에 사용된 알고리즘으로 잠재적 위치 군집모형에서 쓰인 Handcock's MH 알고리즘(HMH), Salter-Townshend, Murphy(2013)가 제안한 알고리즘(STM) 그리고 본 논문에서 제안한 수정된 Handcock's MH 알고리즘(MHMH)이다. 모의실험을 위한 통계 패키지로는 R을 사용하였다.

모의실험을 위해선 소셜네트워크자료인 $n \times n$행렬 Y를 생성해야 한다. 우선 노드와 군집의 수가 작은 자료에 먼저 적용하기 위해 2차원 공간을 바탕으로 6개의 노드와 2개의 군집으로 설정했다. 모수 추정 결과의 비교를 위해 설정한 모수의 참 값은 <표 1>과 같다.

<표 1> 모의실험에 사용된 모수의 참 값

β	μ_{11}	μ_{12}	μ_{21}	μ_{22}	σ_1^2	σ_2^2	λ_1	λ_2
1.5	3	3	0	0	0.1	0.5	0.3	0.7

추가적으로 잠재 위치 z_i도 결정해 주어야 한다. 이를 위해 <표 1>과 식 (8)을 바탕으로 생성했다. 이 때 잠재 위치는 랜덤하게 생성되며 그 중 하나를 사용하였다. 그 결과 생성된 잠재 위치 z_i의 산점도는 <그림 1>과 같다.



<그림 1> 랜덤 잠재 위치 z_i

모든 모수의 설정이 끝난 뒤 식 (7)을 이용해 노드들 간의 연결 확률이 0.5 이상이면 1로, 0.5 미만은 0으로 정의하였다. 그 결과 생성된 모의실험 자료는 <표 2>와 같다. 모의실험 자료 Y는 6×6인 행렬로 노드 간 관계유무가 0과 1로 되어있는 이항자료이다. 그리고 자신과의 관계는 허용하지 않아 대각원소는 0이다.

<표 2> 모의실험 자료 Y

Y	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	0	1	0	1	1	1
[2,]	1	0	0	0	1	1
[3,]	0	0	0	1	0	0
[4,]	1	0	1	0	0	0
[5,]	1	1	0	0	0	0
[6,]	1	1	0	0	0	0

모의실험 자료를 이용해 추정해야 하는 모수는 총 9개다. μ_{11}과 μ_{12}는 첫 번째 군집의 평균, μ_{21}과 μ_{22}는 두 번째 군집의 평균이다. σ_1^2과 σ_2^2은 각 군집의 분산을 나타낸다. 마지막으로 λ_1과 λ_2는 각 군집에 속할 확률이다. 각 군집의 평균들은 잠재 위치 z_i에 의해 결정되므로 z_i를 정확하게 추정하는 것에 초점을 두었다. 다음으로 모의실험 자료를 이용해 세 가지 알고리즘의 모수 추정 결과를 비교해 보았다.

4.1 Handcock's MH 알고리즘 (2007)

Handcock's MH 알고리즘(HMH)을 통한 모수 추정은 R의 latentnet패키지의 ergmm함수를 사용한다.

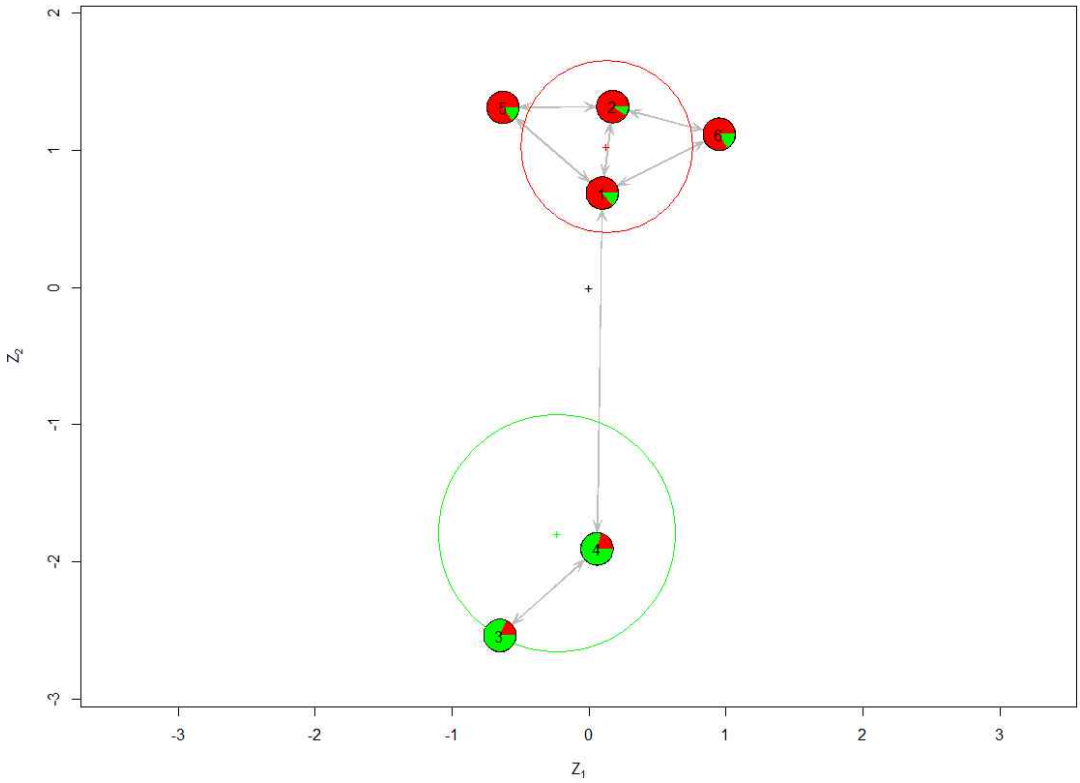
ergmm(y ~ euclidean(d = 2, G = 2), verbose = TRUE)

모형 적합에는 유클리디언 거리를 사용했다. y는 앞에서 생성한 모의실험 자료가 들어가고 공간은 2차원($d=2$) 그리고 군집 수는 2개($G=2$)이다. 그 뒤의 'verbose = TRUE'옵션은 모형 적합 과정을 보여주는 것으로 잘 적합 되고 있는지 확인하기 위해 사용했다. 필수적인 옵션은 아니므로 사용하지 않아도 된다.

모형 적합 후 잠재 위치의 산점도를 보고 노드들의 군집화를 확인할 수 있다. 이 때 필요한 명령어는 R에서 plot으로 다음과 같고 <그림 2>를 통해 확인할 수 있다.

plot(x , pie = TRUE , vertex.cex = 2.5)

x에는 ergmm을 통해 적합한 결과가 들어간다. 'pie = TRUE'는 옵션으로 각 노드가 현재의 군집에 속할 사후 확률(posterior probabilities)을 제공한다. 'vertex.cex = 2.5'는 노드의 크기를 조정하는 옵션이다.



<그림 2> HMH를 통해 적합한 잠재 위치 산점도

<그림 2>를 보면 총 노드의 수는 6개이고 x축과 y축은 각각 2차원 공간의 잠재 위치 Z_1과 Z_2이다. 노드(3,4)가 군집 1, 나머지 노드(1,2,5,6)들이 군집 2로 군집화 되었다. 모수 추정 결과는 <표 3>과 같다. 또한, 노드들이 군집에 속할 사후 확률을 제공한다. 각 노드에서 군집과 동일한 색의 부분이 클수록 군집에 속할 사후 확률이 큰 것을 의미한다.

<표 3> HMH를 통한 모수 추정 결과

$\hat{\beta}$	$\widehat{\mu_{11}}$	$\widehat{\mu_{12}}$	$\widehat{\mu_{21}}$	$\widehat{\mu_{22}}$	$\widehat{\sigma_1^2}$	$\widehat{\sigma_2^2}$	$\widehat{\lambda_1}$	$\widehat{\lambda_2}$
1.3985	-0.2127	-0.3486	0.2572	0.3695	0.1107	0.0873	0.4942	0.5058

4.2 Salter-Townshend, Murphy 알고리즘 (2013)

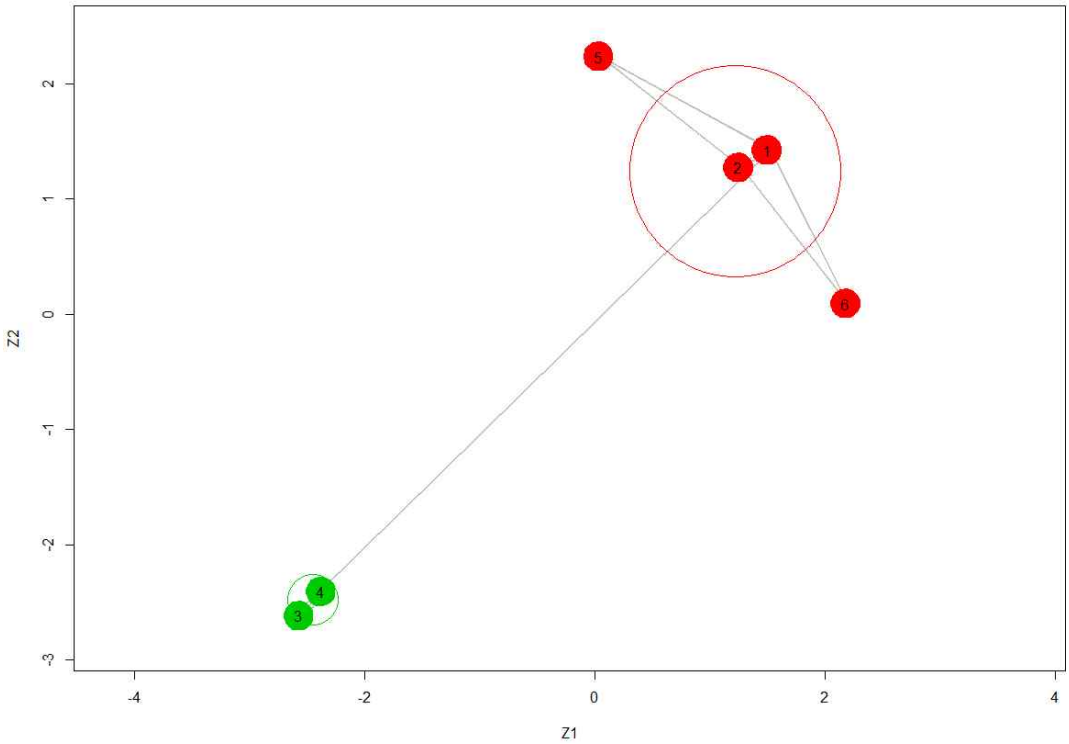
변형 방법 중 Salter-Townshend, Murphy(2013)가 제안한 알고리즘(STM)은 R의 VBLPCM패키지의 vblpcmfit함수를 이용한다. STM은 모수 추정 시 기존의 잠재적 위치 군집모형에 사용한 MCMC 방법이 아닌 근사적인 방법을 통한 최적화 방법을 사용한다. 그로 인해 대용량 네트워크에서도 빠른 속도로 적용 가능하다.

vblpcmfit((vblpcmstart(y, d = 2, G = 2)), STEPS = 20)

vblpcmfit함수에서 맨 처음 들어가는 문장은 모형 적합을 실시하기 위한 모수들의 초기 값이 들어가야 한다. 이를 위해 vblpcmstart함수가 사용된다. vblpcmstart함수는 빠르게 모수들의 값을 계산하며 안에 들어가는 y는 소셜네트워크자료이다. 잠재적 위치 군집모형과 마찬가지로 2차원($d=2$) 공간과 군집수는 2개($G=2$)이다. vblpcmfit함수에서 'STEPS = 20'옵션은 최대 반복 횟수를 지정하는 옵션이다. 잠재 위치의 산점도를 그리기 위한 명령어는 잠재적 위치 군집모형과 다르게 plot.vblpcm을 사용한다.

plot.vblpcm(x, R2 = 0.5, colours = 2:3, xlab = 'Z1', ylab = 'Z2')

x에는 vblpcmfit함수를 통해 적합한 결과가 들어간다. 'R2 = 0.5'는 각 노드의 크기를 지정해주고 'colours = 2:3'은 군집별 색상을 부여하는 옵션이다. 'xlab = 'Z1''과 'ylab = 'Z2''는 각각 x축과 y축의 이름을 지정한다.



<그림 3> STM을 통해 적합한 잠재 위치 산점도

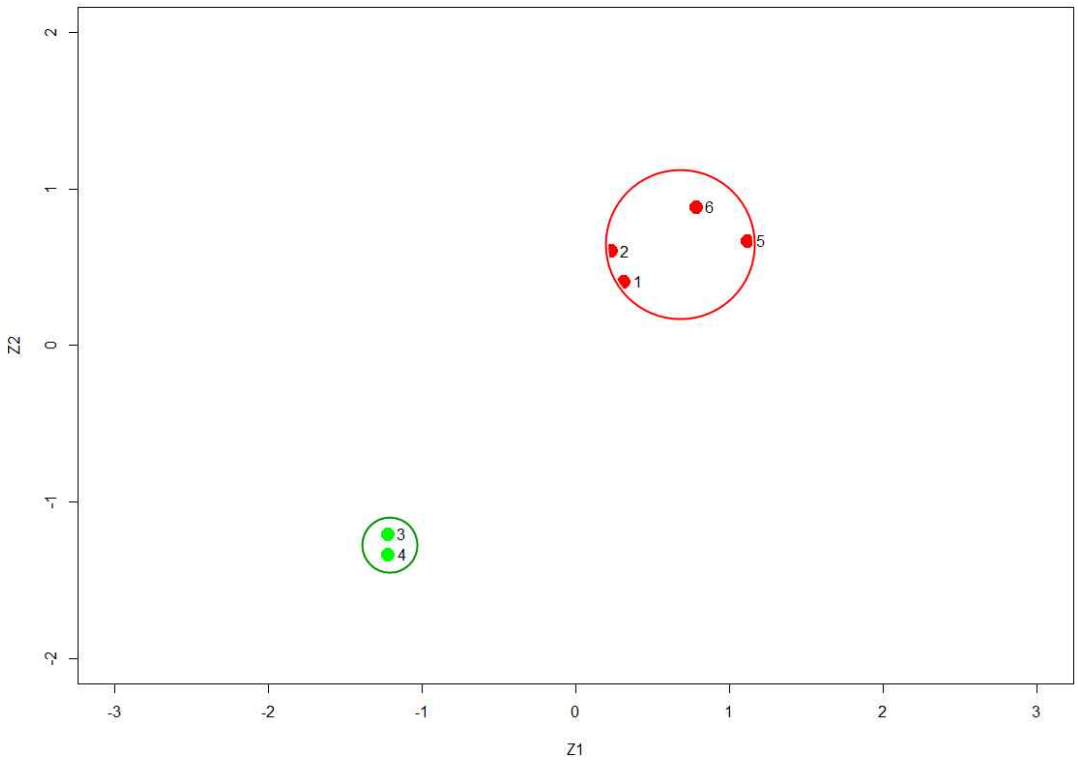
<그림 3>은 <그림 2>와 비슷한 형태를 보이지만 차이점은 각 노드가 현재의 군집에 있을 사후 확률을 제공하지 않는다. 또한, 빨간색으로 표시 된 군집 2의 경우 노드들이 군집의 평균과 분산을 바탕으로 그려진 원 안으로 밀집 된 정도가 약한 것을 확인할 수 있다. 모수 추정 결과는 <표 4>와 같다.

<표 4> STM을 통한 모수 추정 결과

$\hat{\beta}$	$\widehat{\mu_{11}}$	$\widehat{\mu_{12}}$	$\widehat{\mu_{21}}$	$\widehat{\mu_{22}}$	$\widehat{\sigma_1^2}$	$\widehat{\sigma_2^2}$	$\widehat{\lambda_1}$	$\widehat{\lambda_2}$
2.3023	0.3714	-0.0290	-0.8461	0.3304	0.0666	0.0796	0.6667	0.3333

4.3 수정된 Handcock's MH 알고리즘

수정된 Handcock's MH 알고리즘(MHMH)은 HMH에서 동일한 분산을 가지는 표본생성 분포를 군집별 분산을 가지는 새로운 표본생성 분포로 수정한 방법이다. 이를 바탕으로 더 개선된 모수 추정이 기대되며 추정된 잠재 위치 z_i의 산점도는 <그림 4>와 같다.



<그림 4> HMHM을 통해 적합한 잠재 위치 산점도

다만 패키지에서 제공하는 결과가 아니라 다른 두 모형과 비교해 시각적으로 부족하지만 군집화를 확인하기에는 부족함이 없고 모수 추정 결과는 <표 5>와 같다.

<표 5> MHMH를 통한 모수 추정 결과

$\hat{\beta}$	$\widehat{\mu_{11}}$	$\widehat{\mu_{12}}$	$\widehat{\mu_{21}}$	$\widehat{\mu_{22}}$	$\widehat{\sigma_1^2}$	$\widehat{\sigma_2^2}$	$\widehat{\lambda_1}$	$\widehat{\lambda_2}$
2.0878	1.3364	1.7339	-0.6787	-0.7479	0.2704	0.7275	0.5037	0.4963

세 가지 알고리즘의 모수 추정 결과와 평균제곱오차(Mean Squared Error, MSE)를 각각 <표 6>과 <표 7>에 정리했다.

<표 6> 모의실험 모수 추정 결과

모수	참 값	편의 (Bias)		
		HMH	STM	MHMH
$\hat{\beta}$	1.5	0.1015	-0.8023	-0.5878
$\widehat{\mu_{11}}$	3	3.2127	2.6286	1.6636
$\widehat{\mu_{12}}$	3	3.3486	3.0290	1.2661
$\widehat{\mu_{21}}$	0	-0.2572	0.8461	0.6787
$\widehat{\mu_{22}}$	0	-0.3695	-0.3304	0.7479
$\widehat{\sigma_1^2}$	0.1	-0.0107	0.0334	-0.1704
$\widehat{\sigma_2^2}$	0.5	0.4127	0.4204	-0.2275
$\widehat{\lambda_1}$	0.3	-0.1942	-0.3667	-0.2037
$\widehat{\lambda_2}$	0.7	0.1942	0.3667	0.2037
절대값 평균		0.9001	0.9804	0.6388

<표 6>은 각 모형을 총 10번 적합하여 추정한 모수들의 편의의 절대값 평균이다. 그 결과 제안된 방법(HMHM)이 기존의 방법들보다 평균적으로 작은 편의를 보였다. 또한, MCMC 방법을 이용한 알고리즘들(HMH, MHMH)이 최적화 방법(STM)을 이용한 것보다 정확한 모수 추정의 결과를 보였다.

<표 7> MSE 추정 결과

모수	MSE		
	HMH	STM	MHMH
β	0.1854	0.6673	0.4373
μ_{11}	8.3081	7.7074	3.0381
μ_{12}	9.4743	10.2531	1.9157
μ_{21}	1.7957	3.6412	0.6742
μ_{22}	2.2199	4.1442	1.4733
σ_1^2	0.6371	0.0014	0.0578
σ_2^2	0.7885	0.1809	0.3383
λ_1	0.3762	0.1344	0.0563
λ_2	0.3850	0.1344	0.0563
평균	**2.6856**	**2.9849**	**0.8941**

<표 7>은 알고리즘들의 MSE를 비교한 결과이다. <표 6>과 <표 7>의 결과를 통해 노드의 수가 적은 소셜네트워크 자료는 최적화를 통한 방법(STM)보다 MCMC 방법(HMH, MHMH)이 더 우수한 모수 추정이 이뤄진 것을 확인할 수 있었다. 또한 산점도와 MSE를 비롯한 결과로부터 본 연구에서 제안한 알고리즘이 기존의 알고리즘보다 더 효율적인 추정치를 제공하는 방법임을 알 수 있다. 이를 토대로 노드의 수가 증가한 실 자료에 적용하여 모수 추정 결과를 비교해 보았다.

제 5 장 실증분석

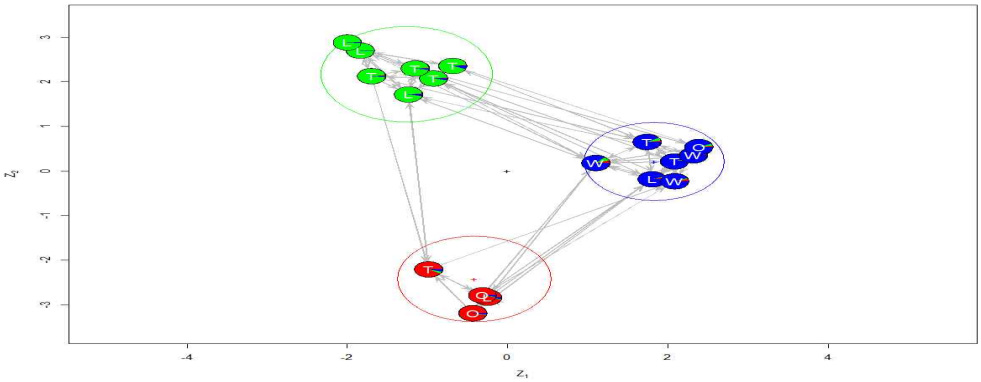
소셜네트워크분석에서 가장 대표적으로 사용되는 예제는 Sampson(1968)의 수도승 자료이다(Krivitsky, Handcock 2008). Sampson은 고립 된 미국 수도원에서 수도승들과 머무르면서 같이 생활하고 인터뷰 및 관찰을 통해 18명의 수도승 사이의 사회적 관계 자료를 얻었다. Sampson은 사회적 관계 중 다른 수도승에게 가지는 '선호도'에 집중하였다. 18명의 수도승에게 3번에 걸쳐 가장 선호하는 수도승 3명을 조사하였다. 각각의 자료는 'samplk1', 'samplk2', 'samplk3'에 저장되었으며 모두 이항자료이다. 또한 Sampson은 최종 수집한 자료를 바탕으로 3분류로 노드를 군집화 하였다. 젊은 터키인(the Young Turks), 충성스러운 야당(the Loyal Opposition) 그리고 추방자들(the Outcasts)로 나누었다.

분석에 사용한 자료는 'samplike'로 3번에 걸쳐 조사한 자료들의 합이다. 3개의 자료를 합한 뒤 이항 값을 가지는 소셜네트워크 자료로 전환했다. 그 결과 <그림 5>와 같이 0과 1로 이루어진 18×18 행렬이 생성되었다. 또한 자기 자신과의 관계는 제한하여 대각원소가 0을 가진다.

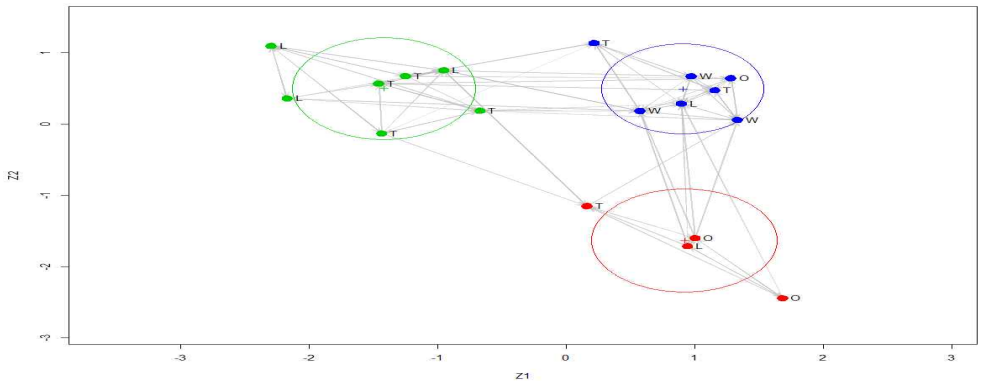
	John Bosco	Gregory	Basil	Peter	Bonaventure	Berthold	Mark	Victor	Ambrose	Romauld	Louis	Winfrid	Amand	Hugh	Boniface	Albert	Elias	Simplicius
John Bosco	0	1	1	0	1	0	0	1	0	0	0	1	0	1	0	0	0	0
Gregory	1	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	0	0
Basil	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1
Peter	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0
Bonaventure	1	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	0
Berthold	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0
Mark	1	1	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0
Victor	1	1	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0
Ambrose	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0
Romauld	0	0	0	1	1	0	0	1	1	0	0	0	1	1	0	0	0	0
Louis	0	0	0	1	1	0	0	1	0	0	0	0	0	1	0	1	0	0
Winfrid	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
Amand	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1
Hugh	1	1	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0
Boniface	1	1	0	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0
Albert	1	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0
Elias	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
Simplicius	1	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0

<그림 5> Sampson의 수도승 자료

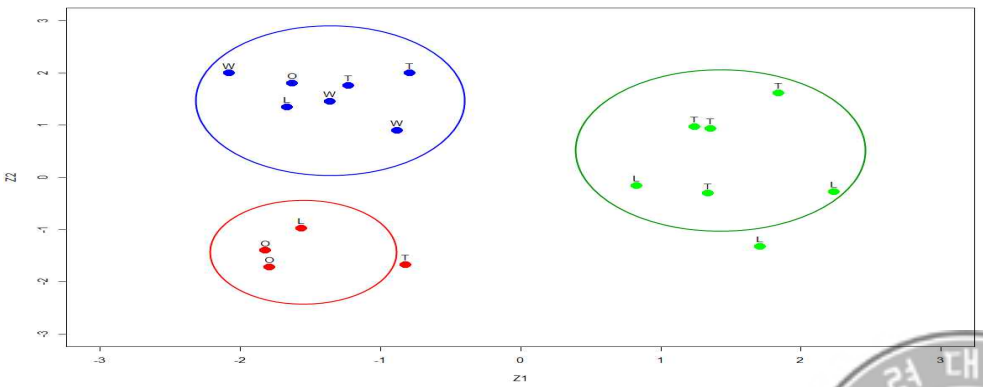
Sampson이 제공한 군집에 관한 정보를 고려하여 2차원 공간($d=2$)과 군집의 수는 3개($G=3$)로 설정하였다.



(a) HMH를 통한 잠재 위치 산점도



(b) STM을 통한 잠재 위치 산점도



(c) MHMH를 통한 잠재 위치 산점도

<그림 6>

각 군집은 T(the Young Turks), L(the Loyal Opposition), O(the Outcasts)로 표현하였고 군집화는 세 가지 방법 모두 우수하게 진행되었다. <그림 6>을 보면 HMH와 MHMH방법이 STM방법보다 군집을 나타내는 원 안으로 노드들이 더 밀집되어 있다. 이는 MCMC방법이 최적화 방법보다 우수하게 모수 추정을 수행하여 군집화가 잘 된 것을 의미한다. 다만 실 자료에서 모수의 참 값을 알 수가 없어 편의와 MSE를 구할 수 없었다. <표 8>은 실 자료의 모수 추정 결과를 비교한 것이다. MHMH방법을 통한 모수 추정 결과가 HMH와 STM방법의 결과와 크게 차이가 나지 않는 것을 확인할 수 있다.

<표 8> 수도승 자료 모수 추정 결과

모수	HMH	STM	MHMH
$\hat{\beta}$	1.1946	1.6967	1.4624
$\widehat{\mu_{11}}$	-0.3642	0.7098	-0.4824
$\widehat{\mu_{12}}$	-0.6699	1.1711	-1.5123
$\widehat{\mu_{21}}$	0.3880	1.5496	-0.1535
$\widehat{\mu_{22}}$	2.4463	-1.2558	0.2532
$\widehat{\mu_{31}}$	-3.1061	-1.5877	-0.2378
$\widehat{\mu_{32}}$	0.7911	-0.4254	0.2939
$\widehat{\sigma_1^2}$	0.1724	0.0345	0.3765
$\widehat{\sigma_2^2}$	0.2253	0.0453	0.2587
$\widehat{\sigma_3^2}$	0.3643	0.0286	1.5260
$\widehat{\lambda_1}$	0.3620	0.3889	0.1028
$\widehat{\lambda_2}$	0.3076	0.2222	0.1219
$\widehat{\lambda_3}$	0.3304	0.3889	0.7753

제 6 장 결 론

소셜네트워크 분석을 위해 여러 가지 통계적 모형이 사용되고 있다. 다만 소셜네트워크분석에 많이 사용되고 있는 고전적 모형인 지수족 랜덤 그래프 모형과 잠재변수를 고려한 모형은 모형이 복잡하고 모수 추정에 어려움이 많다. 이러한 이유로 모수 추정보단 군집화에 초점을 두는 경우가 많다. 하지만 본 논문에선 모수 추정을 정확하게 한다면 군집화 역시 효율적으로 진행될 것으로 기대하고 모수 추정에 초점을 두었다.

본 논문에선 기존의 알고리즘(HMH)이 일정한 분산을 갖는 표본생성분포를 사용한 것과 다르게 각 군집별 정보를 이용하는 새로운 표본생성 방법을 제시하였다. 즉, 노드별 군집의 분산을 고려한 분포에서 새로운 표본을 생성하는 방법을 적용했다. 또한 잠재 위치 z_i에 제약을 주지 않아 더 유연하게 모수 추정을 하도록 했다. 그 결과 제안된 방법은 모의실험을 통해 다른 방법보다 더 개선된 추정치를 제공하는 우수성을 보였다.

다만 새로운 표본생성 방법을 노드의 수가 많은 자료에 적용하기 위해선 충분히 반복해야 하는데 통계 프로그램 중 R을 사용하여 시간이 오래 걸리는 한계점이 존재하였다. 그럼에도 실 자료에 적용한 결과 제안한 방법의 모수 추정 결과가 HMH나 STM의 결과와 큰 차이가 나지 않았다. 따라서 만약 C나 Java와 같은 다른 프로그램을 사용해 구현한다면 대용량 자료에도 개선된 모수 추정의 결과를 얻어 효율적인 군집화가 가능할 것으로 기대된다.

참 고 문 헌

[1] Airoldi, E., Blei, D., Fienberg, S., Xing, E. (2008). Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research,* 9, 1981-2014.

[2] Banfield, J. D., Raftery, A. E. (1993). Model-based Gaussain and non-Gaussian clustering. *Biometrics,* 49(3), 803-821.

[3] Bouveyron, C., Latouche, P., Zreik, R. (2016). The Stochastic Topic Block Model for the Clustering of Networks with Textual Edges, *Statistics and Computing* (in press).

[4] Celisse, A., Daudin, J. J., Pierre, L. (2011). Consistency of maximum-likelihood and variational estimators in the Stochastic Block Model. *Electronic Journal of Statistics,* 6, 1847-1899.

[5] Daudin, J. J., Picard, F., Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing,* 18(2), 173-183.

[6] Fraley, C., Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association,* 97(458), 611-631.

[7] Frank, O., Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association,* 81(395), 832-842.

[8] Gormley, I. C., Murphy, T. B. (2010). A mixture of experts latent position cluster model for social network data, *Statistical Methodology,* 7(3), 385-405.

[9] Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks, *Journal of the Royal Statistical Society Series A*, 170, 301-354.

[10] 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.

[11] Hoff, P. (2005). Bilinear Mixed-Effects Models for Dyadic Data. *Journal of the American Statistical Association,* 100(469), 286-295.

[12] Holland, P. W., Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs(with discussion). *Journal of the American Statistical Association,* 76(373), 33-50.

[13] Hunter, D. R., Krivitsky, P. N., and Schweinberger, M. (2012). Computational Statistical Methods for Social Network Models, *Journal of Computational and Graphical Statistics,* 21(4), 856-882.

[14] Jernite, Y., Latouche, P., Bouveyron, C., Rivera, P., Jegou, L., Lamassé, S. (2014). The random subgraph model for the analysis of an acclesiastical network in merovingian gaul. Annals of Applied Statistics, 8(1), 55-74.

[15] Koskinen, J. (2009). The Linked Importance Sampler Auxiliary Variable Metropolis Hastings Algorithm for Distributions with Intractable Normalising Constants. *Working paper.*

[16] Krivitsky, P. N., Handcock, M. S. (2008). Fitting Position Latent Cluster Models for Social Networks with latentnet, *Journal of Statistical Software,* 24(5).

[17] Lusher, D., Koshinen, J., Robins, G. (2013). *Exponential Random Graph Models for Social Networks,* Cambridge : Cambridge University Press.

[18] Mariadassou, M., Robin, S., Vacher, C. (2010). Uncovering latent structure in valued graphs: a variational approach. *Annals of Applied Statistics,* 4(2), 715-742.

[19] Nowicki, K., Snijders, T. A. B. (2001). Estimation and Prediction for Stochastic Blockstructures. *Journal of the American Statistical Association,* 96(455), 1077-1087.

[20] Ryan, C., Wyse, J., Friel, N. (2017). Bayesian model selection for the latent position cluster model for social networks, *Network Science,* 5(1), 70-91.

[22] Salter-Townshend, M., Murphy, T. B. (2013). Variational Bayesian inference for the Latent Position Cluster Model for network data, *Computational Statistics and Data Analysis,* 57(1), 661-671.

[21] Sampson, S. (1968). A novitiate in a period of change: An experimental and case study

of social relationships, PhD thesis, Cornell University, September.

[22] Schweinberger, M., Snijders, T. A. B. (2003). Settings in Social Networks: A Measurement Model. *Sociological Methodology,* 33(1), 307-341.

[23] Snijders, T. A. B., Nowicki, K. (1997). Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification,* 14(1), 75-100.

[24] Tallberg, C. (2005). A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology,* 29(1), 1-23.

[25] van Duijn, M. A. J. (1995). Estimation of a Random Effects Model for Directed Graphs. in *Toeval zit overal: programmatuur voor random-coëfficiënt modellen,* eds. Snijders, T. A. B., Engel, B., Van Houwelingen, J. C., Keen, A., Stemerdink, G. J., Verbeek, M., Groningen: IEC ProGAMMA, 113-131.

[26] van Duijn, M. A. J., Snijders, T. A. B., Zijlstra, B. H. (2004). p_2: a random effects model with covariates for directed graphs. *Statistica Neerlandica,* 58(2), 234-254.

[27] Wasserman, S., Faust, K. (1994). *Social Network Analysis: Methods and Applications(Structural Analysis in the Social Sciences) First Edition,* Cambridge : Cambridge University Press.

[28] Wyatt, D., Choudhury, T., Bilmes, J. (2008). Learning Hidden Curved Exponential Random Graph Models to Infer Face-to-Face Interaction Networks from Situated Speech Data. *Proceedings of AAAI,* 2, 732-738.

[29] Zijlstra, B. J. H., van Duijn, M. A. J., Snijders, T. A. B. (2009). MCMC estimation for the p_2 network regression model with crossed random effects. *British Journal of Mathematical and Statistical Psychology,* 62(1), 143-166.