creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

이학석사학위논문

Overlapping Community Detection with Application to the 17th National Assembly of Republic of Korea

2020년 2월

서울대학교 대학원
통계학과
김성홍

Overlapping Community Detection with Application to the 17th National Assembly of Republic of Korea

지도교수 장 원 철

이 논문을 이학석사 학위논문으로 제출함

2019년 12월

서울대학교 대학원

통계학과

김성홍

김성홍의 석사 학위논문을 인준함

2019년 12월

위 원 장	오 희 석	(인)
부위원장	장 원 철	(인)
위 원	이 재 용	(인)

초 록

본 논문은 대한민국 17대 국회 공동발의 네트워크에서의 중첩 커뮤니티 탐지 방법에 대한 연구이다. 가중 네트워크 데이터에서의 중첩 커뮤니티 탐지 모형인 CPMw, Link community, EAGLE의 이론적 배경에 대해 설명하고, 대한민국 17대 국회 공동발의 네트워크에 모형을 적합시켜보았다. 적합된 모형들의 modified modularity, 커뮤니티 커버리지와 중첩 커버리지를 비교해 모형의 성능을 비교해보았다. 공동발의 지원행위는 국회의원간의 정치적 관계를 나타내기 때문에, 도출된 커뮤니티로부터 국회의원들의 관계망을 발견할 수 있다. 따라서, 최적 모형으로부터 도출된 커뮤니티를 국회의원들의 기존 정당과 비교해보고 이를 바탕으로 17대 국회의원들의 정당별 성격에 대해 알아보고자 한다.

주요어 : 네트워크, 중첩 커뮤니티, 대한민국 국회

학번 : 2018-24939

Contents

I. Introduction **1**

II. Statistical models **6**

2.1 Notation 6

2.2 CPMw 7

2.3 Link Community 8

2.4 EAGLE 10

III. Case study **13**

3.1 17대 국회 공동법안 발의 데이터 13

3.2 분석결과 14

IV. Conclusion **24**

Abstract **28**

그 림 목 차

그림 1. 방향 및 가중 네트워크 예시 2

그림 2. 17대 국회 공동발의 네트워크 15

그림 3. 시기별 17대 국회 공동발의 네트워크 16

그림 4. 시기 3 Link community Dendrogram과 분할 밀도 plot 18

그림 5. 시기 3 Link community 적합 결과 19

그림 6. 메타 커뮤니티 그래프 20

그림 7. 커뮤니티별 정당의 히스토그램 21

그림 8. 정당별 커뮤니티의 히스토그램 22

표 목 차

표 1. 시기별 노드와 엣지 갯수 14
표 2. Modified modularity 17
표 3. 커뮤니티 갯수 K와 커버리지 O^c, O^{ov} 17
표 4. 정당별 노드가 속한 커뮤니티 갯수에 따른 노드의 비율. 21

제 1 장

Introduction

네트워크 이론은 꼭지점인 노드(node)와 두 노드를 잇는 엣지(edge)를 구성요소로 하는 네트워크 자료분석 방법에 관한 내용을 다룬다. 네트워크 분석은 시각화를 통해 노드 쌍들의 관계를 살핌으로써 해당 네트워크의 관계들에 대한 직관적인 이해를 얻을 수 있으며 유전학, 생태학 또는 사회적 관계망 등 다양한 분야에서 응용이 되고 있다. 노드간의 관계인 엣지는 방향성을 가지거나, 0이나 1 이외의 가중치를 가질 수 있다. 그림 1처럼 엣지가 방향성을 가지는 경우 방향(directed) 네트워크라 정의하며, 그림 1의 엣지에 표현된 숫자와 같은 가중치를 가지는 경우를 가중(weighted) 네트워크라 한다.

일반적으로 커뮤니티는 사람 혹은 개체들의 군집으로 사회적 관계를 파악할 수 있는 중요한 요소이다. 네트워크 이론에서는 노드의 군집을 커뮤니티로 정의하고, 이 군집을 발견하는 방법을 커뮤니티 탐지(community detection)이라고 한다. 커뮤니티 탐지 모형에 대한 많은 연구는 커뮤니티의 노드들이 서로 겹치지 않는(disjoint) 네트워크를 가정한다. 그러나 실상은 대부분의 네트워크에서 개체들이 여러 개의 커뮤니티에 동시에 속할 때가 많다. 본 연구에서는 노드가 단순히 하나의 커뮤니티에만 속하는 경우에서 더욱 확장해 여러 개의 커뮤니티에 동시에 속하는 중첩 커뮤니티 탐지(overlapping community detection)모형을 적용해보고자 한다.

중첩 커뮤니티 탐지와 관련한 선행 연구들은 크게 클릭 여과(clique

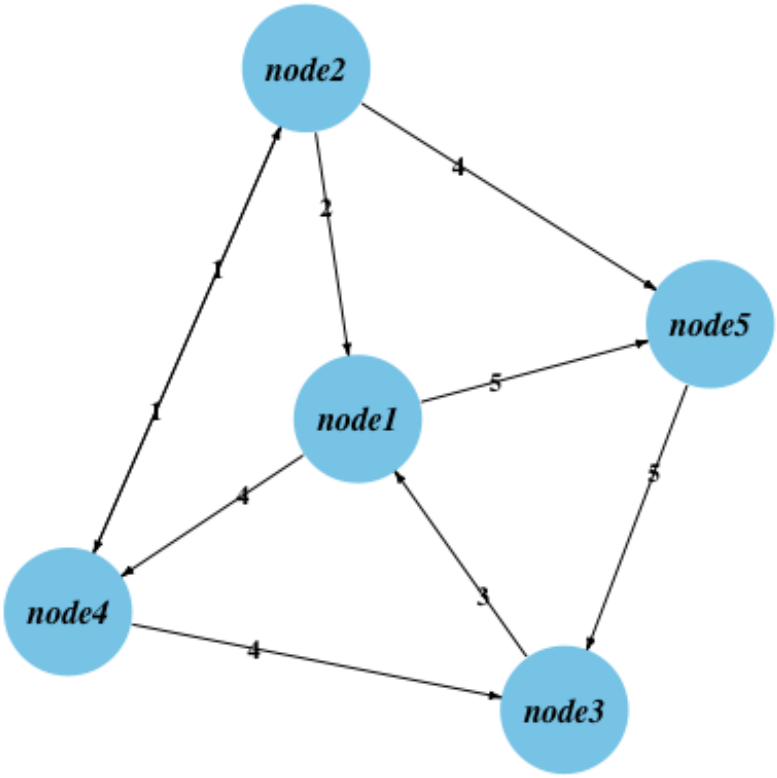


그림 1: 방향 및 가중 네트워크 예시

percolation), 선 분할(line partitioning), 국소 확장(local expansion), 퍼지 탐지(fuzzy detection)와 동적 알고리즘(dynamic algorithms)으로 분류된다. 클릭 여과 방법은 내부의 모든 노드들이 엣지로 연결되어 있는 sub-graph인 클릭을 이용해 커뮤니티를 탐지하는 방법이다. 대표적 방법인 CPM(Clique Percolation Method)는 내부의 노드가 k개인 k-클릭을 찾고 $k-1$개의 노드를 공유하는 k-클릭의 쌍들을 하나의 커뮤니티로 정의한다(Palla et al., 2005). Farkas et al. (2007)은 가중 네트워크에서 k-클릭 내

부 엣지들의 가중치를 고려한 밀도(intensity)를 정의한 후, 밀도가 기준치(threshold)보다 높은 k-클릭만 커뮤니티로 간주하는 CPMw(Clique Percolation Method Weighted)를 제시하였다. 선 분할은 노드의 커뮤니티가 아닌, 엣지의 커뮤니티를 찾음으로써, 하나의 노드에 연결되는 엣지들이 다른 커뮤니티에 속하는 경우를 이용해 해당 노드의 중첩 커뮤니티를 도출해낸다. 선 분할 방법 중 하나인 Link community는 두 엣지간 유사도(Similarity)를 기반으로 계층적 군집화 방법을 적용해 커뮤니티를 발견하는 방법이다(Ahn et al., 2010). Evans et al. (2009)는 기존 네트워크에서 노드와 엣지를 맞바꾸어 가중 선 그래프(weighted line graph)로 변환한 후, 서로 분리된 커뮤니티를 찾아내는 방식을 이용하였다. Brutz et al. (2015)는 엣지들의 지역 집합(local sets)을 노드로 간주하고 개별 노드, 커뮤니티, 그리고 네트워크 수준에서의 3가지 척도를 정의한 후, 이 척도를 최소화하는 유연 다중척도(flexible multiscale) 모형을 제시하였다. 국소 확장은 초기(seed) 커뮤니티와 커뮤니티의 성능측도인 목적함수(objective function)를 정의하고, 최적화 과정을 거쳐 초기 커뮤니티를 확장해가는 방식이다. Lancichinetti et al. (2009)은 국소 확장의 한 방법으로 초기 노드를 임의로 하나를 고른 후, 목적함수를 최대화할 때까지 다른 노드들을 커뮤니티에 추가시키는 알고리즘을 제시하였다. 일단 하나의 커뮤니티가 형성이 되면 남은 노드 중에 다시 하나를 임의로 선택해 과정을 반복한다. EAGLE(EagglomerativE hierarchicAl clusterinG based on maximaL cliquE)은 다른 클릭에 포함되지 않는 최대 클릭(maximal clique)을 초기 커뮤니티로 정의하고, 커뮤니티 집합의 성능 측도인 Extended modularity를 목적함수로 설정하여 중첩 커뮤니티를 발견한다(Shen et al., 2009). Xing et al. (2015)은 국소 구조 정보(local structural information)를 이용해 노드들을 작은 규모의 국소 커뮤니티에 할당하여 초기 커뮤

니티로 설정한 후, 커뮤니티 연결성을 측도로 하여 커뮤니티를 확장시키는 OCDLCE(Overlapping Community Detection algorithm by Local Community Expansion)를 제시하였다. 퍼지 탐지 방법의 경우, 전체 커뮤니티 숫자를 초기값으로 제한하고, 각각의 노드가 특정 커뮤니티에 속하는 정도를 나타내는 belonging factor를 기반으로 커뮤니티를 찾아내는 방법이다. 대표적으로 SPAEM(Simple Probabilistic Algorithm for detecting community structure which employs Expectation-Maximization)모형에서는 belonging factor의 다항분포를 가정하고 엣지의 확률분포를 belonging factor들의 혼합모형(Mixture model)으로 설정해 EM 알고리즘을 통해 belonging factor를 추정한다(Ren et al., 2009). 동적 알고리즘에서 가장 많이 쓰이는 방법은 COPRA(Community Overlap PRopagation Algorithm)와 SLPA(Speaker-listener Label Propagation Algorithm)이다. COPRA는 각 단계별로 노드들의 belonging factor를 이웃 노드들의 belonging factor의 평균으로 업데이트한 후, 특정값으로 수렴할 때까지 반복한다(Gregory, 2010). SLPA는 speaker-listener 과정을 통해 단계별로 노드간의 상호작용 규칙(interaction rule)에 따라 각각의 노드를 커뮤니티에 할당하며, 단계마다 어떤 커뮤니티에 속해있는지를 바탕으로 belonging factor를 설정한다(Xie et al., 2011).

많은 중첩 커뮤니티 탐지 모형은 엣지가 0, 1 값만 가지는 이진수일 때인 비가중 네트워크에서의 모형이다. 그러나 현실에서는 네트워크들이 가중 네트워크를 가정할 때 실제 관계망을 더 잘 반영하는 경우들이 있다. 예를 들어, 국회의원의 공동 발의 법안 네트워크에서 국회의원간의 공동발의행위(Cosponsorship) 여부만을 엣지로 설정하는 것보다는 공동발의 지지 횟수를 엣지로 설정하는 것이 관계망을 더욱 잘 반영할 것이다. 또한 국회의원들은 소속 정당 의원들과의 커뮤니티와는 별개로 개인적 친분 혹

은 정치적 이유로 인해 또 다른 커뮤니티를 형성하고 있을 가능성이 높다. 본 연구에서는 이러한 경향을 반영하기 위해, 가중 네트워크에서의 중첩 커뮤니티 탐지 모형인 CPMw, Link community, EAGLE에 대한 이론을 소개하고 이러한 방법을 17대 국회의원 공동발의 네트워크에 적용해보도록 한다.

본 논문의 나머지 부분은 다음과 같이 구성된다. 2장에서는 네트워크 이론의 수식적 개념들을 먼저 소개하고 모형들을 설명한다. 3장에서는 17대 국회의 공동발의 네트워크 데이터를 소개한 후, 2장에서 소개한 모형들의 적합과 시각화 결과를 제시한다. 각 모형의 성능을 비교한 후, 최적 모형으로부터 도출된 커뮤니티와 실제 정당을 비교하고 이를 바탕으로 결과에 대한 해석을 진행한다. 이후 결론으로 4장을 마친다.

제 2 장

Statistical models

2.1 Notation

네트워크 혹은 그래프는 수식적으로 $G = \{V, E\}$으로 나타낼 수 있으며, 여기서 V는 n개의 노드 집합, E는 V의 노드를 있는 m개의 엣지 집합으로 가정한다. 네트워크의 구조는 비가중 네트워크일 경우 $n \times n$ 인접 행렬 (Adjacency matrix) A, 가중 네트워크일 경우 가중치 행렬 W에 의해 정의된다. A의 원소 a_{ij}는 i 노드와 j 노드를 잇는 엣지가 존재할 때 1, 아니면 0값을 가진다. W의 원소 w_{ij}는 i와 j의 연결 강도를 나타내는 가중치로 음이 아닌 실수값을 가진다. 엣지의 방향성도 고려하는 방향 네트워크는 A, W가 비대칭행렬이나 본 연구에서는 비방향 네트워크인 경우만 고려하기 때문에 A, W는 대칭행렬을 가정한다.

중첩 커뮤니티 탐지 방법에서 전체 커뮤니티 갯수를 K개라고 했을 때, 커뮤니티들의 집합(cover)을 $C = \{c_1, c_2, \cdots, c_K\}$로 정의한다. 노드 i는 K개의 커뮤니티에 속해있는 정도인 belonging factor $b_{i1}, \cdots, b_{iK}$를 가지며, b_{ic}는 노드 i가 커뮤니티 c에 속해있는 정도이다. b_{ic}에 대해서는 다음과 같은 가정을 한다.

$$0 \leq b_{ic} \leq 1, \quad \forall i \in V, \forall c \in C,$$
$$\sum_{c=1}^{K} b_{ic} = 1.$$

b_{ic}가 0과 1값을 가지는 경우 crisp하다고 정의하며, 0과 1 이외의 값을 가질 때, fuzzy하다고 정의한다.

2.2 CPMw

CPM은 Palla et al. (2005)에 의해 제시된 방법으로, 비가중 네트워크에서 k개의 노드로 구성되며, 내부의 모든 노드들이 연결되어 있는 subgraph인 k-클릭들이 인접하는(adjacent) 경우를 커뮤니티로 정의한다. 2개의 k-클릭이 $k-1$개의 노드를 공유할 때, 인접하다고 한다. k-클릭 커뮤니티는 연속적으로 인접한 k-클릭들로 이루어진 커뮤니티로, k-클릭에 동시에 속해있을 수 있기 때문에, 중복된 커뮤니티가 가능하다.

Palla et al. (2005)는 가중 네트워크에서 기준치 T^*를 도입해 엣지의 가중치가 T^*미만이면 엣지를 제거하고, 남은 엣지들은 전부 1로 취급해 비가중 엣지로 변환하는 방식을 제시하였다. 하지만 이러한 방법은 가중치가 높은 엣지와 낮은 엣지가 이웃하는(disassortative) 경우, 가중치가 높은 엣지만 남게 되어 커뮤니티를 발견하지 못한다는 단점이 있다.

Farkas et al. (2007)는 k-클릭에 밀도(intensity)란 개념을 도입해 CPMw을 제시하였다. k-클릭의 밀도는 k-클릭에 속한 노드들의 엣지 가중치의 기하평균으로 다음과 같다

$$I(c) = \left(\prod_{\substack{i<j \\ i,j \in c}} w_{ij} \right)^{\frac{2}{K(K-1)}} .$$

CPMw는 밀도가 기준치 I 이상인 k-클릭만을 커뮤니티에 포함을 시킨다. 가중치가 낮은 엣지를 단순히 제거하는 것이 아니라 커뮤니티의 전

체 엣지도 고려함으로써, 가중치가 높은 엣지와 낮은 엣지가 이웃한 경우도 커뮤니티를 적절히 발견해낼 수 있다.

CPMw는 클릭 크기 k와 밀도 기준 I를 조절모수로 갖는다. I가 너무 크면 커뮤니티가 생성되지 않으며, 반면에 I가 너무 작은 경우에는 하나의 지나치게 큰 커뮤니티가 생성될 수 있다. 적당한 크기의 I를 선택하기 위해 Farkas et al. (2007)는 다음과 같은 방법을 제시하였다. 네트워크의 모든 엣지의 가중치들의 순서통계량을 $w_{(m)} \geq w_{(m-1)} \geq \cdots \geq w_{(1)}$이라고 했을 때, 먼저 $I = w_{(m)}$로 설정을 한 후, 가장 큰 두개의 커뮤니티 크기 비율 n_1/n_2이 2가 되는 지점인 여과 전이 지점(Percolation transition point)까지 I를 가중치의 순서통계량을 따라 줄여나간다. 이렇게 k값마다 최적의 I^*를 찾은 후, 가장 고른 커뮤니티 크기의 분포를 주는 k를 선택한다.

2.3 Link Community

Ahn et al. (2010)은 엣지의 유사도를 기반으로 계층적 군집화 방법을 적용해 엣지의 커뮤니티를 찾는 알고리즘인 Link community를 제시하였다. Link community는 노드 대신 엣지의 커뮤니티를 찾는 방법인 선 분할 기반의 방법이다. 노드에 연결된 엣지들이 다른 커뮤니티에 속할 경우 해당 노드는 중복된 커뮤니티에 속하게 된다.

Link community는 k 노드를 공유하는 엣지 e_{ik}와 e_{jk}의 유사도가 노드를 공유하지 않는 엣지들보다 더 높은 유사도를 가질 거라는 가정하에서, e_{ik}와 e_{jk}의 유사도 $S(e_{ik}, e_{jk})$를 계층적 군집화를 위한 dendrogram의 시작점으로 설정한다. 비가중 네트워크일 때, $S(e_{ik}, e_{jk})$는 다음과 같이 Jaccard 유사도로 정의한다

$$S_J(e_{ik}, e_{jk}) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|},$$

여기서 $|\cdot|$은 집합의 원소의 갯수인 cardinality를 의미하며, N_i 는 i 노드와 인접해 있는 노드 집합(neighborhood)이다. 가중 네트워크와 방향 네트워크의 경우, $S(e_{ik}, e_{jk})$를 Jaccard index의 확장형인 Tanimoto coefficient로 정의해 알고리즘을 적용할 수 있다. 먼저 노드 i에 대하여 벡터 $\vec{a}_i = (\tilde{A}_{i1}, \cdots \tilde{A}_{i|N_i|})$를 다음과 같이 정의한다

$$\tilde{A}_{ij} = \frac{1}{|N_i|} \sum_{i' \in N_i} w_{ii'} \delta_{ij} + w_{ij},$$

여기서 δ_{ij}는 $i = j$일 때 1, 아니면 0이다. Tanimoto 유사도 $S_T(e_{ij}, e_{jk})$는 다음과 같이 정의된다

$$S_T(e_{ij}, e_{jk}) = \frac{\vec{a}_i \cdot \vec{a}_j}{|\vec{a}_i|_2^2 + |\vec{a}_j|_2^2 - \vec{a}_i \cdot \vec{a}_j}.$$

위와 같이 정의된 Tanimoto 유사도를 기반으로 dendrogram을 만들어 계층적 군집화 모형을 적용할 수 있다. 계층적 군집화 모형에는 single linkage 방법이 일반적으로 쓰이지만, complete linkage나 average linkage로도 적용이 가능하다. 분할 지점을 정해 Dendrogram을 분할하면 Link community가 생성이 된다.

적절한 분할지점을 정하기 위해 분할지점의 성능을 측정하는 분할밀도 D(partition density)를 정의한다. $\{c_1^e, \cdots, c_K^e\}$를 분할지점에서의 엣지들의 커뮤니티라고 했을 때, c_k^e의 엣지 갯수는 $m_k = |c_k^e|$이며, 노드의 갯수는 $n_k = |\bigcup_{e_{ij} \in c_k^e} \{i, j\}|$이다. 이 경우, m_k를 최댓값, 최솟값으로 정규화한 값을 D_k라고 했을 때, D_k는 다음과 같이 정의된다

$$D_k = \frac{m_k - (n_k - 1)}{n_k(n_k - 1)/2 - (n_k - 1)} = \frac{1}{2} \frac{m_k - (n_k - 1)}{(n_k - 2)(n_k - 1)}.$$

분할 밀도 D는 $M = \sum_k m_k$라고 정의할 때, m_k/M을 가중치로 하는 D_k의 가중평균으로 나타낼 수 있다

$$D = \sum_k \frac{m_k}{M} D_k = \frac{2}{M} \sum_k m_k \frac{m_k - (n_k - 1)}{(n_k - 2)(n_k - 1)},$$

D의 최대값은 1이며, 이 때 모든 커뮤니티는 완전 연결된(fully-connected) 클릭이다. $D = 0$일 때, 모든 커뮤니티는 개별적인 트리가 된다. 커뮤니티 내의 subraph들이 연결되지 않는 요소를 가질 때 D는 음수가 되며 최소값은 $-2/3$이다. Dendrogram의 분할은 D를 최대화 하는 지점에서 이루어진다.

Link community의 장점은 계층적인 커뮤니티와 중복되는 커뮤니티를 동시에 발견할 수 있다는 점이다. 현실에서 커뮤니티는 상, 하위 커뮤니티로 이루어지는 경우가 많으나, 계층적 커뮤니티는 일반적으로 중복된 커뮤니티 분할을 허용하지 않는다. Link community는 엣지의 계층적 커뮤니티를 발견함과 동시에 노드의 중복된 커뮤니티를 발견함으로써 계층적 커뮤니티의 구조적 단점을 보완할 수 있다.

2.4 EAGLE

Shen et al. (2009)은 다른 더 큰 클릭에 포함되지 않는 클릭인 최대 클릭을 기반으로 한 계층적 군집에서 국소 팽창 방법을 이용해 커뮤니티를 발견하는 알고리즘인 EAGLE을 제안하였다. EAGLE은 Link community와 마찬가지로 dendrogram을 구성한 이후에 커뮤니티 집합의 성능 측도인 extended modularity를 최대화하는 지점에서 분할해 커뮤니티를 생성한다. Extended Modularity란 도출된 중첩 커뮤니티 집합의 성능을 측정

하기 위해 집합내 커뮤니티 내부 엣지의 연결 정도와 무작위로 구성된 엣지의 연결 정도를 비교한 측도로, 중첩되지 않은 커뮤니티에서 쓰이는 성능 측도인 modularity의 확장형이며 다음과 같이 정의된다.

$$Q(C) = \frac{1}{2m} \sum_{c_i \in C} \sum_{v,w \in c_i} \frac{1}{O_v O_w} \left(a_{vw} - \frac{k_v k_w}{2m} \right),$$

여기서 a_{vw}는 인접 행렬의 (i, j)번째 원소이며, k_v는 v 노드의 degree로 v 노드로 연결되는 엣지의 갯수를 의미한다. O_v는 v 노드가 속한 커뮤니티의 갯수이다.

EAGLE을 이용하여 커뮤니티를 구하는 알고리즘은 다음과 같다.

1. 모든 최대 클릭을 찾아 커뮤니티 초기값으로 설정한다. 이 때, 최대 클릭의 크기가 임계치 k 이상인 경우만 고려하고 클릭에 포함되지 않는 노드도 커뮤니티 초기값에 포함을 시킨다. 모든 커뮤니티 쌍의 유사도를 계산한다.

2. 가장 큰 유사도를 가지는 커뮤니티 쌍을 찾아 하나의 커뮤니티로 합치고 다시 커뮤니티 간의 유사도를 계산한다.

3. 하나의 커뮤니티가 남을 때까지 1과 2 과정을 반복한다.

두 커뮤니티 c_1, c_2 간의 유사도 M은 다음과 같이 정의되었다.

$$M(c_1, c_2) = \frac{1}{2m} \sum_{v \in c_1, w \in c_2, v \neq w} \left(a_{vw} - \frac{k_v k_w}{2m} \right).$$

EAGLE은 기본적으로 비가중, 비방향 네트워크에서 쓰이는 방법으로 가중 네트워크에서는 가중치에 기준치를 설정한 후, 기준치 이상인 엣지는 1, 미만인 경우 0으로 엣지를 변환해 비가중 네트워크로 전환한다. 이 때,

엣지의 임계치는 $Q(C)$를 최대화하는 수로 설정한다.

EAGLE은 최대 클릭의 위계적 구조를 고려함과 동시에 한 커뮤니티에 속하는 클릭 안에 노드들의 중복을 허용함으로써, Link community와 마찬가지로 노드의 계층적 구조와 중복된 커뮤니티를 동시에 발견할 수 있다는 장점이 있다.

제 3 장

Case study

3.1 17대 국회 공동법안 발의 데이터

본 연구에서 사용된 데이터는 국회의안정보시스템에서 구할 수 있는 17대 국회의원들의 당적 데이터와 대표발의자, 공동발의자 정보를 포함하는 법안 발의 데이터이다. 대표발의자란 법안을 발의하는 국회의원을 말하며, 공동발의자는 법안을 지지하는 국회의원을 말한다. 대한민국 국회에서 법안의 발의가 유효하기 위해서는 대표발의자를 포함해 최소 10명 이상의 국회의원이 지지를 해야하며, 이러한 지지행위를 공동발의라고 정의한다. 법안에 대한 공동발의 네트워크를 이용하여 커뮤니티를 발견한다면 정당 이외의 국회의원들간 정치적 네트워크를 유추할 수 있다.

공동발의 데이터는 국회의원을 노드로 하고, 국회의원 간의 공동발의 지원횟수를 엣지 가중치로 설정해 가중 네트워크 자료로 변환하였다. 공동발의 네트워크는 공동발의자에서 대표발의자로 방향성을 가지지만 본 연구에서는 방향성을 고려하지 않은 가중 네트워크에서의 중첩 커뮤니티 탐지 알고리즘을 적용하기 위해, 두 노드 사이 엣지의 가중치를 합하는 방식으로 방향성을 제거하고 분석을 진행하였다.

연구에서는 17대 국회의원들의 정당을 열린우리당, 한나라당, 민주노동당과 그 외의 당은 기타로 분류하였다. 보다 나은 분석을 위해 공동발의 네트워크를 Chang (2011)이 제안한 것과 같이 정국의 변화에 따라 4개의 시기로 나누었다. 시기 1은 2004년 6월부터 2005년 1월까지로 열린

우리당이 4대 개혁입법을 추진하던 시기, 시기 2는 2005년 2월부터 2006년 1월까지로, 노무현 대통령과 열린우리당의 관계가 악화되었던 시기이다. 시기 3은 2006년 2월부터 2007년 1월까지로 신자유주의적 개혁의 시대이며, 시기 4는 2007년 2월부터 2008년 5월까지 17대 국회의 끝무렵이다.

3.2 분석결과

분석에는 공동발의 의원 수가 대표발의자를 포함해 10명 이상인 법안들만 이용하였다. 시기별 노드와 엣지 갯수는 표 1과 같다.

표 1: 시기별 노드와 엣지 갯수

	1st	2nd	3rd	4rd
노드	296	309	304	300
엣지	7280	11458	9143	7877

전체 시기에서의 그래프는 그림 2와 같으며, 각 시기에서의 그래프는 그림 3과 같다. 국회의원의 정당별로 노드의 색을 다르게 표현하였고, 엣지의 가중치는 엣지의 굵기를 통해 표현하였다. 노드의 크기는 노드의 degree와 비례하도록 설정하였다. 그림 2를 보았을 때, 전체 시기에서 네트워크가 정당에 따라 뚜렷하게 구분되는 경향을 볼 수 있다. 전체적으로 열린우리당과 한나라당의 군집이 발견되고, 민주노동당과 기타당이 중간에 위치한 모습을 보여준다. 그림 3의 각 시기에서도 이러한 경향은 유지된다.

모형 선택 기준으로는 Chen et al. (2010)에 의해 제시된 가중 네트워크에서의 중첩 커뮤니티 성능 측도인 modified modularity Q^w를 이용하였

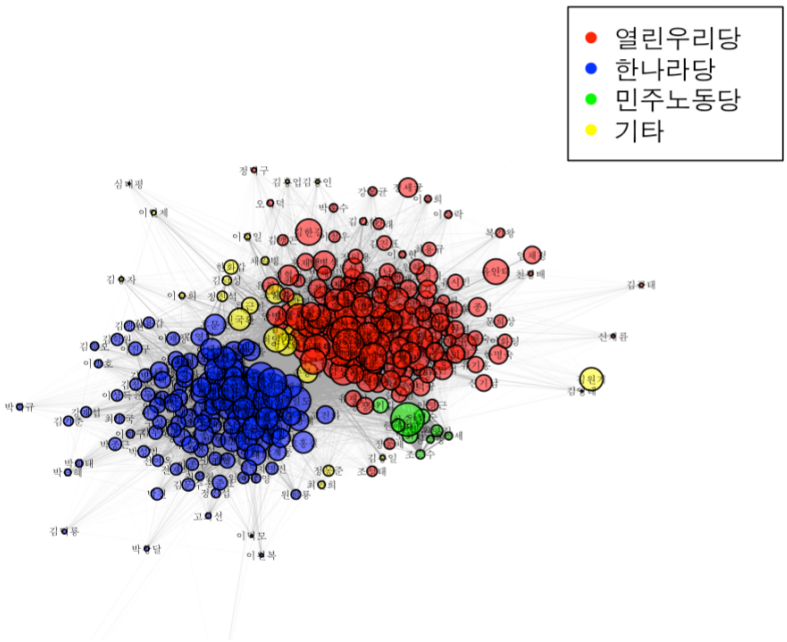


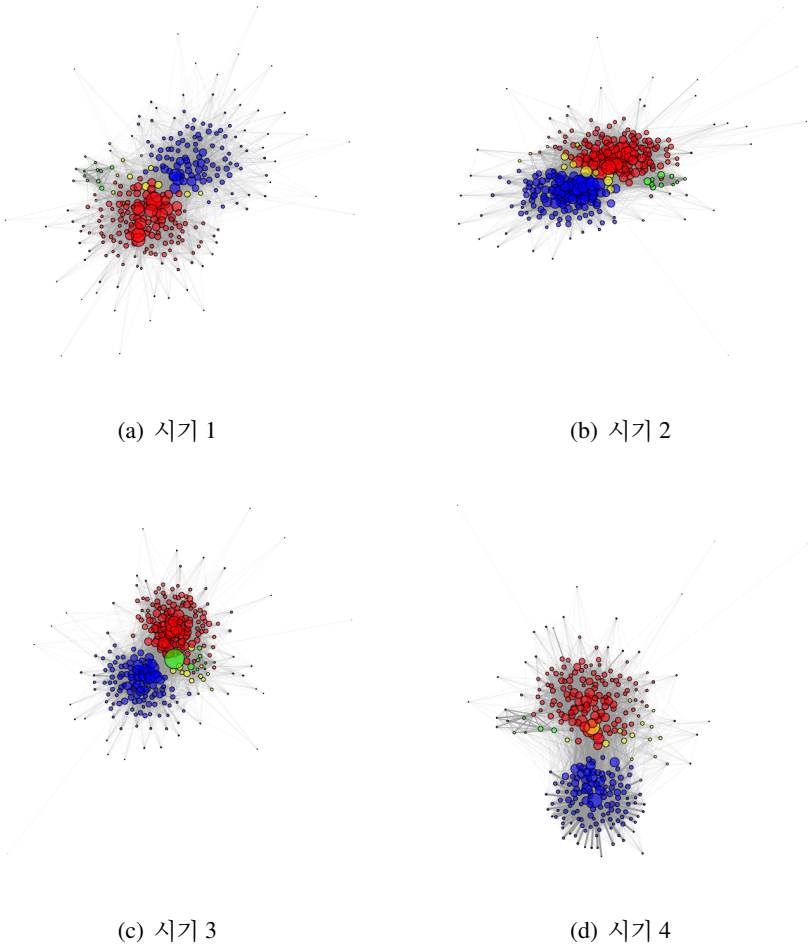
그림 2: 17대 국회 공동발의 네트워크

다. Q^w의 수식적 정의는 다음과 같다

$$Q^w(C) = \frac{1}{2m} \sum_{c_i \in C} \sum_{v,w \in V} b_{vi} b_{wi} \left(a_{vw} - \frac{k_v^w k_w^w}{2m} \right),$$

여기서 $k_v^w = \sum_{i \in V} w_{iv}$는 가중 네트워크에서의 노드 강도(node strength)로 노드 v에 연결되는 엣지 가중치의 합이다. b_{vi}는 노드 v가 커뮤니티 c_i에 속할 belonging factor이며, $b_{vi} = k_{vc_i}^c / \sum_{i'=1}^{K} k_{vc_{i'}}^c$로 정의된다. 이 때, $k_{vc_i}^c = \sum_{j \in c_i} w_{vj}$로 커뮤니티 c_i에 있는 노드들로 연결되는 가중치의 합이다.

하지만 모형의 커뮤니티 성능 측도가 높더라도 적은 수의 노드만을 커뮤니티로 분류하거나, 노드의 중첩을 많이 발견해내지 못하는 경우가 생길 수 있다. 도출된 커뮤니티의 이러한 성질들 또한 고려하기 위해 Q^w



(a) 시기 1 (b) 시기 2

(c) 시기 3 (d) 시기 4

그림 3: 시기별 17대 국회 공동발의 네트워크

이외에도 Ahn et al. (2010)이 제시한 커뮤니티 커버리지(community coverage) O^c와 중첩 커버리지(overlap coverage) O^{ov}를 함께 비교해보았다. O^c는 커뮤니티에 속하는 노드의 비율이고, O^{ov}는 커뮤니티에 포함된 노드들의 속한 커뮤니티 갯수의 평균이다. Ahn et al. (2010)의 기준을 따라 커뮤니티가 유의미한 최소한의 크기를 3으로 설정해 그 이상인 커뮤니티

만 고려하였다.

표 2: Modified modularity

Algorithms	1st	2nd	3rd	4rd
CPMw	0.08	-	0.090	0.033
Linkcomm	0.163	0.192	0.249	0.190
EAGLE	0.122	0.187	0.165	0.208

각 시기별 모형들의 modified modularity는 표 2와 같다. CPMw의 경우 각 시기마다 최적의 I^*와 k를 도출해 적합하였다. Link community는 single linkage 방법이 모든 시기에서 average linkage와 complete linkage에 비해 좋은 성능을 보였기 때문에 single linkage 방법을 이용하였다. 결과를 보았을 때, 시기 3에 Link community를 적합했을 때 modified modularity가 가장 높았다. CPMw가 비교적 낮은 성능을 보여주었고, Link community가 전체 시기에서 가장 높은 성능을 보여주었다. CPMw는 시기 2와 같이 모든 엣지 가중치에 대해서 n_1/n_2가 여과 전이 지점에 도달하지 못할 때, 커뮤니티가 생성되지 않는 문제점이 있었다.

표 3: 커뮤니티 갯수 K와 커버리지 O^c, O^{ov}

	CPMw			Linkcomm			EAGLE		
	K	O^c	O^{ov}	K	O^c	O^{ov}	K	O^c	O^{ov}
1st	32	0.917	2.567	15	0.986	1.798	17	0.934	2.401
2nd	-	-	-	53	0.994	3.905	28	0.978	3.250
3rd	74	0.928	4.691	96	0.970	6.712	24	0.956	4.790
4th	8	0.794	1.679	34	0.986	3.534	19	0.912	2.901

표 3의 커뮤니티 갯수와 커버리지를 보았을 때, 커뮤니티의 크기와 중첩 커버리지 O^{ov}는 시기 3의 Link community에서 가장 높았으며, 커뮤니티 커버리지 O^c는 시기 2의 Link community에서 가장 높았다. 모형

에 따라 중첩 커버리지의 큰 차이는 없었으나, 커뮤니티 커버리지는 Link community가 전반적으로 높았다. 시기 4에서 CPMw는 충분히 많은 커뮤니티가 생성되기 전에 n_1/n_2가 여과 전이 지점에 도달해버려 1/5에 해당하는 노드들이 커뮤니티에 속하지 못하는 문제점이 있었다.

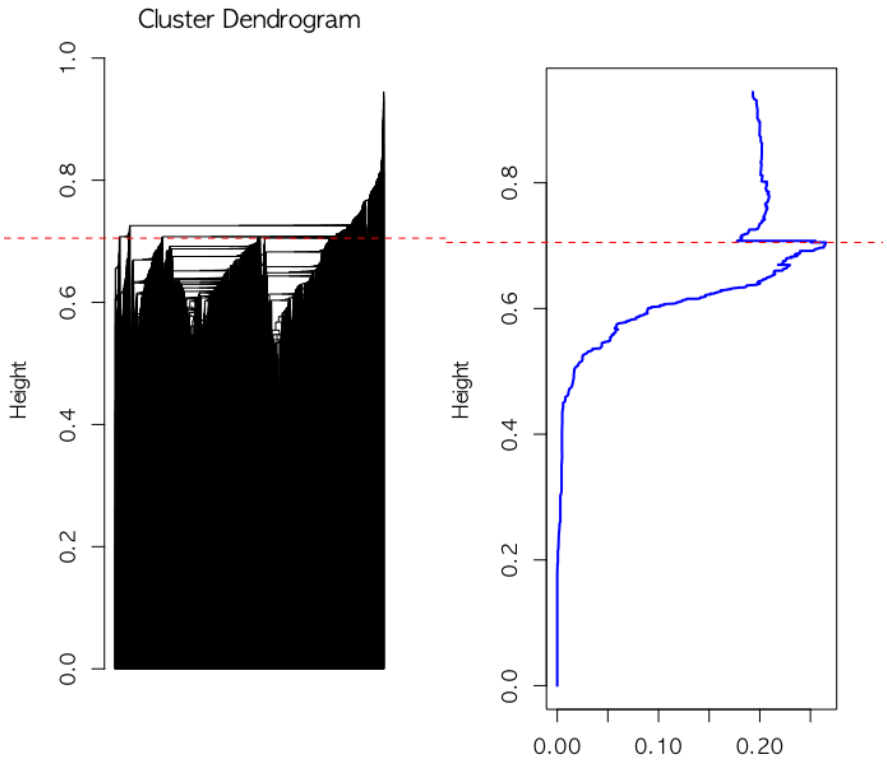
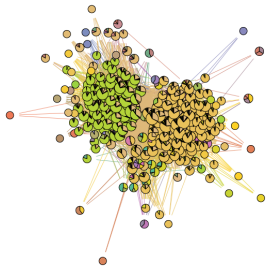
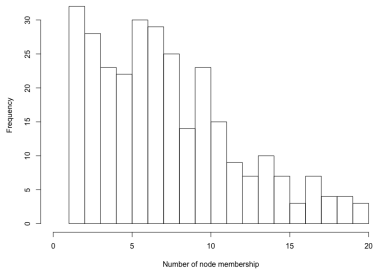


그림 4: 시기 3 Link community Dendrogram과 분할 밀도 plot

종합해보았을 때, 세 모형 중 Link community가 modified modularity와 커버리지에서 가장 높은 성능을 보여주었다. 그러나 도출된 커뮤니티가 높은 성능을 보여줌에도 불구하고, 그림 5의 시각화 결과를 보았을 때 지나치게 많은 수의 커뮤니티로 인해 해석의 어려움이 있었다. 과도한 커



(a) 적합 결과 네트워크 그래프



(b) 노드들이 속한 커뮤니티 갯수 히스토그램

그림 5: 시기 3 Link community 적합 결과

뮤니티의 갯수는 커뮤니티의 직관적인 이해에 어려움을 주기 때문에, 결과 해석을 위하여 커뮤니티의 갯수를 줄일 필요가 있다고 판단하였다. 크기가 큰 커뮤니티만 고려하는 방법도 있지만, 커뮤니티에 포함이 되지 않는 노드들이 생겨 커뮤니티 커버리지가 낮아질 수 있다는 단점이 있다. 따라서, 커뮤니티의 성능을 최대한 보존하면서 커뮤니티의 갯수를 줄이기 위해 커뮤니티의 군집인 메타 커뮤니티(Meta community)를 구하였다. 이를 위해, 커뮤니티간의 Jaccard 유사도를 기반으로 한 Ward linkage 계층적 군집화 모형을 적용하였고 커뮤니티간의 Jaccard 유사도는 다음과 같이 정의하였다,

$$S_J(c_i, c_j) = \frac{|c_i \cap c_j|}{|c_i \cup c_j|}.$$

메타 커뮤니티로부터 도출한 커뮤니티의 시각화 결과는 그림 6과 같았다. 커뮤니티의 갯수를 96개에서 3개로 통합시킴으로써 보다 직관적인 이해가 가능해졌다. 도출된 커뮤니티는 A, B, C로 정의하였다. 네트워크 그래프를 보았을 때, 커뮤니티 B와 C가 반대에 위치하고 커뮤니티 A는 뭉치는 경향이 B와 C보다 약하다. 커뮤니티 C가 B보다 좀 더 배타적인

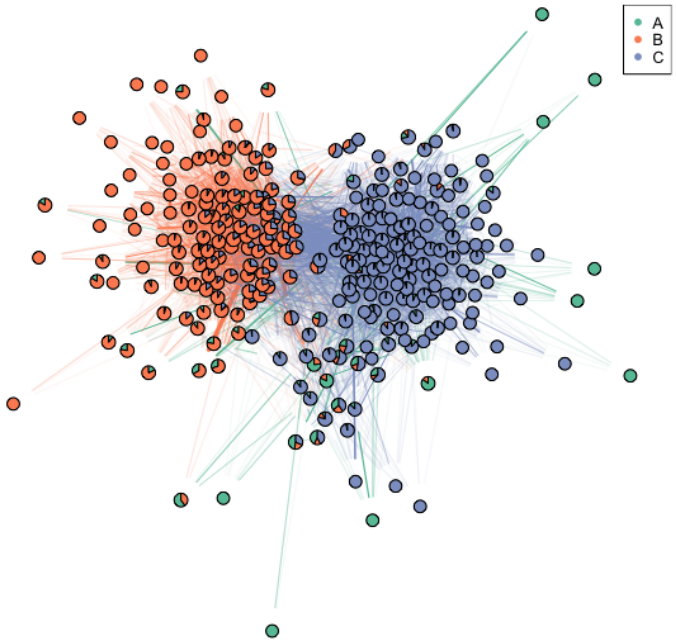


그림 6: 메타 커뮤니티 그래프

성향이 강하고 B의 많은 노드들이 C와 중첩되는 모습을 보인다.

그림 7의 히스토그램을 보면 커뮤니티별로 정당의 분포를 알 수가 있다. 커뮤니티 A와 C는 열린우리당이 대다수를 차지하고 대체로 비슷한 정당 구성을 가지고 있지만, A가 C에 비해 정당의 고른 분포를 보여준다. 이에 반해, 커뮤니티 B는 한나라당이 대다수를 차지한다. 이를 통해 커뮤니티 A와 C가 열린우리당을 중심으로 가까운 커뮤니티이고, B는 한나라당을 중심으로 두 커뮤니티와 대립하고 있음을 알 수 있다.

그림 8의 히스토그램은 정당별 커뮤니티의 분포를 나타낸다. 열린우

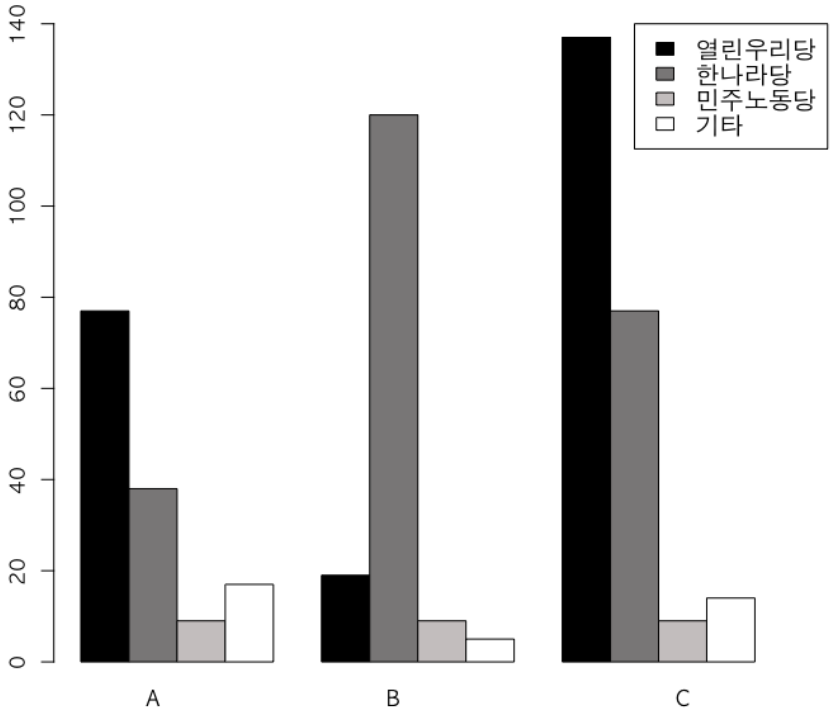


그림 7: 커뮤니티별 정당의 히스토그램

표 4: 정당별 노드가 속한 커뮤니티 갯수에 따른 노드의 비율.

단위 : %	커뮤니티 갯수				
	0	1	2	3	합
열린우리당	2.1	43.35	50.35	4.2	100
한나라당	3.85	20	60.77	15.38	100
민주노동당	0	0	0	100	100
기타	9.1	36.35	45.45	9.1	100

리당은 커뮤니티 C에 대다수가 속하며, 한나라당은 커뮤니티 B에 대다수가 속해있다. 민주노동당은 의원 전체가 커뮤니티 모두에 속하고, 기타

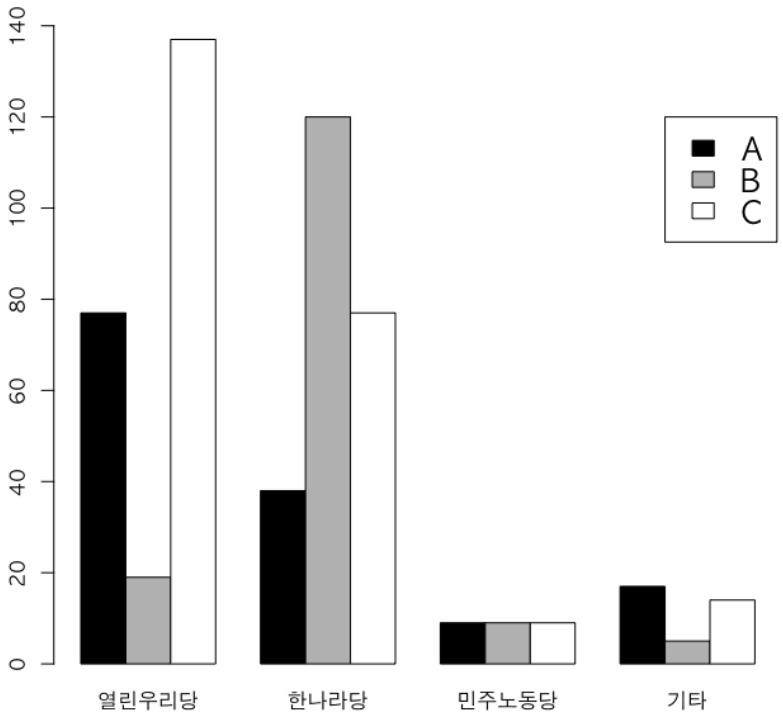


그림 8: 정당별 커뮤니티의 히스토그램

당원들은 커뮤니티 A와 C에 비슷하게 속하며, B에는 거의 없었다. 이로부터 모형 적합 전 그림 3에서 관측되었던 양상을 모형이 적절하게 나타내고 있음을 알 수 있다.

종합해보았을 때, 커뮤니티 A는 비록 열린우리당이 의원 자체가 많은 까닭에 열린우리당원의 비율이 높지만 비교적 모든 정당들이 조화된 커뮤니티로 볼 수 있다. 커뮤니티 B는 한나라당이 주도적인 커뮤니티로 볼 수 있으며, 다른 당들의 낮은 비율을 알 수 있다. 커뮤니티 C는 열린우리당이 주도하는 커뮤니티이지만, 한나라당도 다수 속해있다. 전반적

으로 한나라당과 열린우리당으로 대표되는 두 커뮤니티 B와 C가 대립적 구도를 보이며, 비교적 중도적인 커뮤니티 A가 C와 가까운 관계에 있다는 것을 알 수 있다.

한나라당은 커뮤니티 A와 C에 속하는 경우도 많은 데에 비해, 열린우리당은 커뮤니티 B에 거의 속하지 않는 모습을 보여준다. 또한, 표 4에서 하나의 커뮤니티에만 속하는 국회의원의 비율은 열린우리당이 43.35 퍼센트로 비율이 제일 높은걸로 보아 열린우리당 국회의원들이 다른 정당에 비해 배타적이었음을 알 수 있다. 민주노동당의 경우 소속 의원이 9명이지만 전원이 3개의 커뮤니티에 속하는 걸로 보아 사안에 따라 여야와 협력하는 정당으로 볼 수 있다. 기타당의 국회의원들은 한나라당보다는 열린우리당에 가까웠음을 유추할 수 있다.

제 4 장

Conclusion

본 논문에서는 가중 네트워크에서의 중첩 커뮤니티 탐지 모형인 CPMw, Link community, EAGLE을 17대 국회의원들의 공동발의 네트워크에 적용해보았다. 분석에 앞서 네트워크 자료형에서의 통계적 개념과 모형들의 이론적 내용에 대해 소개를 하였다. 17대 국회를 Chang (2011)이 제안한 4개의 시기로 나누었고 모형 적합을 위해 방향 네트워크를 노드 사이의 엣지 가중치를 합치는 방식으로 비방향 네트워크로 변환하였다. 가중 네트워크에서의 중첩 커뮤니티 성능 측도인 modified modularity를 이용해 4개 시기에 적합한 모형들의 성능을 비교하였으며, 커뮤니티 커버리지와 중첩 커버리지를 통해 중첩 커뮤니티가 많은 노드들을 포함하는지와 중첩을 많이 발견하는지를 살펴보았다.

분석 결과 모든 시기에서 Link community가 높은 modified modularity를 보여주었으며, 커뮤니티 커버리지와 중첩 커버리지 또한 대부분의 시기에서 제일 높았다. CPMw의 경우 여과 전이 지점에 따라 적합이 안되거나, 커뮤니티 커버리지가 낮은 불안정한 결과를 보여주었다.

그러나 모형의 성능이 좋음에도 불구하고, 지나치게 많은 커뮤니티 갯수로 인해 해석의 어려움이 있었다. 결과 해석을 위해 도출된 커뮤니티들의 Jaccard 유사도를 기반으로 계층적 군집화 방법을 이용해 3개의 메타 커뮤니티를 생성하였다. 2개의 커뮤니티는 각각 열린우리당과 한나라당이 주도적인 커뮤니티로 대립적인 구도를 보이며, 나머지 1개의 커뮤니티

는 열린우리당이 주도적인 커뮤니티와 가까우나 비교적 중도적인 모습을 보여주었다. 정당별로 보았을 때는 열린우리당은 한나라당이 주도적인 커뮤니티에는 거의 참여가 없었고 커뮤니티 갯수가 1명인 국회의원의 비율이 다른 정당에 비해 높은 걸로 보아 배타적인 성격을 가지고 있었음을 알 수 있다. 한나라당은 주도적인 커뮤니티 이외에 2개의 커뮤니티에서도 높은 비율을 차지해 열린우리당에 비해 좀더 균형잡힌 분포를 보여주었다. 민주노동당의 경우, 모든 의원들이 3개의 커뮤니티에 속해있는 걸로 보아 여야 구분없이 협력하는 정당이었으며, 기타당의 의원들은 열린우리당이 주도적인 커뮤니티에서 좀 더 비중이 높았으므로 열린우리당 의원들과 가까웠다고 볼 수 있다.

본 연구의 한계점으로는 방향 네트워크에서의 구현 가능한 중첩 커뮤니티 탐지 모형을 찾지 못해 방향 네트워크를 비방향 네트워크로 전환한 점과 최적 모형의 커뮤니티 갯수가 많을 때의 해석상의 어려움이 있었다. 그러나 커뮤니티들의 유사도를 기준으로 메타 커뮤니티를 생성해 커뮤니티 커버리지를 보존하면서 최적 모형으로부터 해석을 도출해낼 수 있었다. 메타 커뮤니티의 분석 결과는 기존 네트워크 그래프의 시각화 결과와 유사했으며, 커뮤니티와 정당별로 히스토그램을 이용해 각각의 특징을 분석할 수 있었다. 추후 연구에서 방향 네트워크에서의 중첩 커뮤니티 탐지 모형을 적용한다면 국회의원들의 공동발의 네트워크에 대하여 더욱 정확한 분석이 가능할 것이라 예상한다.

Reference

Ahn, Y., Bagrow, J. P. and Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, **466**(7307), 761-764.

Brutz, M. and François, G. M. (2015). A Modular Multiscale Approach to Overlapping Community Detection. *Arxiv*, **abs/1501.05623**

Chang, D. (2011). Policy Networks in the 17th National Assembly of Korea: An Analysis of Co-sponsorship and Friendship Ties. *Korean Political Science Review*, **10**(2), 157-187.

Chen, D., Shang, M., Zehua, L. and Fu, Yan. (2010). Detecting overlapping communities of weighted networks via a local algorithm. *Physica A: Statistical Mechanics and its Applications*, **389**(19), 4177-4187.

Evans, T. S. and Lambiotte, R. (2009). Line graphs, link partitions, and overlapping communities. *American Physical Society*, **80**(1).

Farkas, I., Ábel, D., Palla, G. and Vicsek, T. (2007). Weighted network modules. *New Journal of Physics*, **9**(6), 180.

Gregory, S. (2011). Finding overlapping communities in networks by label propagation. *New Journal of Physics*, **12**(10).

Lancichinetti, A., Fortunato, S. and Kertész, J. (2009). Detecting the Overlapping and Hierarchical Community Structure in Complex Networks. *New Journal of Physics*, **11**(3).

Palla, G., Derényi, I., Farkas, I. and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**(7043), 814-818.

Ren, W., Yan, G., Liao, X. and Xiao, L. (2009). Simple probabilistic algorithm for detecting community structure. *Physical Review E*, **79**(3).

Shen, H., Cheng, X., Cai, K. and Hu, M. (2009). Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications*, **388**(8).

Xie, J., Kelley, S. and Boleslaw K. S. (2011). Overlapping Community Detection in Networks: the State of the Art and Comparative Study. *ACM Computing Surveys*, **45**(4).

Xie, J., Boleslaw K. S. and Liu, X. (2011). SLPA: Uncovering Overlapping Communities in Social Networks via a Speaker-Listener Interaction Dynamic Process. *2011 IEEE 11th International Conference on Data Mining Workshops*, 344-349.

Xing, Y., Fanrong, M., Yong, Z. and Ranran, Z. (2015). Overlapping Community Detection by Local Community Expansion. *Journal of Information Science and Engineering*, **31**, 1213-1232.

Abstract

Seonghong Kim
Department of Statistics
The Graduate School
Seoul National University

In this paper, we detect the overlapping communities in the 17th National Assembly of Republic of Korea. The theoretical backgrounds of CPMw, Link community, and EAGLE, which are overlapping community detection models in weighted networks are presented, and the model is fitted to the cosponsorship network of the 17th National Assembly. The performances of the models are compared using the modified modularity, community coverage and overlapping coverage. Since cosponsorship networks represent political relationships among the legislators, the political networks can be found from statistical communities. Therefore, we compare the communities derived from the optimal models with the political parties of legislators and examines characteristics of the parties in the 17th National Assembly.

Keywords : Network, Overlapping community, National Assembly of Republic of Korea

Student Number : 2018-24939