응용통계연구 (2011)
24(2), 401–409
DOI: 10.5351/KJAS.2011.24.2.401

사회네트워크분석에서 몬테칼로 방법의 활용

허명회[1] · 이용구[2]

[1]고려대학교 통계학과, [2]중앙대학교 통계학과

(2011년 1월 접수, 2011년 3월 채택)

요 약

사회네트워크분석(social network analysis)은 l개 연결선을 갖는 n개 노드의 자료를 대상으로 한다. 기본적인 자료기술로서 노드 간 최단거리(shortest distance), 근접 중심성(closeness centrality), 중개 중심성(betweenness centrality) 등을 산출한다. 기존의 사회학적 연구에서 다룬 네트워크는 대개 노드 수 n이 수십 또는 수백 정도였으나 최근에는 그 크기가 수십만 또는 수백만에 이르는 경우가 드물지 않다. 이에 따라 사회네트워크분석에서도 자료규모성(data scalability)의 이슈가 생겼다. 본 연구에서는 몬테칼로(Monte Carlo) 방법을 활용하여 $n = 100{,}000$ 규모의 임의 네트워크의 작은 세상(small world) 성질을 실증적으로 탐구하고 그 정도 규모에서의 중개 중심성과 근접 중심성의 산출 방법을 제안하고자 한다.

주요용어: 사회네트워크분석, 근접 중심성, 중개 중심성, 작은 세상, 자료 규모성, 몬테칼로 방법.

1. 연구 배경과 목적

사회네트워크는 다수의 연결된 개인(또는 기관)으로 구성된 사회적 구조로, 연결 유형은 친구, 친족, 공통 관심, 금융 거래, 성 관계 등 다양하다 (Wikipedia, “social network”, 2010/08). 사회네트워크분석은 다수의 점(vertices, nodes)과 이들은 연결하는 선(edges, lines)으로 구성된 망(網, 네트워크)에 대한 사회과학적·통계적 분석이다 (허명회, 2010).

사회네트워크분석은 노드(개인) 간 가장 짧은 길이의 연결인 최단경로(shortest path, geodesic path)와 노드들의 중심성(centrality)에 대한 측도의 산출에서 출발한다 (Wasserman과 Faust, 1994). 그런데, 네트워크의 노드 수가 n이면 모든 노드 쌍의 수는 $n(n-1)/2$이므로 모든 노드 쌍 간 최단거리의 산출은 매우 큰 계산시간을 요구한다. 또한 모든 노드 쌍에 대하여 메모리를 할당하는 경우 메모리 요구량을 시스템이 받쳐 주지 못할 수 있다. 이런 문제들 때문에 개인용 PC에서는 메모리로만 작업하는 경우 $n = 10{,}000$ 규모까지만 최단경로 및 중심성 지표들을 산출할 수 있다.

본 연구에서는 이런 문제들을 해결하기 위한 몬테칼로(Monte Carlo) 방법의 활용을 제안한다. 적용 예로서 $n = 100{,}000$ 규모의 임의 네트워크에 대하여 노드 간 최단거리의 분포를 찾고 아울러 중개 중심성과 근접 중심성을 산출함으로써 작은 세상 성질을 탐구한다.

2. 작은 세상

미국의 사회심리학자 Stanley Milgram은 한 사회에서 임의의 두 사람을 연결하기 위해 필요한 중개자 수는 의외로 매우 적다는 큰 네트워크의 “작은 세상(small world)” 성질을 입증하고자 실제 실험적 조

[1]교신저자: (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 통계학과, 교수. E-mail: stat420@korea.ac.kr

표 **2.1.** 임의 네트워크($n_0 = 1{,}000$, $\mu = 10$)에서 노드 간 최단거리 D의 분포

1) 완전열거에 의한 노드 간 최단거리의 분포 (상대빈도, %)

$D =$	1	2	3	4	5	∞	합계
	1.0	9.5	53.2	35.9	0.3	0.0	100.0

∗경과시간 0.81초

2) 몬테칼로 표본추출에 의한 노드 간 최단거리의 분포 (상대빈도, %)

$D =$	1	2	3	4	5	∞	합계
	1.2	9.9	52.5	36.1	0.3	0.0	100.0

∗ 경과시간 9.05초, 반복수 $N_{rep} = 10{,}000$

표 **2.2.** 임의 네트워크($n_0 = 1{,}000$, $\mu = 5$)에서 노드 간 최단거리 D의 분포

1) 완전열거에 의한 노드 간 최단거리의 분포 (상대빈도, %)

$D =$	1	2	3	4	5	6	7	8	∞	합계
	1.5	2.6	11.7	36.1	38.8	8.7	0.7	0.0	0.9	100.0

∗ 경과시간 0.58초

2) 몬테칼로 표본추출에 의한 노드 간 최단거리의 분포 (상대빈도, %)

$D =$	1	2	3	4	5	6	7	8	∞	합계
	0.6	2.4	12.0	35.5	38.7	8.8	0.7	0.0	1.2	100.0

∗ 경과시간 9.39초, 반복수 $N_{rep} = 10{,}000$

사연구를 한 바 있다 (Travers와 Milgram, 1969). 그의 연구에 취약점이 있어 한계가 있기는 하지만 그 충격은 매우 컸다.

이 절에서는 임의 네트워크에 대한 몬테칼로 분석결과를 제시하고자 한다. 연구 설계는 다음과 같다 (허명회, 2010). 구성원 수 n이 $n_0 + 1$인 사회에서 각 구성원이 평균 μ명을 안다고 가정한다. n_0은 1,000 또는 10,000으로 놓았고 μ는 10 또는 5로 놓았다. 그리고 모든 쌍의 구성원 i와 $j\,(\neq i)$에 대하여 i가 j를 아는 사건은 독립적으로 성공확률 $\theta\,(= \mu/n_0)$의 베르누이 분포를 따라 임의로 결정된다고 가정한다. 이렇게 생성된 임의 네트워크를 $\mathbf{A}$로 표기한다.

이와 같은 설정 하에서 구성원의 임의 쌍 (I, J)에 대하여 I로부터 $J(\neq I)$에 도달되는 경로 $I \to J$의 최단길이 $D_{I,J}\,(= D)$를 구하여 그 분포를 밝히기로 한다.

임의 네트워크 $\mathbf{A}$로부터 D의 분포를 추정하는 방법으로는 1) 완전열거(complete enumeration)와 2) 몬테칼로 표본추출(Monte Carlo sampling)을 고려한다. 여기서 전자의 완전열거는 모든 노드 쌍에 대하여 최단경로의 거리를 구하는 방법이고 후자의 몬테칼로 표본추출은 모든 노드 쌍 가운데 중복을 허락하여 임의추출된 N_{rep}개의 노드 쌍 각각에 대하여 최단경로의 거리를 구하는 방법이다.

연구는 R의 `igraph` 팩키지를 활용하여 수행되었다 (http://igraph.sourceforge.net/). 즉, 임의 네트워크 생성을 위해서는 `igraph` 라이브러리의 `random.graph.game` 함수를, 구성원(노드) 간 최단거리의 분포를 구하기 위해서는 `path.length.hist` 함수를, 2개 노드 간 최단거리를 구하기 위해서는 `get.shortest.paths` 함수를 사용하였다.

결과를 표 2.1, 2.2, 2.3, 2.4에 정리하였다. 표 2.1과 2.2는 $n_0 = 1{,}000$이고 μ가 각각 10과 5인 임의 네트워크에서 노드 간 최단거리의 분포를 보여준다.

$n_0 = 1{,}000$인 경우 μ 값에 무관하게 완전열거법에 의한 최단거리분포와 몬테칼로 방법에 의한 최단거리분포가 유사하게 나타났다. 표 2.1의 완전열거법에 의하면 자신 외 평균 10명(=μ)을 아는 사회에서는

표 **2.3.** 임의 네트워크($n_0 = 10{,}000$, $\mu = 10$)에서 노드 간 최단거리 D의 분포

1) 완전열거에 의한 노드 간 최단거리의 분포 (상대빈도, %)

$D =$	1	2	3	4	5	6	7	∞	합계
	0.1	1.0	9.3	52.5	36.6	0.5	0.0	0.0	100.0

∗ 경과시간 71.8초

2) 몬테칼로 표본추출에 의한 노드 간 최단거리의 분포 (상대빈도, %)

$D =$	1	2	3	4	5	6	7	∞	합계
	0.1	1.1	9.0	52.2	37.1	0.6	0.0	0.0	100.0

∗ 경과시간 29.6초, 반복수 $N_{rep} = 10{,}000$

표 **2.4.** 임의 네트워크($n_0 = 10{,}000$, $\mu = 5$)에서 노드 간 최단거리 D의 분포

1) 완전열거에 의한 노드 간 최단거리의 분포 (상대빈도, %)

$D =$	1	2	3	4	5	6	7	8	9	10	∞	합계
	0.1	0.3	1.2	5.9	22.6	43.5	22.1	2.9	0.2	0.0	1.4	100.0

∗ 경과시간 50.7초

2) 몬테칼로 표본추출에 의한 노드 간 최단거리의 분포 (상대빈도, %)

$D =$	1	2	3	4	5	6	7	8	9	10	∞	합계
	0.0	0.2	1.2	6.0	23.1	43.0	22.0	2.8	0.2	0.0	1.4	100.0

∗ 경과시간 32.6초, 반복수 $N_{rep} = 10{,}000$

임의 2인 사이의 거리가 3단계인 빈도가 가장 높았다(53.2%). 불과 2명을 사이에 두면 두 사람이 연결될 수 있다는 "작은 세상" 현상을 볼 수 있다. 표 2.2의 완전열거법에 의하면 자신 외 평균 5명(=μ)을 아는 사회에서는 임의 2인 사이의 거리가 5단계인 빈도가 가장 높았다(38.8%).

$n_0 = 1{,}000$인 경우 μ 값에 관계없이 요구되는 계산 시간은 완전열거법이 몬테칼로 방법보다 짧게 나타났다.

표 2.3과 2.4는 $n_0 = 10{,}000$이고 μ가 각각 10과 5인 임의 네트워크에서 노드 간 최단거리의 분포를 보여준다. 이런 경우에도 분포를 얻는 방법에 관계없이 두 결과가 유사하였다. 자신 외 평균 10명(= μ)을 아는 사회에서는 임의 2인 사이의 거리가 4단계인 경우가 가장 많았고(52.5%, 완전열거법), 자신 외 평균 5명(= μ)을 아는 사회에서는 임의 2인 사이의 거리가 6단계인 경우가 가장 많았다(43.5%, 완전열거법).

$n_0 = 10{,}000$인 경우에는, $n_0 = 1{,}000$인 경우와 달리, μ 값에 관계없이 요구되는 계산 시간은 완전열거법에 비해 몬테칼로 방법이 짧게 나타났다. 완전열거법이 노드 수의 제곱 n^2에 비례하는 계산시간을 요구하여 컴퓨터에 부담을 주지만 몬테칼로 표본추출법은 그렇지 않은 것이다.

더 나아가, n_0를 100,000으로, 자신 외 아는 사람의 평균 수 μ를 10으로 설정해 보았다. 이 경우에는 연구자의 노트북(Intel Core2 Duo CPU T7500 @2.2GHz, 2.00G RAM @789MHz)에서 완전열거법은 사실상 작동하지 않았다. 경과시간 5,400초에서 계산결과를 내지 못한 것이다. 반면 몬테칼로 표본추출법은 표 2.5의 결과를 산출하였다($N_{rep} = 10{,}000$).

몬테칼로 연구의 결과, $n_0 = 100{,}000$이고 $\mu = 10$인 네트워크에서는 임의 2인 사이의 거리가 5단계인 경우가 가장 많았다(52.4%). $n_0 = 100{,}000$이고 $\mu = 5$인 네트워크에서는 임의 2인 사이의 거리가 7단계인 경우가 가장 많았다(36.7%).

표 **2.5.** 임의 네트워크($n_0 = 100{,}000$)에서 노드 간 최단거리 D의 분포 – 몬테칼로 표본추출의 결과

1) $\mu = 10$인 경우, 노드 간 최단거리의 분포 (상대빈도, %)

$D =$	1	2	3	4	5	6	7	8	∞	합계
	0.0	0.1	0.9	9.1	52.4	37.0	0.5	0.0	0.0	100.0

$*$ 경과시간 337초, 반복수 $N_{rep} = 10{,}000$

2) $\mu = 5$인 경우, 노드 간 최단거리의 분포 (상대빈도, %)

$D =$	1	2	3	4	5	6	7	8	9	10	∞	합계
	0.0	0.0	0.1	0.8	3.2	13.3	36.7	35.5	8.1	0.8	1.3	100.0

$*$ 경과시간 392초, 반복수 $N_{rep} = 10{,}000$

3. 중개 중심성

사회네트워크에서 각 노드의 중개 중심성(betweenness centrality)은 모든 노드 쌍의 최단경로에서 해당 노드를 통과하는 경우의 상대적 빈도로 정의된다. 즉, 노드 v의 중개 중심성 $C_B(v)$는

$$C_B(v) = \frac{1}{n(n-1)} \sum_i \sum_j \frac{g_{ij}^{(v)}}{g_{ij}} I\left(g_{ij} \geq 1\right), \quad v = 1, \ldots, n$$

이다. 여기서 g_{ij}는 i에서 j로 들어가는 최단경로의 수이고 $g_{ij}^{(v)}$는 i에서 j로 들어가는 최단경로 중 노드 v를 통과하는 경로의 수이다. 따라서 모든 노드 쌍 (i, j)에 대하여 $g_{ij} = 1$ 또는 0인 경우 $C_B(v)$는

$$C_B^*(v) = \frac{1}{n(n-1)} \sum_i \sum_j I\left(g_{ij}^{(v)} = 1\right), \quad v = 1, \ldots, n$$

과 일치한다. 노드 수 n의 네트워크에서 가능한 노드 쌍의 수가 $n(n-1)$이므로 $C_B^*(v)$는 네트워크 내 임의 2개 노드 간 연결 경로에서 노드 v가 놓이는 상대적 빈도로 해석될 수 있다. 그러나 n이 큰 경우 모든 쌍의 노드 간 최단경로를 산출하여 노드 v의 중개 중심성을 산출하는 방식은 막대한 부담을 초래한다.

대안으로서 다음과 같은 몬테칼로 방식의 중개 중심성 평가를 제안한다.

0) $v = 1, \ldots, n$에 대하여 $T(v)$를 0으로 놓는다.

1) 노드 쌍 (i, j)를 단순임의추출한다.
   $i \to j$ 최단경로를 찾는다.
   $v = 1, \ldots, n$에 대하여 그 중에서 노드 $v = 1, \ldots, n$가 포함되면 $T(v)$를 1만큼 증가시킨다.

2) 단계 1을 N_{rep}회 반복한다.

3) $v = 1, \ldots, n$에 대하여 $C_B(v)$를 $T(v)/N_{rep}$로 평가한다.

표 3.1은 $n_0 = 1{,}000$인 임의 네트워크에서 임의 노드 V에 대한 중개 중심성 $C_B^*(V)$를 산출하여 평균과 표준편차를 본 것이다. 자신 외 연결되는 평균 노드 수 μ는 10 또는 5로 놓았다. $\mu = 10$의 네트워크에서는 완전열거법의 경우 중개 중심성의 평균이 0.225%, 몬테칼로법에서는 0.227%로 상당히 일치하는 것으로 나타났다. 다만, 표준편차에서는 차이를 보였다. 이는 몬테칼로법이 $C_B(V)$ 대신 $C_B^*(V)$를 평가하기 때문으로 생각된다. 그러나 두 방법이 산출한 지표 간 상관은 0.72로서 비교적 일치하는 결과를 보였다. $\mu = 5$인 네트워크에서는 완전열거법의 경우 중개 중심성의 평균이 0.090%, 몬테칼로법에서는

표 **3.1.** 임의 네트워크($n_0 = 1{,}000$)에서 중개 중심성 $C_B(V)$의 평균과 표준편차 (단위: %), 경과시간 (단위: 초)

1) $\mu = 10$인 경우

	평균	표준편차	경과시간
– 완전열거	0.225	0.100	1.97
– 몬테칼로	0.227	0.139	9.82

2) $\mu = 5$인 경우

	평균	표준편차	경과시간
– 완전열거	0.0899	0.0123	10.4
– 몬테칼로	0.0897	0.1892	9.0

표 **3.2.** 임의 네트워크($n_0 = 10{,}000$)에서 중개 중심성 $C_B(V)$의 평균과 표준편차 (단위: %), 경과시간 (단위: 초)

1) $\mu = 10$인 경우

	평균	표준편차	경과시간
– 완전열거	0.0326	0.0149	218
– 몬테칼로	0.0327	0.0246	31

2) $\mu = 5$인 경우

	평균	표준편차	경과시간
– 완전열거	0.0048	0.0032	126
– 몬테칼로	0.0048	0.0039	33

표 **3.3.** 임의 네트워크($n_0 = 100{,}000$)에서 중개 중심성 $C_B(V)$의 평균과 표준편차 (단위: %), 경과시간 (단위: 초)

1) $\mu = 10$인 경우

	평균	표준편차	경과시간
– 완전열거	N/A	N/A	2,740
– 몬테칼로	0.00426	0.00684	347

2) $\mu = 5$인 경우

	평균	표준편차	경과시간
– 완전열거	N/A	N/A	6,820
– 몬테칼로	0.00624	0.00902	406

0.090%로 일치하였다. 표준편차에서는 차이가 있었으나 두 방법이 산출한 지표 간 상관은 0.92로서 일치도가 $\mu = 10$인 네트워크에 비해 향상되었다. 이는 $\mu = 5$인 네트워크가 $\mu = 10$인 네트워크에 비해 최단경로의 중복성이 작기 때문으로 생각된다.

표 3.2는 $n_0 = 10{,}000$인 임의 네트워크에서 나온 결과이다. 결과에서 나타난 전체적 양상은 $n_0 = 1{,}000$인 경우, 즉 표 3.1과 유사하다. 그러나 계산 시간에 있어서는 완전열거법과 몬테칼로법이 역전되어, 몬테칼로법이 완전열거법에 비해 상대적으로 훨씬 작게 걸렸다.

$n_0 = 100{,}000$인 임의 네트워크에 대하여는 완전열거법 계산이 불가능하였다. 그러나 표 3.3에서 보듯 몬테칼로법은 600초 이내에 결과를 산출하였다.

4. 근접 중심성

사회네트워크에서 근접 중심성(closeness centrality)은 임의 노드에 이르는 거리들의 평균한 다음 그 역

을 취하여 중심에 있을수록 지표 값이 커지도록 정의된다. 이 연구에서는 조화평균에 의한 정의

$$C_K(v) = \frac{1}{n-1}\sum_j \frac{1}{d_{vj}}, \quad v = 1, \ldots, n$$

을 사용하기로 한다. 산술평균에 의한 정의를 피하려는 이유는 노드 간 거리가 하나라도 무한(∞)인 경우 지표 값이 0이 되기 때문에 지표로서의 변별력이 떨어지기 때문이다.

모든 노드 v에 대하여 $C_K(v)$를 산출하려면 (i,j)의 모든 쌍에 대하여 최단경로 d_{ij}가 계산되어야 하므로 n이 커짐에 따라 계산 부담이 커질 수밖에 없다. 이를 타개하는 방안으로 크기 n_1의 노드 표본을 추출하여 "훈련용" 부(副)네트워크 $\mathbf{S} = \{(i,j) \,|\, i \le n_1, j \le n_1\}$를 만들고 이 네트워크의 개별 노드들과 테스트 노드 v 간 평균 최단거리를 노드 v의 근접 중앙성 지표로 정의하는 것이 어떨까? 즉, 테스트 노드 $v\,(\ge n_1+1)$의 근접 중심성을 다음으로 제안한다.

$$C_K^*(v) = \frac{1}{n_1}\sum_{j=1}^{n_1} \frac{1}{d_{vj}^*}, \quad v = n_1+1, \ldots, n.$$

여기서 d_{vj}^*는 훈련용 부네트워크 $\mathbf{S}$에 테스트 노드 v가 붙은 확장 부네트워크에서 노드 v와 부네트워크 $\mathbf{S}$의 노드 j 사이 최단거리이다.

d_{vj}^*의 산출을 위해 테스트 노드 v로부터 부네트워크의 노드 j까지의 최단경로를 매번 새로 계산해야 할 필요는 없다. 다음 관계를 활용할 수 있기 때문이다.

정리 4.1 노드 $v\,(=n_1+1, \ldots, n)$과 노드 $j\,(=1, \ldots, n_1)$에 대하여

$$d_{vj}^* = \min_{i=1,\ldots,n_1;\, a_{vi}=1} d_{ij} + 1$$

이다. 여기서 a_{vi}는 $\mathbf{A}$의 (v,i) 요소이다.

증명: 노드 v로부터 노드 j에 도달하려면 부네트워크 $\mathbf{S}$에 착지를 해야 하는데 착지가 가능하다면 착지 노드는 노드 v와 연결되는 노드여야 한다. 그 노드를 i라고 하자. 노드 v로부터 노드 j에 이르는 최단경로는 노드 v에서 노드 i에 이르고 다시 노드 i에서 노드 j에 이르는 경로 중 하나에서 나온다. 만약 착지가 가능하지 않다면 $d_{vj}^* = \infty$이다. 따라서 어느 경우이든 위 식이 성립한다. □

n = 10,000의 임의 네트워크에서 $\mu = 10$인 경우, 몬테칼로 알고리즘(n_1 = 1,000)을 적용한 결과 임의 노드 V의 근접 중심성 $C_K^*(V)$는 평균과 표준편차가 각각 0.00376, 0.00571로 나타났고 경과시간은 2,298초였다. $\mu = 5$인 경우는 근접 중심성 $C_K^*(V)$의 평균과 표준편차가 각각 0.000688, 0.00108이었다. 경과시간은 2,315초였다.

n = 100,000의 임의 네트워크에서 $\mu = 10$인 경우, 몬테칼로 알고리즘(n_1 = 1,000)을 적용한 결과 근접 중심성의 평균과 표준편차가 각각 0.000106, 0.000340으로 나타났고 경과시간은 24,851초였다. $\mu = 5$인 경우는 근접 중심성의 평균과 표준편차가 각각 0.0000116, 0.0001127로 나타났고 경과시간은 5,405초였다.

5. 응용사례

이 절에서는 Zachary (1977)의 가라테 클럽(karate club) 네트워크에 몬테칼로 방법을 적용해보기로 한다. 이 네트워크는 2년에 걸쳐 34명의 구성원간 인간 관계에 대한 연구에서 나왔다. 그림 5.1은 가라테 클럽 자료에 대한 네트워크 그래프이다.

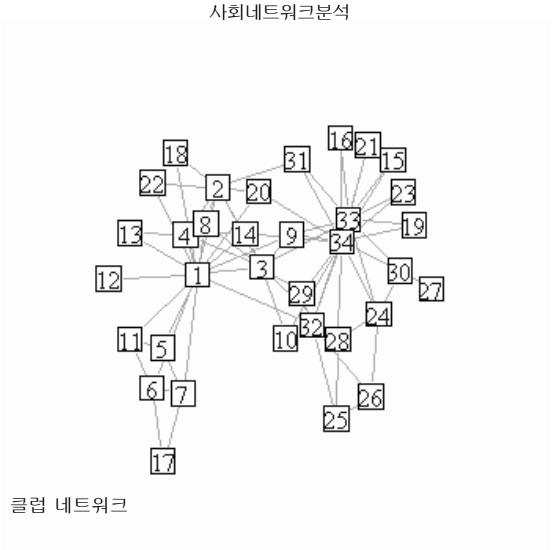

그림 **5.1.** Zachary의 가라테 클럽 네트워크

표 **5.1.** 가라테 클럽 네트워크에서 노드 간 최단거리의 분포(%)

거리	1	2	3	4	5	6+
몬테칼로 방법	12.6	48.3	24.7	12.5	1.9	0.0
완전열거법	13.7	47.4	24.4	13.0	1.4	0.0

표 **5.2.** 가라테 클럽 네트워크에서 34개 노드의 중개 중심성

노드	1	2	3	4	···	31	32	33	34
몬테칼로 방법	0.46	0.03	0.29	0.00	···	0.00	0.07	0.25	0.14
완전열거법	0.41	0.06	0.13	0.01	···	0.01	0.13	0.15	0.28

노드 수가 크지 않으므로 굳이 몬테칼로 방법을 적용하지 않고 기존의 완전열거법으로 분석이 가능하다. 그럼에도 불구하고 이 사례를 연구하고자 하는 이유는 몬테칼로 방법을 완전열거법과 비교하기 위해서이다.

표 5.1은 노드 간 거리의 분포를 몬테칼로 방법과 완전열거법으로 구한 결과이다. 몬테칼로 방법에서 반복 수는 1,000회였다. 두 방법의 결과가 상당히 일치하는 것을 확인할 수 있다.

표 5.2는 34개 노드에 대한 중개 중심성을 몬테칼로 방법과 완전열거법으로 구한 결과이다. 몬테칼로 방법에서 반복 수는 1,000회였다.

표 5.2에서 2개의 결과 간 상관계수는 0.90으로 나타났다. 노드 1과 노드 33, 노드 34의 중심성이 두드러지게 높은 점 등 두 결과가 상당히 일치하는 것을 볼 수 있다. 결국 이 클럽은 노드 1과 노드 34를 중심으로 분열되었다고 한다.

마지막으로, 노드 수 17의 부(副)네트워크를 임의추출하여 나머지 노드들에 대한 근접 중심성을 산출하고 완전열거법에 의한 결과와의 일치도로 상관계수를 구하여 보았다. 그림 5.2가 독립적으로 얻은 1,000개 상관계수들의 히스토그램이다. 상관계수들의 중간값이 0.78, 평균이 0.77로 나타나, 몬테칼로 방법에 의한 중심성 값이 근사적 목적으로 활용될 수 있다고 하겠다.

6. 맺음말

십수년 전까지만 하더라도 사회학자들의 연구에서 다루어진 사회네트워크는 100개 정도의 구성되었다.

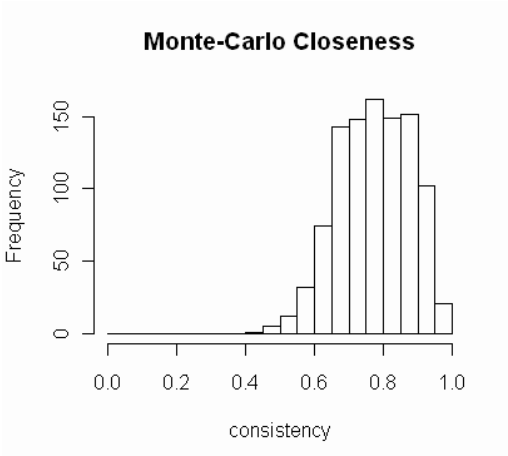

그림 5.2. 몬테칼로 근접 중심성과 완전열거 근접 중심성 간 일치도(상관계수)

그러나 최근에는 정보기술의 보편화에 따라 이보다 훨씬 큰 대규모의 사회네트워크가 전자적으로 생성되고 있다.

이에 따라 사회네트워크분석에서 규모성(scalability) 문제는 필히 해결이 필요한 과제가 되었다. 이 연구에서는 메모리 작업의 전제 하에서 완전열거방법에 대한 대안적으로 노드 간 거리의 분포, 중개 중심성, 근접 중심성 등의 근사값을 산출해주는 몬테칼로 알고리즘을 제안하였다.

참고문헌

허명회 (2010). <R을 활용한 사회네트워크분석 입문>, 자유아카데미, 서울.

Travers, J. and Milgram, S. (1969). An experimental study of the small world problem, *Sociometry*, **32**, 425–443.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*, Cambrdge University Press.

Zachary, W. (1977). An information flow model for conflict and fission in small groups, *Journal of Anthropological Research*, **33**, 452–473.

Monte-Carlo Methods for Social Network Analysis

Myung-Hoe Huh[1] · Yonggoo Lee[2]

[1]Department of Statistics, Korea University; [2]Department of Statistics, ChungAng University

(Received January 2011; accepted March 2011)

Abstract

From a social network of n nodes connected by l lines, one may produce centrality measures such as closeness, betweenness and so on. In the past, the magnitude of n was around 1,000 or 10,000 at most. Nowadays, some networks have 10,000, 100,000 or even more than that. Thus, the scalability issue needs the attention of researchers. In this short paper, we explore random networks of the size around $n = 100{,}000$ by Monte-Carlo method and propose Monte-Carlo algorithms of computing closeness and betweenness centrality measures to study the small world properties of social networks.

Keywords: Social network analysis, closeness centrality, betweenness centrality, small world, data scalability, Monte Carlo method.

[1]Corresponding author: Professor, Department of Statistics, Korea University, 5-1 Anam-Dong, Sungbuk-Gu, Seoul 136-701, Korea. E-mail: stat420@korea.ac.kr