

## 공저자 네트워크를 활용한 응용통계연구 분석

이민희<sup>1</sup> · 박미라<sup>2</sup> · 이효정<sup>3</sup> · 진서훈<sup>4</sup>

<sup>1</sup>고려대학교 통계학과, <sup>2</sup>을지대학교 예방의학교실, <sup>3</sup>고려대학교 통계학과

<sup>4</sup>고려대학교 정보통계학과

(2011년 10월 접수, 2011년 11월 채택)

---

### 요약

본 연구는 사회 네트워크 분석을 활용하여 응용통계연구에 논문을 게재한 저자 간 공동 연구의 양상을 분석하였다. 2000년부터 2010년까지 총 664개의 논문을 대상으로 하였으며 네트워크 분석을 통해 공저자 네트워크의 중심을 구성하는 연구자를 찾아보았고 하위 네트워크 분석을 통해 연구 분야 및 공동연구 집단이 차별화되는 네트워크를 살펴 보았다. 또한, 공저관계를 반응변수로 하고 소속집단을 설명변수로 하는 로지스틱 회귀분석을 수행하여 소속집단이 공동연구에 미치는 영향정도를 분석해 보았다.

주요용어: 사회네트워크분석, 공저자네트워크, 중심성, 응용통계연구.

---

### 1. 서론

최근 여러 분야에서 사회 네트워크(social network)에 대한 관심이 높아지고 있다. 사회 네트워크란 다수의 연결 또는 연결되지 않은 개인(또는 기관)으로 이루어진 사회적 구조를 이르는데, 이 때 연결 여부는 친구관계, 성관계, 금융거래, 통신 빈도, 직무적 접촉, 정보 제공 등 관심 주제에 따라 다양하게 정의될 수 있다. 사회 네트워크 분석(social network analysis; SNA)은 이들 간의 구조를 파악하기 위한 방법으로서 다수의 점(點, point)과 이들을 연결하는 선(線, line)으로 구성된 망(網, network)에 대한 사회과학적 통계적 분석이다 (허명희, 2010). 사회 네트워크 자료는 크게 두 가지로 나눌 수 있다. 일반적인 자료 형태는 노드라 불리는  $n$ 개의 점간의 관계를 정방행렬로 나타낼 수 있는 네트워크 자료이다.  $n \times n$  행렬에 칸  $(i, j)$ 는  $i$ 와  $j$ 간의 연결여부 혹은 연결강도를 표현한다. 예를 들어 학교 내 학생 간의 교우관계 네트워크, 대기업의 자회사들 간의 자본소유관계 등을 표현하는 것이다. 두 번째는 사회 구성원과 조직(구성체) 간 관계를 나타낼 수 있는 2부 네트워크(bipartite network)이다. 일명 연결네트워크(two-mode network)라고 불리는 2부 네트워크는  $n$ 개 행 노드와  $m$ 개 열 노드 간 연결로 구성되어  $n \times m$  행렬로 나타낼 수 있다. 예컨대  $n$ 명 개인에 대하여 중복가입을 허용하여 그들이 가입한  $m$ 개 클럽을 조사한 클럽 멤버십 자료로부터 2부 네트워크를 생성할 수 있다. 개인 2명이 어느 한 클럽에 가입되어 있다면 소통 관계인 것으로 보고 2개 클럽이 한 개인을 회원으로 공유하면 연결된 것으로 본다. 이와 같은 식으로 개인 간, 그리고 클럽 간 네트워크가 형성될 수 있다. 공저자 네트워크 분석은 이러한 사회 네트워크 분석의 한 응용 사례로서 논문의 저자를 각각 점으로 놓고 이들이 연결되어 있는 상호구조를 파악하기 위한 분석이다. 공저자 네트워크 분석에서 노드는 저자 개인이 되고 연결선은 함께 연구했

이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No.2011-0004376).

<sup>4</sup>교신저자: (339-7007) 충남 연기군 조치원읍, 고려대학교 정보통계학과, 교수. E-mail: seohoon@korea.ac.kr

을 경우에 생기게 된다. 예컨대 A, B, C라는 세 사람이 공동연구를 한 경우 A-B, A-C, B-C의 관계가 성립한다 (정민수와 정동준 2007). 이러한 분석을 통해 저자 간의 관계와 특성을 분석할 수 있다. 관련된 몇몇 선행연구를 살펴보면 Newman (2001)은 물리학, 의학, 전산학 등의 논문에 있어 공저자 네트워크 분석을 연구하였고 Malin과 Carley (2007)는 생물정보학과 의료정보학 간의 관련성을 사회 네트워크 분석을 통해 연구하였다.

응용통계연구는 한국통계학회에서 발간되는 논문집으로 1987년 창간되었다. 2003년까지 연 2회 발간되던 것이 2004년부터 연 3회로 늘었으며 2008년부터는 격월로 매년 6회 발간되고 있다. 최초로 국문 논문집이었던 것이 현재는 국문과 영문 논문이 모두 출판되고 있다. 최근에는 한 호에 약 20편의 논문이 게재되고 있어 연간 100편 이상의 논문이 응용통계연구를 통해 발표되고 있다. 국내의 통계학 학술 지로는 역사가 오래되었으며 발표 논문의 규모도 상위에 있는 논문집이라 할 수 있다. 이에 통계학 연구자들 간의 공저 네트워크 분석에 적합하다고 판단되어 사회 네트워크 분석을 활용한 연구자 간의 공저 관련성을 분석하였다.

2절에서는 공저자 네트워크 분석 방법과 네트워크 분석에 이용되는 측도를 간략히 소개하였다. 3절에서는 본 연구에 이용한 응용통계연구 게재 논문의 공저자 데이터에 대해 소개하고 2000년부터 2010년까지 응용통계연구에 게재된 논문을 대상으로 한 분석결과를 제시하였다. 공저자 네트워크에서 발견된 하위 네트워크(sub-network)에 대해 상세 분석하였으며 아울러 2000년과 2010년의 공저자 네트워크를 비교함으로써 두 시점 간 공동연구 경향의 변화를 살펴보았다. 끝으로 4절에 연구의 결론을 요약하였다.

## 2. 공저자 네트워크 분석 방법론

공저자 네트워크 분석 자료는 저자와 논문 간 관계를 나타낼 수 있는 2부 네트워크의 형태이다. 2부 네트워크는  $n$ 개의 행 노드와  $m$ 개의 열 노드 간 연결로 구성되어  $n \times m$  행렬로 나타낼 수 있다. 예를 들어 2개의 논문이 있을 때 그 중 하나는 저자 A와 B가 썼고 나머지 하나는 저자 A, C, D가 함께 썼다면 다음과 같은 행렬로 표현할 수 있다.

	A	B	C	D
논문1	1	1	0	0
논문2	1	0	1	1

이 때 논문에 대한 저자별 가중치를 달리 줄 수도 있다. 예를 들어 공저 논문에 있어 주저자는 1, 공동저자는 0.5의 가중치를 부여하여 자료를 생성할 수 있다. 앞의 예에서 논문1과 논문2의 주저자가 각각 A와 C였다면 가중치가 반영된 자료행렬은 아래와 같이 표현된다.

	A	B	C	D
논문1	1	0.5	0	0
논문2	0.5	0	1	0.5

네트워크 분석의 주요 요소 중 하나는 네트워크 데이터를 2차원 또는 3차원 공간에 시각화하는 것이다. 시각화 알고리즘으로는 force-directed 알고리즘 (Fruchterman와 Reingold, 1991), Kamada-Kawai 알고리즘 (Kamada와 Kawai, 1989), MDS(multi-dimensional scaling) 등 여러가지가 있다. 표 2.1은 R의 sna 라이브러리에서 제공되는 시각화 알고리즘 옵션들이다(R package 'sna', 2011). 본 연구에서는 이 중 circle 옵션을 이용하여 네트워크의 시각화를 시도하였다.

표 2.1. R의 네트워크 시각화 알고리즘 옵션들

option	description
fruchtermanreingold	A variant of Fruchterman and Reingold's force-directed layout algorithm (default).
kamadakawai	A version of the Kamada-Kawai force-directed layout algorithm.
spring	A spring embedder algorithm.
circle	Nodes are placed in a circle, arranged clockwise by their order in the adjacency matrix.
eigen	Node placement is based on the eigenstructure of the adjacency matrix.
hall	Node placement based on the last two eigenvectors of the Laplacian of the input matrix.
mds	Node placement based on multidimensional scaling of a specified distance matrix(matrix of rows and columns used by default).
princoord	Node placement based on the eigenstructure of a given correlation/covariance matrix(matrix of rows and columns used by default).
target	Arranges nodes on the radii of concentric circles, based on a vector of node values(affine-transformed Freeman degree centrality scores used by default).
random	Nodes are placed randomly. Uses a uniform distribution by default, but a Gaussian or "Gaussian donut" distribution may be specified.

네트워크의 전역적 특성을 나타내는 지표인 네트워크 밀도(density)는 '가능한 총 연결선의 수' 대비 '총 연결선의 수'로 정의된다. 네트워크 내 노드 수를  $n$ 이라고 할 때 '가능한 총 연결선의 수'는  $n \times (n - 1)$ 이다. 그리고 총 연결선 수는 입선수(indegree) 또는 출선수(outdegree)의 합과 같다 (허명희, 2010). 즉, 식 (2.1)과 같이 두 값의 비로 네트워크 밀도를 구하게 되는데, 여기서  $g_k^{[I]}$ 는 노드  $k$ 로의 입선수이고  $g_k^{[O]}$ 는 노드  $k$ 로 부터의 출선수이다. 이러한 밀도는 네트워크 내 전체 구성원이 서로 간 얼마나 많은 관계를 맺고 있는가를 표현하기 위한 개념이다 (손동원, 2002).

$$\text{density} = \frac{\sum_{k=1}^n g_k^{[I]}}{n(n-1)} = \frac{\sum_{k=1}^n g_k^{[O]}}{n(n-1)}. \quad (2.1)$$

네트워크의 중심성 지수는 그 행위자가 전체 네트워크에서 중심에 위치하는 정도를 나타낸다. 이러한 중심성 지수는 중심에 대한 정의에 따라 그 계산방법을 달리할 수 있다. 본 연구에서는 근접 중심성과 중개 중심성이라는 두 가지 다른 관점에서 중심성 수치를 살펴보았다. 첫 번째로 근접 중심성은 특정 노드가 그를 제외한 다른 노드와 얼마나 가까운가를 평가하는 것이다. 도달 가능한 모든 노드의 평균거리를 이용하는데, 고립 노드가 있을 경우를 감안하여 노드 간 최단거리의 역수(이는 노드 간 친밀도 또는 유관도라 볼 수 있음)를 평균(조화평균)하는 것으로 정의할 수 있다. 예를 들어, 그림 2.1의 A, B, C노드로 이루어진 왼쪽의 네트워크에서 노드 A의 중심성은 A와 B의 최단거리가  $d_1$ 이고 A와 C의 최단거리가  $d_1 + d_2$ 이므로,  $1/2(1/d_1 + 1/(d_1 + d_2))$ 로 구해진다. 근접 중심성을 일반화 하여 표현하면,  $d_{ki}$ 를 노드  $k$ 와  $i$ 의 거리라고 할 때 노드  $k$ 의 근접 중심성은  $C_k^{[1]} = 1/(n-1)(\sum_{i \neq k} 1/d_{ki})$ 로 표현할 수 있

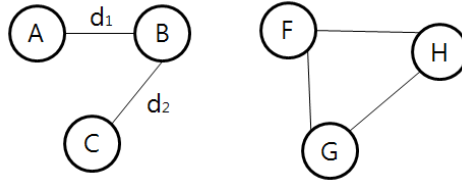


그림 2.1. 네트워크의 예

표 3.1. 2000년에서 2010년까지 논문에 나타난 상위 20개 키워드

순위	키워드	빈도	순위	키워드	빈도
1	로지스틱회귀	16	11	회귀의사결정나무	9
2	데이터마이닝	15	11	의사결정나무	9
2	군집분석	15	11	결측치	9
4	EM알고리즘	12	14	혼합모형	8
5	반복측정	11	14	이상점	8
5	깁스샘플러	11	14	신뢰구간	8
6	K-평균군집	10	14	로그선형모형	8
6	붓스트랩	10	18	포함확률	7
6	몬테칼로방법	10	18	정확도	7
6	마이크로어레이	10	18	주성분분석	7

다. 두 번째로 중개 중심성은 네트워크 내에서 한 행위자가 다른 행위자들 사이에서 매개자 혹은 중개자의 역할을 하는 정도를 측정하는 것이다. 특정 노드  $i$ 의 중개 중심성은 노드  $i$ 를 제외한 다른 두 노드 사이의 최단 경로들 중 노드  $i$ 를 지나는 경로의 비율을 모든 노드 쌍들에 대해 구하여 더한 값이다. 즉,  $g_{jk}$ 를 노드  $j$ 에서  $k$ 로 가는 최단경로의 수이고  $g_{jik}$ 를 노드  $j$ 에서  $k$ 로 가는 최단경로 중 노드  $i$ 를 거쳐 가는 최단경로의 수라고 두면 노드  $i$ 의 중개 중심성은  $C_i^{[2]} = \sum_{j \neq i} \sum_{k \neq i, \neq j} g_{jik} / g_{jk}$ 로 구하게 된다.

네트워크 분석에서 노드들이 서로 연결되어 있는가 여부를 분석하는 것은 중요하다. 네트워크의 모든 노드가 서로 연결되어 있을 수도 있으나 경우에 따라 일부 노드들은 연결되지 않고 별도의 그룹을 형성할 수 있다. 네트워크 분석에서 서로 연결된 노드들의 묶음을 컴포넌트라고 한다. 즉, 컴포넌트란 한 네트워크에 소속된 노드들끼리 하나의 연결체계를 형성하는 하위 네트워크(sub-network)를 말한다(손동원, 2002). 예를 들어 그림 2.1은 A, B, C와 F, G, H의 두 개의 컴포넌트로 이루어진 네트워크이다. 컴포넌트의 분석을 통해 전체 네트워크가 몇 개의 하위 네트워크로 이루어져 있는지를 파악할 수 있고 이를 통해 네트워크 내 정보의 흐름이 원활한지 또는 그렇지 않아 하위집단 간 파편화 가능성이 있는지 여부 등을 분석할 수 있다.

### 3. 응용통계연구분석

#### 3.1. 자료 설명

본 연구에 이용한 자료는 2000년부터 2010년까지의 응용통계연구에 실린 논문이다. 논문수는 664개이고 저자수는 747명이었으며 저자들의 소속기관 수는 319개였다. 총 1,989개의 키워드가 등장했으며 표 3.1은 빈도 기준으로 상위 20개의 키워드를 정리한 것이다. 표 3.2에는 각 연도별 논문수와 저자수, 그리고 소속기관의 현황을 정리하였다.

전체 논문 자료로부터 논문(664)×저자(747), 소속(370)×저자(747)의 2부 네트워크 구조의 데이터를

표 3.2. 응용통계연구 연도별 논문게재 현황

년도	논문수	연간호 수	저자수	소속기관 수
2000	47	2	86	44
2001	39	2	82	41
2002	39	2	63	41
2003	35	2	70	37
2004	48	3	91	41
2005	54	3	99	45
2006	48	3	91	41
2007	50	3	98	43
2008	93	6	177	64
2009	109	6	209	72
2010	102	6	187	74
2000~2010	664	38	747	370

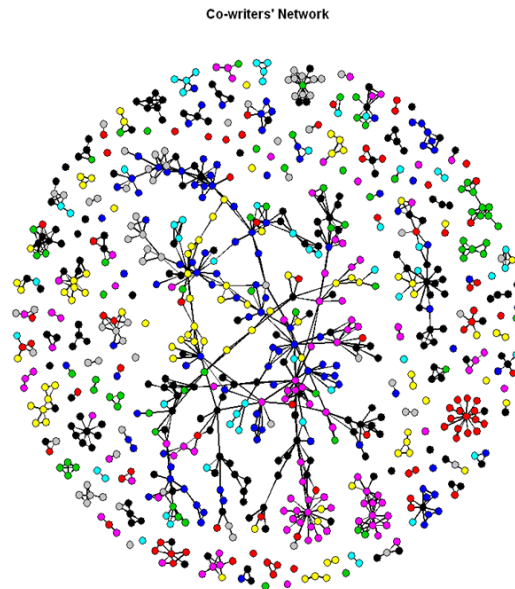


그림 3.1. 공저자 네트워크(2000년~2010년)

구성하였다. 이 중 네트워크 분석에는 공저자 데이터인 논문×저자를 이용하였다. 논문×저자 데이터 행렬의  $(i, j)$  원소는  $i$ 논문에 저자  $j$ 가 주저자로 참여하면 1, 공동저자로 참여하면 0.5의 값을 갖는다. 여기서 주저자는 제1저자 또는 교신저자를 의미하는 것으로 제1저자와 교신저자가 다를 경우 한 논문에 2명의 주저자가 존재하게 된다. 이 데이터 행렬을  $A$ 로 두고  $A'A = C$ 의 연산을 행하면  $747 \times 747$ 의 저자 간 네트워크 데이터 행렬  $C$ 를 얻을 수 있고 이 데이터를 이용해 공저자 네트워크 분석을 실시하였다.

### 3.2. 전체 네트워크 분석

2000년에서 2010년까지 총 11년 동안의 응용통계연구에 실린 논문에 대한 공저자 네트워크 분석을 통해 얻어진 네트워크가 그림 3.1과 같다. 색깔은 해당 네트워크 내에 같은 소속 저자를 구분한 것으로 연

표 3.3. 하위 네트워크 현황

크기	1	2	3	4	5	6	7	8
수	55	30	18	15	10	3	4	3
비율(%)	36.91	20.13	12.08	10.07	6.71	2.01	2.68	2.01
크기	9	10	11	13	16	30	288	
수	2	3	1	1	2	1	1	
비율(%)	1.34	2.01	0.67	0.67	1.34	0.67	0.67	

표 3.4. 2000년~2010년 공저자 네트워크에서 주요 저자

순위	연결가중치 기준		연결선 기준		근접 중심성 기준	
	이름	연결가중치합	이름	연결선 수	이름	근접 중심성지수
1	황선영	16.75	황선영	22	송석현	0.08
2	송혜향	16	홍중선	21	정병철	0.08
3	홍중선	14.75	송석현	15	황선영	0.08
4	임동훈	14	송혜향	15	박진아	0.07
5	신기일	12	허명희	13	이동희	0.07
6	허명희	12	김진흠	12	최문선	0.07
7	송석현	11.75	김혜중	12	백지선	0.07
8	박유성	10.75	신기일	12	이재원	0.07
9	여인권	10.5	임동훈	12	전명식	0.07
10	정병철	9.25	김기환	11	구자용	0.06

결된 네트워크가 같은 색깔로 구성되면 같은 소속의 연구자들이 공동연구 성향이 높음을 보여주는 것이다. 전체 네트워크 특성치의 하나인 밀도는 0.00215로 구해졌다. 또한 이 네트워크를 구성하고 있는 하위 네트워크의 현황은 표 3.3과 같다. 747명의 저자 중 288명이 하나의 하위 네트워크를 이루고 있고 나머지 저자들은 30명 이하 규모의 소그룹의 이루고 있다. 특히 4명 이하의 소그룹에 다수가 포함되어 있어 많은 연구자들이 혼자 연구를 하거나 공동연구를 하더라도 몇몇의 특정 연구자와만 공동연구를 수행하고 있음을 알 수 있다. 한편 이 네트워크에서 관심을 끄는 것은 중심부에 있는 288명으로 구성된 하위 네트워크이다. 이는 이후 별도의 상세분석을 통해 그 내용을 자세히 살펴보기로 한다.

공저자 네트워크의 각 노드별 연결선 수와 연결 가중치를 구해 각각의 상위 10개 노드에 해당하는 저자를 표 3.4에 정리하였다. 또한 노드 별 근접 중심성을 구하여 상위 10개 노드를 정리하였다. 연결선의 수는 해당 저자에 있어 공동연구자의 규모를 나타내며 연결 가중치의 합은 논문의 개수와 논문에서의 역할이 동시에 반영되어 있는 수치이다. 연결 가중치 합에 비해 연결선의 수가 상대적으로 많다면 공동연구를 많이 수행하며 주저자 보다는 공동저자로 참여하는 비율이 높다고 볼 수 있다. 반대로 연결선의 수에 비해 연결 가중치의 합이 크면 소수의 제한된 연구자와 공동연구를 진행하며 주저자로 논문을 작성하는 비율이 높다고 볼 수 있다. 예를 들어 “박유성”은 연결 가중치 순위에 포함되어 있지만 연결선 수 순위에는 빠져있는데 연결선 수는 10으로 17위에 해당되었다. 또한, 총 10편의 논문을 게재하였는데 이중 7편의 논문에 주저자로 참여하여 주로 주저자의 역할을 하고 있다.

그림 3.1의 공저자 네트워크는 총 149개의 하위 네트워크로 구성되어 있다. 이렇게 하위 네트워크가 많은 경우는 네트워크의 중심성 분석이 크게 의미를 갖지 못하는데 그 이유는 중심성 상위 노드가 가장 규모가 큰 하위 네트워크를 구성하는 노드에서 중심에 위치하는 노드로 결정되어지기 때문이다. 실제로 표 3.4의 근접 중심성을 보면 “송혜향”의 경우 연결 가중치나 연결선 수에서 상위에 있음에도 불구하고 근접 중심성은 상위에서 빠져 있다. 이는 이 저자의 경우 가장 큰 하위 네트워크에 속해있지 않을 것이

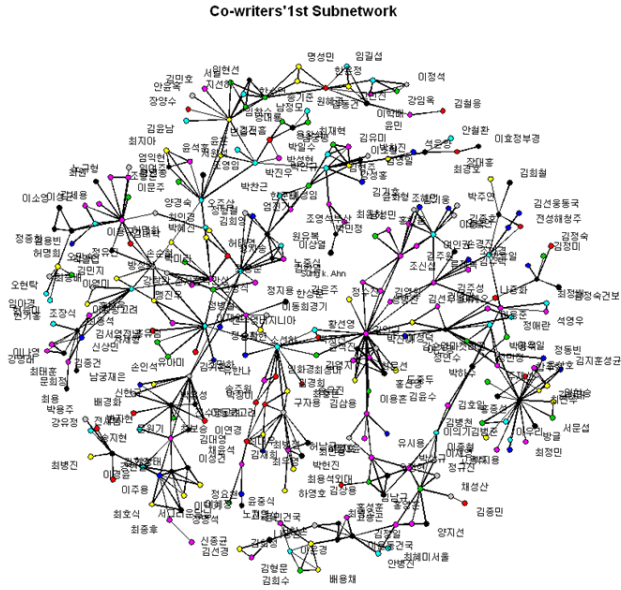


그림 3.2. 제 1 하위 네트워크(288명)

표 3.5. 288명 하위 네트워크에서 주요 저자

순위	연결가중치 기준		연결선 수 기준		근접 중심성 기준		중개 중심성 기준	
	이름	연결가중치 합	이름	연결선 수	이름	근접중심성 지수	이름	중개 중심성 지수
1	황선영	16.75	황선영	22	송석헌	0.21	송석헌	5.7
2	홍중선	14.75	홍중선	21	정병철	0.21	황선영	4.67
3	허명희	12	송석헌	15	황선영	0.21	이재원	3.3
4	송석헌	11.75	허명희	13	박진아	0.19	전명식	3.29
5	박유성	10.75	김진흠	12	이동희	0.19	정형철	2.42
6	여인권	10.5	김기환	11	최문선	0.19	박만식	2.25
7	정병철	9.25	박만식	11	백지선	0.18	진서훈	2.21
8	한상태	8.5	이성덕	11	이재원	0.18	윤석진	1.95
9	김기환	8.25	이재원	11	전명식	0.18	김영지	1.93
10	박만식	8.25	전명식	11	구자용	0.17	이성덕	1.67

라는 것을 짐작케 해 준다. 따라서 중심성 분석은 가장 큰 하위 네트워크인 288명의 저자로 구성된 네트워크에서 수행하였다. 실제로 이후 분석에서 하위 네트워크에서의 근접 중심성 상위 10인이 전체네트워크에서의 근접 중심성 상위 10인과 일치함을 확인할 수 있다.

**3.3. 하위 네트워크 분석**

2000년부터 2010년까지 논문의 저자로 참여한 747명 중 288명이 하나의 하위 네트워크를 이루고 있다. 이 288명으로 이루어진 네트워크를 따로 그린 것이 그림 3.2이다. 또한 이 네트워크에서 연결선 수와 연결가중치 합이 상위 10위에 속하는 저자들을 정리한 것이 표 3.5이다. 전체 네트워크에서는 연결선 수 및 연결가중치 합이 상위에 속했으나 288명 하위 네트워크에서는 빠져있는 저자는 공동연구를 행

표 3.6. 제 1 하위 네트워크의 로지스틱 회귀 결과

Coefficients	Estimate	Exp(b)	p-value
(intercept)	-26.566	2.90E-12	< 0.001
(동일기관 소속 여부)	53.132	1.19E+23	< 0.001

Co-writers'2nd Subnetwork

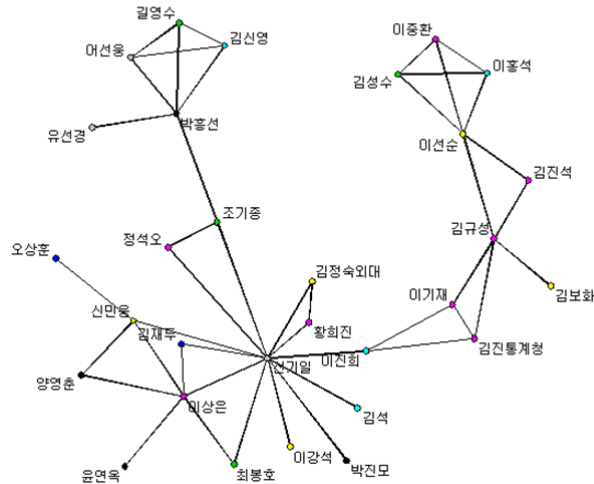


그림 3.3. 제 2 하위 네트워크(30명)

하는 연구자 그룹이 다른 연구자들과 차별되는 특정 학문 영역의 연구자일 것으로 생각된다. 또한 표 3.5에는 이 네트워크의 노드들 중 근접 중심성과 중개 중심성 상위에 속하는 노드들도 정리되어 있다. 다른 노드들로부터 가까운 정도를 나타내는 근접 중심성을 보면 “송석현”, “정병철”, “황선영”이 네트워크의 가장 중심에 있다고 볼 수 있고 노드와 노드를 연결하는 중개 중심성의 관점에서 보면 “송석현”, “황선영”, “이재원”의 순으로 네트워크의 중심에 있다고 볼 수 있다.

네트워크 분석을 함에 있어 자연스러운 접근은 노드 간의 연결을 설명할 수 있는 다른 변수를 확보하여 이들 변수와 연결 간의 연관모형을 도출함으로써 노드 간의 연결이 무엇에 영향 받고 있는지를 설명하는 것이다. 네트워크의 로지스틱 회귀분석은 이와 같은 관계를 찾아보는데 이용할 수 있는 분석방법으로 주요 관계의 유무를 종속변수로 그 외 부속 관계 혹은 인구학적 특성을 설명변수로 하여 분석하는 방법이다. 본 연구에서는 공저자 간 네트워크에서 가중치를 무시하고 공저자 관계 여부만을 종속변수로 하여 로지스틱 회귀분석을 수행하였다. 독립변수로는 각 저자의 소속기관을 이용하였는데, 즉, 두 저자 사이에 공동 논문의 존재 여부를 종속변수로 하고 두 저자가 같은 소속인지 여부를 독립변수로 한 로지스틱 회귀모형을 분석한 것이다. 표 3.6은 288명 하위 네트워크에서의 로지스틱 회귀모형의 결과이다. 모형의 결과로부터 동일 기관에 소속되어 있을수록 공동연구를 수행하는 경향이 매우 유의하다는 것을 알 수 있다.

두 번째로 큰 하위 네트워크는 30명의 저자로 이루어진 네트워크로 그림 3.3과 같다. 연결 가중치, 연결 선 수, 근접 중심성, 중개 중심성 모두 “신기일”이 최상위에 있다(연결 가중치 = 12, 연결선 수 = 12, 근접 중심성 = 0.54, 중개 중심성 = 63.6). 이 네트워크를 구성하는 30명의 소속을 보면 한국외대 13명,



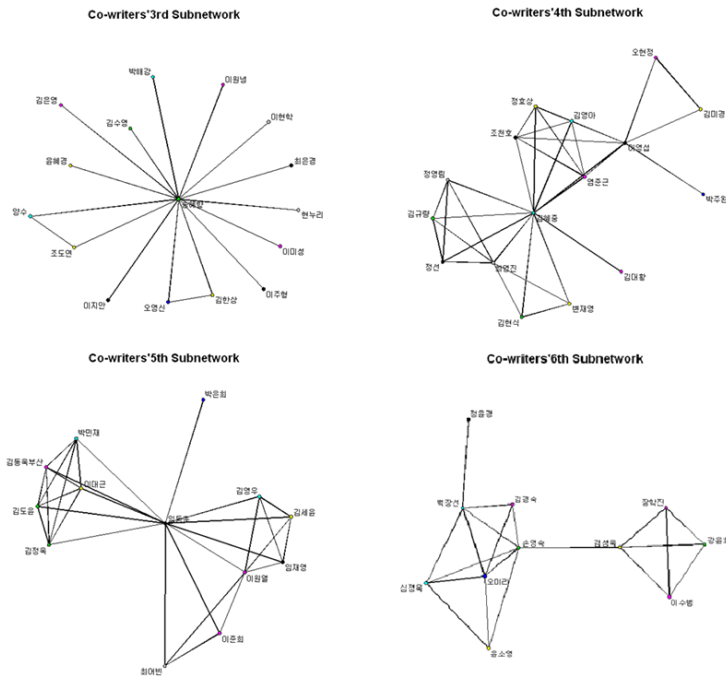


그림 3.4. 제 3, 4, 5, 6 하위 네트워크

통계청 3명, 경기대 3명, 서울대 3명, 고려대, 성균관대, 충북대, (주)니즈아이가 각 1명으로 되어 있다. 따라서, 한국외대의 연구자들을 중심으로 이루어진 네트워크로 볼 수 있다.

그림 3.4는 제 3, 4, 5, 6 하위 네트워크 그림이다. 제 3 하위 네트워크는 16명으로 구성되어 있는데 이 네트워크의 저자들은 전원 카톨릭대학교 의학통계학과 및 간호대학에 소속되어 있다. 즉, “송혜향”을 중심으로 한 카톨릭대학의 교내 연구 집단으로 특성을 분류할 수 있다. “송혜향”의 근접 중심성과 중개 중심성은 각각 1.07과 2.06으로 계산되었다. 제 4 하위 네트워크는 16명의 저자로 구성된 하위 네트워크이다. 이 네트워크를 구성하고 있는 저자들의 소속분포를 보면 동국대 통계학과 소속이 10명, 기상청 소속이 6명으로 되어 있다. 따라서 동국대 통계학과와 기상청의 공동연구 네트워크로 해석할 수 있다. 한편, 네트워크의 중심성을 구해보면 “김혜중”이 네트워크의 중심에 있는데 근접 중심성과 중개 중심성은 각각 0.48과 12.4로 구해졌다. 제 5 하위 네트워크는 13명으로 구성된 하위 네트워크이다. 이 네트워크는 “임동훈(경상대)”을 중심으로 하고 있는데(근접 중심성 = 1.17, 중개 중심성 = 8.2) “임동훈”과 연결되어 있는 12명의 공저자의 소속을 살펴보면 한국과학영재학교 소속이 9명, 경상대와 부산대가 각 1명이다. 즉, 이 네트워크는 “임동훈”을 중심으로 한 과학영재학교 연구자들과의 공동연구 네트워크라고 해석할 수 있겠다. 제 6 하위 네트워크는 11명으로 구성된 하위 네트워크이다. 이 네트워크를 구성하는 저자의 소속을 보면 전남대 통계학과 6명, 한양대 응용수학과 3명, 서울대와 경기통계사무소가 각 1명으로 구성되어 있다. 네트워크 그림을 살펴보면 왼쪽에 전남대 통계학과 소속 저자가 위치하고 오른쪽에 한양대 응용수학과 소속 저자가 위치하고 있는데 이 두 개의 네트워크를 “손영숙”과 “김성욱”이 연결하고 있는 모습이다(이와 같이 두 하위 네트워크를 연결하는 선을 “브릿지”라고 함). 만일 이 연결이 없었다면 이 네트워크는 두 개의 하위 네트워크로 다시 나누어지며 각각은 전남대 통계학과 공동연구 네트워크, 한양대 응용수학과 공동연구 네트워크로 해석할 수 있다. 이 네트워크에서 중심성 값은 “손영

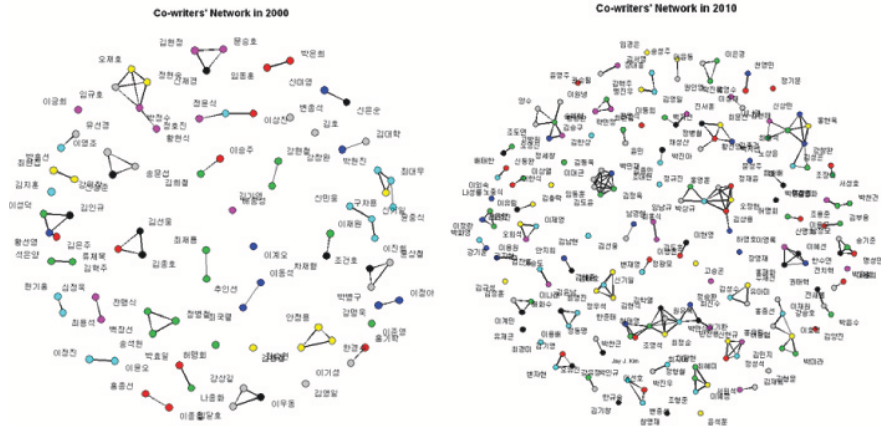


그림 3.5. 2000년과 2010년의 공저자 네트워크

표 3.7. 2000년과 2010년의 하위 네트워크 분포

크기	1	2	3	4	5	6	8	11	합
수(2000년)	10	18	9	2	1	0	0	0	40
수(2010년)	24	27	13	5	1	3	2	1	76

숙”이 근접 중심성 0.9, 중개 중심성 4.8로 가장 상위에 있다. 한편 “김성욱”의 중개 중심성은 4.2로 2위에 해당하는데 브릿지에 연결되어 있는 노드의 중개 중심성이 크게 나오는 것은 자명하다.

3.4. 연도별 네트워크 비교

2000년과 2010년, 두 해의 공저자 네트워크를 분석하였다. 이는 두 시기에 있어 공동연구의 양태가 달라졌는가를 알아보기 위함이다. 물론 표 3.2에서 보는 바와 같이 2000년에는 연 2회의 논문집이 발간되었고 2010년에는 연 6회의 논문집이 발간되어 논문의 편수 및 참여저자의 수가 다소 차이가 있다. 또한 2000년에는 교신저자를 따로 두지 않아 제 1저자에만 가중치를 1로 부여하였다. 그림 3.5는 2000년과 2010년의 공저자 네트워크 그림이다. 2000년 네트워크의 밀도는 0.0071이고 2010년 네트워크의 밀도는 0.0065이며 하위 네트워크의 수는 2000년이 40개 2010년이 76개이다. 하위 네트워크 수의 차이는 논문의 편수가 늘어나면서 나타난 자연스러운 차이로 생각된다. 한편 표 3.7에서 알 수 있듯이 하위 네트워크의 크기도 2000년에 비해 2010년에 커지는 경향이 있어 크기가 6이상 되는 하위 네트워크도 6개가 발견되었다. 이는 저자들 간의 공동연구의 범위가 2000년에 비해 다소 넓어졌음을 암시한다. 2000년의 네트워크에서 “박정수”는 근접 중심성 0.14, 중개 중심성 6.0으로 가장 높은 중심성이 가장 높은 노드였으며, 2010년에는 “박만식”이 근접 중심성 0.10, 중개 중심성 44.0으로 가장 높았다. 2000년 네트워크의 경우 가장 큰 하위 네트워크가 5개의 노드로 구성된 것이었는데 이 하위 네트워크를 구성하는 노드들 중 근접 중심성과 중개 중심성이 높은 노드가 찾아졌다. 또한 2010년의 네트워크를 보면 근접 중심성과 중개 중심성 상위 저자는 11개의 노드로 이루어진 가장 큰 하위 네트워크로부터 발견되어진다.

2000년과 2010년의 공저자 네트워크에 있어 두 저자  $i, j$ 의 공저자 관계를 종속변수로, 같은 소속관계를 설명변수로 고려한 로지스틱 회귀모형의 결과가 표 3.8에 각각 정리되어 있다. 2000년에 비해 2010년에 같은 소속기관 연구자 간 공동연구가 상대적으로 활발해 졌음을 알 수 있다.

표 3.8. 2000년과 2010년의 공저자 네트워크의 로지스틱 회귀분석 결과

	Coefficients	Estimate	Exp(b)	p-value
2000년	(intercept)	-4.6424	0.0096	0.193
	(동일기관 소속 여부)	3.6010	36.6332	< 0.001
2010년	(intercept)	-5.1908	0.0056	0.004
	(동일기관 소속 여부)	3.7461	42.3549	< 0.001

#### 4. 결론

본 연구는 지난 2000년부터 2010년까지 11년간 응용통계연구에 게재된 664개의 논문을 대상으로 747명의 저자 간 공저자 관계를 분석하였다. 전체 네트워크의 분석을 통해 288명의 저자로 구성된 대규모의 하위 네트워크를 발견하였다. 또한 네트워크의 중심성 분석을 통해 네트워크 중심에 위치하는 저자를 찾아보았다. 전체 네트워크의 경우 너무나 많은 하위 네트워크로 구성되어 있어 네트워크의 중심을 찾아보는 것이 큰 의미가 없으나 모든 노드가 서로 연결되어 있는 하위 네트워크에서 중심을 찾아보는 것은 네트워크의 중심을 파악하는데 있어 의미 있는 작업이다. 또한 공동연구에 동일 소속기관 여부가 영향을 미치는지를 로지스틱회귀분석을 통해 살펴보았다. 2000년과 2010년의 공저자 네트워크의 분석을 통해 동일 소속기관 연구자 간의 공동연구가 상대적으로 더 활발해 졌음을 알 수 있었다. 사회 네트워크 분석은 많은 영역에서 그 활용이 기대되어 진다. 특히 IT의 비약적 발전과 함께 쌍방향 소통의 채널이 다양하게 존재하게 됨에 따라 서로 관련을 파악하고 이를 분석하여 사회적 현상의 다양한 해석에 활용하려는 노력이 계속되어질 것으로 생각된다. 향후 연구과제로는 키워드와 연구자 간의 2부 네트워크 자료의 분석을 생각해 볼 수 있다. 이를 통해 연구자 간 연구 분야를 매개로한 네트워크를 분석이 가능할 것이다. 물론, 이러한 연구를 위해서는 키워드의 표준화작업을 통해 같은 내용을 서로 다른 용어로 표현한 키워드들의 통합 및 정제작업이 선행되어야 한다.

#### 참고문헌

손동원 (2002). <사회 네트워크 분석>, 경문사, 서울.

정민수, 정동준 (2007). 국내 예방의학 분야의 공저자-핵심어 네트워크와 군집 양상: 대한예방의학회지(1991~2006) 게재논문의 분석, <예방의학회지>, 14, 1-9.

허명희 (2010). <R을 활용한 사회네트워크분석 입문>, 자유아카데미, 경기.

Fruchterman, T. M. J. and Reingold, E. M. (1991). Graph drawing by force-directed placement, *Software-Practice and Experience*, 21, 1129-1164.

Kamada, T. and Kawai, S. (1989). An algorithm for drawing general undirected graphs, *Information Processing Letters*, 31, 7-15.

Malin, B. and Carley, K. (2007). A longitudinal social network analysis of the editorial boards of medical informatics and bioinformatics journals, *Journal of the American Medical Informatics Association*, 14, 340-348.

Newman, M. E. J. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, *Physical Review E*, 64, 016132-1-016132-6.

# Analysis of Papers in the Korean Journal of Applied Statistics by Co-Author Networks Analysis

M. Lee<sup>1</sup> · M. Park<sup>2</sup> · H. Lee<sup>3</sup> · S. Jin<sup>4</sup>

<sup>1</sup>Department of Statistics, Korea University; <sup>2</sup>Department of Preventive Medicine, Eulji University  
<sup>3</sup>Department of Statistics, Korea University; <sup>4</sup>Department of Informational Statistics, Korea University

(Received October 2011; accepted November 2011)

---

## Abstract

This study analyzed an aspect of co-author relationship in papers published in the Korean Journal of Applied Statistics by social network analysis. The data were extracted from 664 papers in the journal from 2000 to 2010. Authors at center of the network are detected by a network centrality analysis. Sub-network analysis found distinguishable research groups from the point of view of their topics or affiliations. The significance of affiliations to co-author relationship was examined by logistic regression analysis.

Keywords: Social network analysis, co-author networks, centrality, the Korean Journal of Applied Statistics.

---

---

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0004376).

<sup>4</sup>Corresponding author: Professor, Department of Informational Statistics, Korea University, Jociwon-eup, Yoengi-gun, Chungnam 339-700, Korea. E-mail: seohoon@korea.ac.kr