Journal of the Korean Institute of Industrial Engineers
Vol. 40, No. 1, pp. 1-7, February 2014.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2014.40.1.001

<Original Research Paper>

휴대폰 통화기록 기반의 소셜 컨택 네트워크 구성 및 Scale-free 특성에 관한 분석

이 진 호†

해군사관학교 경영과학과

Constructing a Social Contact Network based on Cellphone Call Records and Analysis of its Scale-free Property

Jinho Lee

Department of Management Science, Korea Naval Academy

We consider a human contact social network that has connections through cellphone addresses. To construct such a social network, we use real call records provided by a large carrier, and connect to each other if there exists a call record between any two cellphone users. Due to its huge amount of data, we down-sample it in a way that the smallest-degree nodes are removed, in turn, from the network. For a moderate size of the networks we show that the degree distribution of the network follows a power-law distribution via linear regression analysis, implying the so-called scale-free property. We finally suggest some alternative measures to analyze a social network.

Keywords: Social Network, Scale-Free Property, Degree Distribution

1. 서 론

소셜 네트워크(Social Network)에 대한 분석은 최근 사회적으로 여러 학문 분야에서 많은 관심을 불러 일으키고 있다. 인터넷 보급의 일반화와 개인 컴퓨터 및 스마트폰의 휴대가 인적 교류의 장을 온라인으로 급속도로 빠르게 확산시키고 있으며, 이로 인해 SNS(Social Network Service)상에서 지인들과 지속적으로 교류를 유지하고 정보 교환 및 관리를 활발히 진행하고 있다. 이러한 소셜 네트워크의 성황은 기업들에게 마케팅의 개념도 기존의 전통적인 신문, 방송 매체 등의 방식에서 벗어나 SNS를 활용한 마케팅 전략 개발에 관심을 기울이게 하고 있다(Lee, 2013). 이는 비단 마케팅 전략뿐만 아니라 대통령 선거 등에서도 적극적으로 활용되고 있다(Lim, 2012). 이른바 '빅데이터'를 분석하여 마케팅 또는 선거 전략에 도입함으로서 그 효과를 얻고 있는 셈이다. 마케팅의 측면에서 복잡하고 대형의 네트워크 공간상 어떤 소비자 또는 유권자를 공략하는 것이 제품 광고 또는 인지도 상승의 효과를 가져다 줄 것인가에 대한 연구(Chen *et al*., 2009)를 비롯하여 복잡한 네트워크 상에서 어떤 흐름의 제어 및 통제 또는 탐지 등을 통해 네트워크 또는 시스템의 정상적인 작동 및 운용을 위한 의사결정에도 많은 연구가 진행되고 있다. 대표적으로 컴퓨터 네트워크 상 바이러스의 확산을 조기에 탐지하는 모델에 대한 연구(Balthrop *et al*., 2004), 스마트폰의 일반화로 인한 스마트폰 보안 문제(Fleizach *et al*., 2007), 그리고 전염병의 확산을 방지하기 위한 인적 네트워크에 대한 연구(Dimitrov and Meyers, 2010) 등이 있다.

하지만, 빅데이터로 대표되는 소셜 네트워크는 그 데이터의 확보가 어려우며, 방대한 양으로 인해 실제 네트워크를 구현해 내기가 쉽지 않고 수천만 명의 연결정보를 모두 확보한

본 연구는 국내의 모(某) 이동통신회사의 지원을 받아 수행된 연구임.
† 연락저자 : 이진호 교수, 645-797 경남 창원시 진해구 앵곡동 사서함 88-1-4 해군사관학교 경영과학과, Tel : 055-549-1092, Fax : 055-542-0033, E-mail : jinho7956@hotmail.com
2013년 11월 13일 접수; 2013년 12월 28일 수정본 접수; 2014년 1월 19일 게재 확정.

다고 해도 매우 대형이므로 네트워크의 위상(Topology)에 대한 특성을 파악하기가 어렵다. 따라서 네트워크 내의 흐름의 최적화 등의 연구 이전에 소셜 네트워크 자체에 대한 위상적 분석을 시도하는 연구가 많이 진행되어 오고 있다. 본 논문에서는 접근성이 어려운 실제 데이터를 바탕으로 네트워크를 구성하고 특성을 분석하며 다음의 사항에 대하여 기존 연구와 차별화한다.

1. 휴대전화 사용자간의 실제 통화 기록을 바탕으로 하는 소셜 컨택 네트워크를 정의하고 사용자간의 연결 현황을 통화 기록을 토대로 추출하여 네트워크를 구성한다.
2. 구현된 소셜 컨택 네트워크의 위상적 특징을 Degree(연결정도 : 노드에 연결된 아크의 수 또는 인접한 이웃 노드의 수)의 분포를 분석하여 Barbasi and Albert(1999)가 정의한 Scale free 특성 여부를 확인한다.
3. 네트워크의 크기를 다운 샘플링(Down Sampling)하는 방법을 제안하고, 다운 샘플링된 네트워크에 대한 Scale free 특성 보존 여부를 확인한다.
4. Degree 분석 기반의 Scale free 특성 이외의 네트워크의 위상적 특징을 분석하는 대안들을 제시한다.

따라서 본 연구를 통해 소셜 컨택 네트워크의 Scale free 특성 여부를 확인하고, 네트워크의 크기를 감소하는 적절한 방법에 대해 고찰하며 나아가 Degree 분석을 통한 특성 이외의 대안들에 대해 제안하고자 한다.

2. 네트워크 구성

휴대폰 사용자간 실제 통화 기록을 바탕으로 한 소셜 네트워크를 본 논문에서는 소셜 컨택 네트워크(Social Contact Network)라 정의한다. 본 장에서는 국내의 한 이동 통신회사로부터 연구 목적으로 제공받은 사용자간의 통화 기록을 토대로 소셜 컨택 네트워크를 구성하는 데 그 구체적인 내용은 다음과 같다. 제공받은 데이터는 해당 회사에 가입한 휴대폰 이용자 중 임의로 10%를 선정하고 총 1년간(2009년)의 통화 기록, 문자 메시지 및 휴대폰의 종류, 가입 내용 등이다. 소셜 컨택 네트워크를 구성하는 데 필요한 통화 기록을 별도로 추출하였고 다른 내용은 사용하지 않기로 하였다.

소셜 컨택 네트워크는 일반적인 네트워크와 마찬가지로 노드(Node)와 아크(Arc)로 구성되는데 다음과 같은 가정 사항을 토대로 구성한다.

1. 노드는 휴대폰 사용자로 정의하고 노드간을 연결하는 아크는 특정한 두 사람간 통화 기록이 있으면 아크가 존재하는 것으로 가정한다.
2. 아크의 방향성은 양방향으로 가정하여 통화 기록이 존재하면 두 사람 간 양방향으로 의사소통이 가능한 것으로 가정한다.
3. 문자 메시지의 전송은 불특정 다수에게 홍보 및 광고 또는 캠페인(Campaign), 스팸(Spam) 등에 의해 발송되는 경우가 많은 것을 감안하여 고려하지 않는다.
4. 동일한 통화 대상자간 중복된 기록 및 통화 시간에 대해서는 고려하지 않았고, 최소 한 번 이상의 통화 기록이 있으면 해당하는 두 노드 간을 연결하도록 한다.

가정 2는 통화 기록이 있는 두 사람은 각자 서로의 전화번호를 저장하고 있을 것이라고 가정한 데서 비롯된 것이며, 가정 3은 문자 메시지의 경우만으로 두 사람이 서로 알고 연락하며 지내는 사이라고 보기 어려운 점에서 제외하기로 한다. 또한 본 연구는 컨택 네트워크를 고려하므로 가정 4를 통해 통화의 빈도(Frequency)는 고려하지 않고 연결 여부만을 확인하도록 한다. 컨택 네트워크를 구성하는데 필요한 실제 통화 기록은 통신회사에서 제공하는 데이터에 기반하는데 일반적으로 통신회사는 월별 이용내역을 토대로 요금을 책정하므로 월 단위의 통화 기록을 제공한다. 따라서 전화를 건 송신자의 기록으로만 작성되어 있고 수신자는 해당 통신회사의 고객이 아닐 수도 있다. 컨택 네트워크를 구성하는데 있어서 수신자가 해당 통신회사의 고객이 아닐 경우는 제외하였는데 그 이유는 통신회사 입장에서 해당 고객의 관리를 우선으로 한다는 점과 타 회사 고객의 경우 통화 기록상 나타나는 전화번호를 제외한 다른 정보를 알 수 없기 때문이다. 제공받은 데이터에서 월 평균 약 2,880만건의 통화 기록이 나타나고 월별 별개의(Distinct) 통화 기록의 총 량은 <Table 1>과 같이 나타났다.

Table 1. Number of monthly distinct calls in 2009.

Month	number of distinct call pairs(million)	Month	number of distinct call pairs(million)
12	31.3	6	29.1
11	28.8	5	29.2
10	30.7	4	28.5
9	30.0	3	28.0
8	28.5	2	25.7
7	29.8	1	26.1
Average		28.8	

<Table 1>에서 나타난 통화 기록을 추출한 후 3개월 단위로 통화 기록을 누적해가며 네트워크를 형성한다. 즉, 최초 3개월간 통화 기록을 바탕으로 하나의 네트워크를 형성한 후 다음 3개월을 최초 네트워크에 누적시켜서 추가적인 노드와 아크로 네트워크를 형성하는 것이다. 이렇게 해서 최대 1년간의 통화 기록을 모두 누적한 네트워크를 최종적으로 형성한다. 현재 네트워크에서 누적된 네트워크를 형성할 때에는 먼저 노드가 현재 네트워크에 존재하는지 확인하고 존재하지 않는 노드에 대하여는 노드를 추가하며, 새로운 아크가 현재 네트워크에

이미 존재하는지를 확인한 후 존재하지 않으면 아크를 생성하도록 한다. 이렇게 하여 생성된 네트워크를 G_r로 정의하며 r은 누적된 개월수를 의미한다. 총 5개의 다른 누적 개월수를 고려하는데 그 값은 $r=1, 3, 6, 9, 12$이다. 즉, 처음 1개월 동안의 통화 기록을 기반으로 구성한 컨택 네트워크를 바탕으로 3개월 동안의 네트워크, 6개월, 9개월 그리고 최종적으로 12개월 동안의 총 통화 기록을 모두 누적한 네트워크까지 형성한다. $G_r=(N_r, A_r)$로 구성하는데 N_r은 해당 r개월 동안 누적된 컨택 네트워크의 노드 집합이며 A_r은 아크의 집합이다. <Table 2>는 이렇게 형성된 네트워크의 노드수, 아크수, 평균 및 최대 Degree를 보여준다. 여기서 d_{avg}와 $d_{\max}$는 각각 평균 및 최대 Degree, 즉 평균적으로 하나의 노드당 연결되어 있는 아크의 수 그리고 네트워크 전체에서 연결된 아크의 수가 가장 많은 노드의 아크 수를 의미한다. 컨택 네트워크가 양방향 아크를 가지는 네트워크로 정의되었으므로 $d_{avg}=2\times$노드수/아크수로 계산되었다. <Table 2>에서 나타나는 것처럼 d_{avg}는 통화 기록이 누적될수록 증가하긴 하지만 그 증가비율은 점차 감소한다. 이것은 노드 및 아크수의 증가에서도 나타난다. 놀라운 점은 G_1에서 나타나는 $d_{\max}$가 2,125라는 것인데, 이것이 의미하는 것은 특정 1인이 한 달간 무려 2,125명의 다른 사람에게 통화한 기록이 존재한다는 것으로 하루 평균 약 70명의 매번 다른 사람과 통화를 했으며 하루 24시간 내내 매 20분 간격으로 다른 사람에게 통화를 건 것과 마찬가지이다. 이처럼 실 통화 기록을 바탕으로 컨택 네트워크를 구성한 결과 극단적으로 많은 아크를 보유하고 있는 노드도 존재할 수 있음을 참고하기 바란다.

Table 2. Contact networks with different cumulation periods

G_r	Number of nodes	Number of arcs	d_{avg}	$d_{\max}$
G_1	13,748,461	31,272,417	4.5	2,125
G_3	17,137,361	59,329,219	6.9	4,980
G_6	19,622,089	92,556,677	9.4	8,849
G_9	21,649,040	123,109,834	11.4	12,047
G_{12}	22,931,058	148,721,984	13.0	15,343

마지막으로 총 고객 중 임의로 선택된 10%에 대해 연결성(Connectedness)을 검사해 본다. 네트워크의 연결성은 Ahuja *et al.*(1994)의 정의처럼 모든 노드들이 아크들을 거쳐서 도달될 수 있는지를 나타낸다. 모든 노드가 도달될 수 있다면 그 네트워크는 연결되어 있다고 말할 수 있으나 만약 어떤 특정한 노드들은 도달될 수 없다면 도달이 가능한 노드들 간의 집합을 Component라 정의하며, 모든 네트워크는 하나 이상의 Component가 존재하게 된다. <Table 2>에서 구한 네트워크를 Ahuja *et al.*(1994)에서 제시한 Breadth-first Search(BFS) 방법을 이용하여 연결성을 검사한 후 만약 연결되지 않았다면 가장 큰 Component를 채택하도록 하며, <Table 3>은 그 결과를 보여준다.

< Table 3>에서 나타나듯이 실 통화 기록 기반의 컨택 네트워크는 매우 높은 연결성을 보여주며, 누적 기간이 늘어날수록 그 연결성이 증가함을 알 수 있다. 또 하나의 놀라운 점은 단 1개월 동안의 통화 기록만으로도 약 1,370만 명이 컨택 네트워크 상에서 연결되어 있다는 점이다. 12개월 동안의 총 통화 기록을 모두 누적하여 구한 LC_{12}는 약 2,300만 명이 휴대폰 전화번호들을 통해서 서로에게 도달할 수 있음을 보여주는데 이는 국내 인구의 약 절반에 해당하는 수준이며 휴대폰을 실제 보유하고 있는 사람의 수를 고려할 때 절반 이상이 컨택 네트워크를 통해 연결되어 있음을 알 수 있다. 가장 큰 Component만을 취하여 구성한 최종 컨택 네트워크는 <Table 4>에서 나타난다.

Table 3. Largest component size of contact networks with different cumulation periods

G_r	Number of nodes	Component size	% of Component size to number of nodes
G_1	13,748,461	13,683,151	99.52%
G_3	17,137,361	17,123,988	99.92%
G_6	19,622,089	19,615,666	99.96%
G_9	21,649,040	21,644,618	99.97%
G_{12}	22,931,058	22,927,621	99.98%

Table 4. Connected contact networks after taking the largest components

G_r	Number of nodes	Number of arcs	d_{avg}	$d_{\max}$
G_1	13,683,151	31,231,085	4.6	2,125
G_3	17,123,988	59,321,217	6.9	4,980
G_6	19,615,666	92,552,929	9.4	8,849
G_9	21,644,618	123,107,251	11.4	12,047
G_{12}	22,927,621	148,719,957	13.0	15,343

3. 네트워크 다운 샘플링(Down-Sampling)

제 2장에서 구한 컨택 네트워크는 모평균의 10%를 임의로 추출하여 구성하였으므로 모집단 전체를 고려한 네트워크와 유사한 형태를 가지는 지 알 수 없다. 그렇다고 해서 모집단 전체의 통화 기록을 모두 망라하여 네트워크를 구성하는 데에는 어마어마한 양의 빅데이터를 다루어야 하며 데이터 제공자 측면에서도 총 데이터를 모두 공개하는데 대한 부담이 따를 수 있다. 본 장에서는 네트워크의 크기를 감소시키는 다운 샘플링을 진행하며 위상적 변화를 관찰해 보고자 한다. 또한 이것

은 네트워크 분석 이후 추가적인 문제 해결 또는 의사결정이 필요한 상황(예를 들어, 마케팅 측면에서 네트워크 상에 제품에 대한 이미지가 빠르게 전파되도록 소수의 표적 집단을 선택하는 문제, 인플루엔자 또는 스마트폰내의 악성 바이러스 등의 전파를 신속히 탐지하여 피해를 막기 위해 모니터링하고 있어야 하는 유효 집단을 구하는 문제)에서 본래의 대형 네트워크를 다룰 수 없을 때 효과적인 해법을 제시해 주는 한 가지 방법이 될 수도 있다. 본 연구에서는 Seidman(1983)이 정의한 *k*-core의 개념을 이용한다. *k*-core는 전체 네트워크에서 노드와 아크를 일부 선택하여 해당하는 부분 네트워크의 모든 노드가 Degree *k* 이상이 되는 가장 큰 네트워크를 의미한다. 전체 네트워크의 일부를 선택함으로서 다운 샘플링을 진행하는 방법에는 여러 가지 방법이 있을 수 있다. 예를 들어, 최소 걸침 나무(Minimum Spanning Tree)를 구하는 방법을 적용한다면 노드 수 -1 만큼의 아크만 선택된 후 나머지는 제거되는 방식으로 다운 샘플링을 진행할 수도 있겠지만, 이 경우는 네트워크의 위상적 특징이 매우 다르게 변할 수 있으므로 기존 네트워크의 위상적 변화를 최소화하면서 다운 샘플링 할 수 있는 방법을 적용하기 위하여 core 개념을 바탕으로 한다. 또한 제 4장에서 다룰 Scale-free 특성을 분석할 때 노드의 Degree 분포를 이용하므로 최소 걸침 나무 등의 방법은 적절하지 않을 수 있다.

지금껏 구한 통화 기록 기반의 컨택 네트워크는 노드당 평균 Degree가 약 5이며, 12개월 동안의 총 누적 네트워크에서도 13으로 그다지 높지 않다. 이는 현재 개개인의 휴대폰에 저장되어 있는 지인의 수를 감안하였을 때 매우 낮은 수치라 할 수 있을 것이며, 모집단 전체를 고려한 컨택 네트워크에 가까운 위상적 특징을 가진다고 보기 어려울 것이다. 따라서, 다운 샘플링의 방법으로 <Table 4>에서 최종적으로 채택한 컨택 네트워크를 최초의 네트워크로 고려한 후, Degree가 가장 낮은 노드순으로 노드를 네트워크에서 제거하도록 한다. 즉, 최초 네트워크에서 아크의 수가 오직 하나인 노드를 모두 제거한다. 이러한 노드들의 제거는 또 다른 노드를 제거 대상이 되는 노드로 만들 수 있다. 지속적으로 이러한 노드들을 제거한 후 더 이상 Degree가 1인 노드가 없을 때까지 진행한다. 동일한 방식으로 이번에는 Degree가 2인 노드들을 제거하고 반복적으로 시행한 후 모든 노드가 Degree 3 이상을 가질 때까지 진행한다. 만약 Degree가 최소 10 이상인 노드만 남도록 제거한다면 반복적인 노드 제거 후 최종적으로 남은 네트워크는 모든 노드가 Degree 10 이상이 될 것이다. 그렇게 얻은 네트워크는 연결되어 있지 않을 수도 있으므로 마지막으로 가장 큰 Component를 구하여 채택하도록 한다. <Table 4>에서 구한 다른 누적 기간의 5개의 컨택 네트워크에 대하여 아크 수 10 이상, 15 이상, 그리고 20 이상인 네트워크를 동일한 다운 샘플링 방법을 통하여 구해 본 결과 <Table 5>와 같이 나타났다.

<Table 5>에서 보여주는 결과는 다운 샘플링이 진행될수록 노드의 수가 급격히 감소함을 알 수 있다. 특히 12개월간 통화 기록을 누적한 네트워크 G_{12}는 최초 2,300만 개의 노드를 가지고 있었으나 각 노드의 Degree가 최소 10인 네트워크로 다운 샘플링했을 때 580만 개의 노드가 남아 약 75%의 노드가 제거되었고 최소 15인 네트워크로 변환시에는 약 90%의 노드가 사라지고 마지막으로 최소 20인 경우에는 최초 네트워크의 약 1% 만이 남아 있게 되었다. 그러나 평균 Degree는 13.0에서 50.9까지 증가됨을 알 수 있다. 이러한 급격한 노드 수의 감소는 높은 수치의 Degree를 요구할수록 해당 Degree를 만족시키지 못하는 노드가 더욱 많이 발생하게 되어 네트워크에서 제거되며 이러한 노드들의 제거는 또 다른 노드들을 연쇄적으로

Table 5. Connected contact networks after down-sampling

G_r	min degree	Number of nodes	Number of arcs	d_{avg}	$d_{\max}$
G_1	10	909	9,581	21.0	134
	15	92	1,257	27.3	69
	20	56	752	26.8	60
G_3	10	92,114	870,566	18.9	538
	15	3,217	55,900	34.7	452
	20	1,036	22,479	43.4	361
G_6	10	2,732,757	30,647,776	22.4	6,013
	15	18,179	335,825	36.9	906
	20	5,637	140,890	49.9	799
G_9	10	4,494,438	57,083,008	25.4	9,777
	15	1,410,772	21,543,029	30.5	6,209
	20	12,409	324,567	52.3	1,083
G_{12}	10	5,857,785	80,670,365	27.5	13,422
	15	2,529,342	42,085,201	33.2	10,200
	20	24,167	615,209	50.9	3,266

제거하기 때문에 나타나는 것으로 보인다. 다음 장에서 나타나겠지만, 본 연구에서 진행한 다운 샘플링 방식은 네트워크의 Scale-free 특성을 보존함을 보여주며 이는 또한 제안한 다운 샘플링의 필요성을 더욱 강조하는 부분이라 할 수 있다. 다른 한편으로 제안한 다운 샘플링은 그런 특성을 보존하기 위하여 Degree를 증가시키며 진행할수록 더욱 급격한 노드의 감소를 야기시킨다고 볼 수 있으며 이는 Scale-free 특성을 갖는 네트워크에서 일반적으로 나타나는 현상이라고 할 수 있을 것이다.

4. 컨택 네트워크의 Scale-free 특징

지금까지 Degree가 작은 노드들부터 순차적으로 제거하는 형태의 다운 샘플링을 통하여 효과적으로 네트워크의 크기를 줄일 수 있었다. 본 장에서는 이렇게 줄어든 네트워크가 다운 샘플링을 진행하기 전의 특징을 유지하고 있는지를 살펴보고자 한다.

네트워크의 위상적 특징을 설명하고 비교해 보기 위해서 여러 다양한 방법이나 측정치(Measure)를 적용해 볼 수 있다. 특히 소셜 네트워크와 같은 대형의 네트워크에 대한 특징을 알아보기 위한 방법은 어느 하나의 측정치만으로 설명하기 어렵다고 할 수 있을 것이다. 여러 다양한 방법들 중에서 네트워크를 분석하는데 매우 두드러지게 사용되고 있는 대표적인 측정치는 Barbasi and Albert(1999)가 정의한 Scale-free 네트워크이다. Scale-free의 특성을 가지는 네트워크인지 판별하는 방법은 다양한 측정치 중 Degree의 분포를 확인하는 것이다. Barbasi and Albert(1999)은 Scale-free 네트워크는 Degree의 분포가 거듭제곱 법칙(Power law)의 형태를 따르는 네트워크라고 정의하였다. 즉, 하나의 노드가 k개의 이웃 노드를 가지고 있을 확률을 $P(k)$라고 정의할 때 $P(k)$가 점근적으로 $\alpha k^{-\gamma}$의 값을 가지는 것을 의미하며, 여기서 α는 표준화 상수이며 γ는 파라미터 값이다. Scale-free 개념은 많은 소셜 네트워크의 위상적 특징을 설명하도록 제안되었으며, 실제로 Kempe *et al.*(2003)의 컨택 네트워크 및 Adamic and Huberman(2000)의 World Wide Web에서도 Degree를 바탕으로 한 Scale-free 개념이 소셜 네트워크의 위상적 특징을 잘 대변하였다. 따라서, 먼저 휴대폰 사용자간의 전화번호를 토대로 다운 샘플링을 통해 얻어진 네트워크의 Degree 분포가 거듭제곱 법칙을 따르는지를 살펴본다. 먼저 각 네트워크별로 Degree 분포를 그래프상에 그려보면 <Figure 1>과 같이 나타난다.

<Figure 1>에서 보여지는 그래프의 형태가 거듭제곱의 형태와 유사하게 나타나므로 자연로그함수를 취한 후 선형 회귀분석을 이용하여 다음과 같이 γ값을 측정한다.

$$\ln P(k) = \ln\alpha - \gamma \ln k. \tag{1}$$

<Figure 2>는 해당하는 네트워크에 대하여 log-log 형태의 그래프를 도식한 결과이며, <Table 6>은 선형 회귀분석 결과를 보여준다. 측정된 거듭제곱 파라미터 γ의 95% 신뢰구간과 함께 회귀분석 결과 적합도를 산출하는 R^2의 값을 보여준다. 선형 회귀분석 결과에서 나타나듯 거듭제곱 분포(Power-law Distribution)가 일반적으로 높은 적합도를 보이고 있음을 알 수 있고, R^2의 값이 전반적으로 1에 가까움에서도 확인될 수 있다. 이것은 해당 네트워크가 Scale-free 특성을 가지고 있음을 보여주는 것으로 본 연구에서 구성한 소셜 컨택 네트워크가

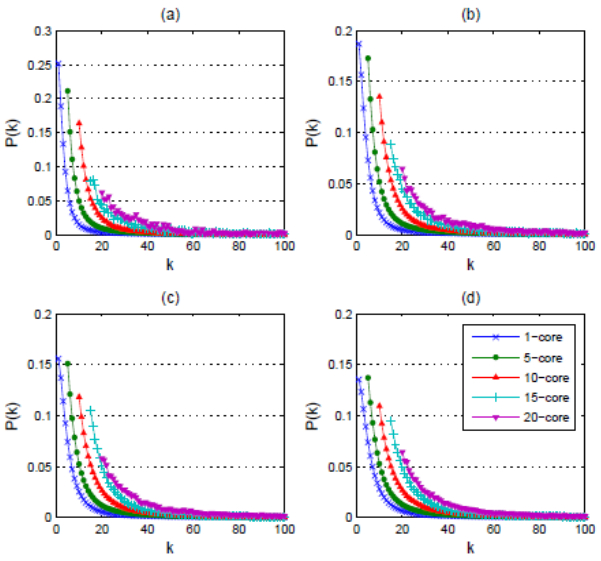

Figure 1. Degree Distributions after obtaining the minimum degree nodes as "core" for (a) G_3, (b) G_6, (c) G_9, (d) G_{12}

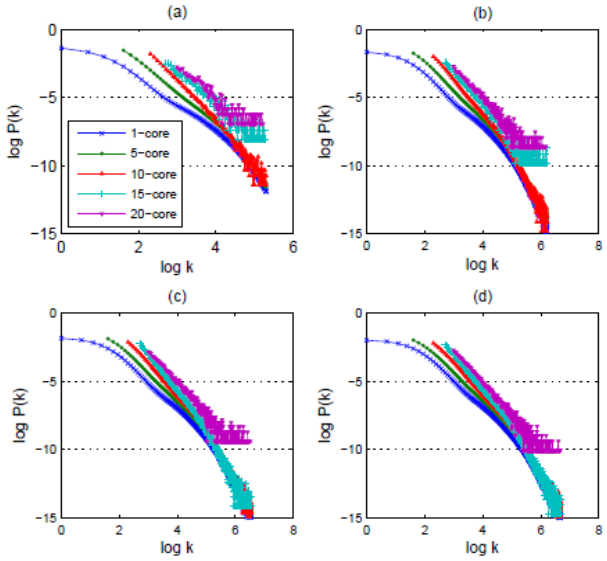

Figure 2. The plot for $(\ln k, \ln P(k))$ after obtaining the minimum degree nodes as "core" for (a) G_3, (b) G_6, (c) G_9, (d) G_{12}

Scale-free 네트워크임을 설명한다. 또한 다운 샘플링을 통하여 얻어진 사이즈가 작은 네트워크도 원래의 네트워크의 특징을 비교적 잘 보존하고 있음을 나타내기도 한다. 추가적으로 Barbasi and Albert(1999)가 제시한 Scale-free 네트워크에서 거듭제곱 파라미터 γ는 통상 2에서 4사이의 값을 일반적으로 갖는다는 점을 근거로 본 컨택 네트워크 역시 근사적으로 2의 값을 가지는 것으로 보아 해당 네트워크가 Scale-free 네트워크의 특징을 가지고 있음을 보여준다.

한 가지 주목할 점은 최소 Degree를 증가시킬수록 다운 샘플링한 네트워크의 γ값은 일정 수준 증가하다가 감소함을 알 수 있다. 같은 현상으로, R^2의 값도 증가하다가 감소함을 알 수 있는데 약 5～15사이의 최소 Degree를 가지는 네트워크 수준으로 다운 샘플링했을 때, Scale-free 네트워크에 가장 근접한 특징을 가짐을 의미하며 <Figure 2>에서도 나타나듯이 최소 Degree가 1 또는 20일 때보다 5～15일 때 더욱 직선에 가까운 형태의 log-log 그래프가 됨을 보여준다. 물론 모든 소셜 네트워크에서 비슷한 현상이 나타날 지에 대해서는 추가적인 연구가 필요하겠지만, 본 연구에서 고려한 컨택 네트워크는 최소 Degree가 5인 수준까지 다운 샘플링을 할 때 가장 Scale-free 네트워크에 근접한 네트워크가 됨을 알 수 있다. 또한 이것은 어느 정도까지 다운 샘플링을 하는 것이 가장 좋은 것인가에 대한 하나의 척도가 될 수도 있을 것이다.

Table 6. Regression analysis results of contact networks for minimum degree nodes

Network	minimum degree	γ(± error)	R^2
G_3	1	2.36(±0.050)	0.958
	5	2.82(±0.038)	0.978
	10	3.30(±0.080)	0.970
	15	2.33(±0.121)	0.917
	20	1.95(±0.173)	0.815
G_6	1	1.81(±0.042)	0.984
	5	2.14(±0.014)	0.999
	10	2.55(±0.019)	0.998
	15	2.45(±0.066)	0.980
	20	2.46(±0.114)	0.941
G_9	1	1.75(±0.046)	0.981
	5	2.06(±0.011)	0.999
	10	2.36(±0.009)	0.999
	15	2.72(±0.016)	0.999
	20	2.29(±0.061)	0.980
G_{12}	1	1.72(±0.049)	0.977
	5	2.01(±0.013)	0.999
	10	2.28(±0.008)	0.999
	15	2.54(±0.009)	0.999
	20	2.36(±0.040)	0.988

5. 소셜 네트워크의 위상적 특징 분석 대안

지금까지 실제 통화 기록을 바탕으로 하여 휴대폰 사용자간의 컨택 네트워크를 구성하고, 통화 기록을 누적함으로서 사용자간의 연결 상태를 더욱 면밀히 관찰한 대형 소셜 네트워크를 구성하였다. 다운 샘플링을 통하여 실험 및 분석이 가능한 형태로 축소하는 방법을 제시하였고 해당 네트워크의 거듭제곱 분포에 대한 회귀분석 결과 Degree 분포가 전반적으로 Scale-free 네트워크로서의 특징에 부합함을 알 수 있었으며, 다운 샘플링을 통한 네트워크 역시 유사한 특징을 보존함을 통해 제안된 축소 방법이 비교적 효과적인 방법임을 보여주었다.

하지만 빅데이터가 사회적 뿐만 아니라 학문적으로 많은 관심과 각광을 받고 있는 지금, 빅데이터로 대표되는 소셜 네트워크에 대한 분석 및 그 위상적 특징을 관찰하고자 할 때 현재까지는 Barbasi and Albert(1999)가 제안한 Scale-free 네트워크 여부임을 판별하는데 너무 의존하는 경향이 있다. Alderson (2008)은 Degree 분포를 통한 네트워크 분석이 소셜 네트워크와 같은 복잡 네트워크(Complex Network)의 특징들을 모두 반영하기에는 충분하지 않음을 지적하였다. 만약 어떠한 실제 소셜 네트워크의 다른 추가적인 정보나 지식이 없이 Degree 분포에 대해서만 알고 있다고 가정하고, 동일한 Degree 분포를 따르는 네트워크를 랜덤으로 생성한다고 하자. 같은 분포를 가지지만 매우 다른 특징들을 가지는 많은 별개의 네트워크들을 생성할 수 있다면 Degree 분포 기반의 네트워크 분석은 위상적 특징을 대변하기에 충분하지 않은 방법일지도 모른다. 또한 Scale-free 네트워크는 노드의 연결정도를 대표할 수 있는 척도가 없이 Degree 분포만으로 정의하는 네트워크라는 점과 파라미터 γ의 값도 실험적으로 얻어진 값이므로 이론적인 수치를 제시하지 못한다. 이러한 점들을 바탕으로 최근에 보다 다양한 방식으로 추가적인 특징을 파악하기 위한 측정치를 제공하는 경우도 있다. Alvarez-Hamelin *et al.*(2012)은 인터넷 네트워크의 Scale-free 특성만으로 파악할 수 없는 구조적인 특징을 분석하기 위하여 Degree 분포에 추가적으로 Neighbor's Average Degree Distribution(NADD) 및 Clustering Spectrum(CS)과 같은 분석기법을 적용하기도 하였다. NADD는 해당 노드의 Degree 분포에 추가하여 그 노드의 이웃 노드들에 대한 평균적인 Degree 분포를 확인하는 것이며, 이웃 노드간의 모든 가능한 연결수에 대한 실제 연결수의 비율을 Clustering Coefficient(CC)라고 정의한 후 Degree k를 가지는 노드들의 평균적인 CC를 CS라 하여 그 값들을 측정하는 것이다.

뿐만 아니라, 네트워크의 흐름(Flow)에 대하여 병목(Bottleneck)을 검사하는 기법을 적용하는 것도 Scale-free 특성에 추가하여 전반적인 네트워크의 위상적 특징을 확인해 볼 수 있는 좋은 측정치가 될 수 있다. Conductance(a.k.a. Cheeger constant), Vertex Expansion, 그리고 Expander Ratio(a.k.a. Isoperime-

tric Number)가 그 대표적인 측정 도구들이며, 이러한 개념들은 원래 마코프 체인(Markov Chain)이 얼마나 빨리 Stationary Distribution에 도달하는지를 측정하기 위해 소개된 것이다(Chung, 1997). 또한 이러한 기법들은 모두 각각의 다른 목적값을 취하기 위해 네트워크를 두 부분으로 나누는(Bipartitioning) 문제들과 많은 연관성을 가지고 있다. 하지만, 일반적으로 네트워크에서 Conductance, Vertex Expansion 및 Expander Ratio를 구하는 문제는 NP-hard로 알려져 있어서(Hochbaum, 2010), 정확한 값을 구하기가 매우 어려우며 특히 소셜 네트워크와 같은 대형 네트워크에서는 더더욱 그렇다. Chung(1997)은 그 대안으로 Laplacian 매트릭스의 Eigenvalue를 이용한 근사기법을 제안하였으나 Laplacian 매트릭스의 크기가 커짐에 따른 Eingenvalue 자체를 구하기가 어려운 점도 고려되어야 한다. 최근에 Hochbaum(2010)이 이 문제를 해결하기 위한 방법으로 Spectral Technique을 적용하긴 하였으나 이것 또한 NP-hard로서 근사치를 통한 접근으로 제한되었다. 이처럼 Scale-free 특징 이외에도 추가적인 분석 방법들이 제시되고 있으나 계산의 복잡성과 네트워크 자체의 방대함 등으로 인해 활발히 진행되지 못하고 있는 실정이다.

6. 결론 및 향후 연구방향

본 연구에서는 대형 컨택 네트워크를 실 데이터를 이용하여 구현하였으며, 통화기록을 바탕으로 노드와 아크를 구성하였으며, 제시한 컨택 네트워크는 Scale-free 특성을 갖는 Scale-free 네트워크임을 보여주었다. 컨택 네트워크의 분석을 위하여 Degree가 가장 작은 노드부터 제외하는 다운 샘플링을 제안하였고, 실험 결과를 통해 제안한 다운 샘플링은 Scale-free 특성을 매우 잘 보존하면서 네트워크의 크기를 효과적으로 축소시킬 수 있는 방법이라 할 수 있다. 이는 랜덤 샘플링으로 보존하지 못하는 특성을 잘 보존해 주는 샘플링 기법이다. 따라서 본 연구에서 제안한 다운 샘플링은 대형 네트워크의 크기도 효과적으로 축소시킬 뿐만 아니라 그 위상적 특성도 보존해 주기 때문에 그 적용 가치가 매우 높다고 할 수 있다. 본 연구에서는 분석 결과만을 제시하였지만, 분석에 추가하여 네트워크상 최적화를 추구하는 문제를 고려할 경우에도 다운 샘플링은 풀 수 있는 네트워크의 크기로 축소할 수 있는 매우 유용한 기법 중 하나가 될 것이다.

네트워크 과학(Network Science)으로 통용되는 소셜 네트워크를 포함한 여러 복잡 네트워크에 대한 분석 및 위상적 특성 파악 분야는 여러 수학적, 통계적 기반의 분석적인 연구가 진행되어 왔지만, 그 응용에 대한 연구는 그리 많지 않은 편이다. 네트워크에 대한 연구는 경영과학 및 산업공학에서도 매우 전통적인 분야 중 하나임이 분명하고 네트워크상 제한된 자원 및 이용 가능한 경로 등을 바탕으로 흐름을 통제하고 비용-효과 측면에서의 최적화에 대한 연구는 많은 공헌을 해 온 것도 사실이다. 따라서 소셜 네트워크 자체에 대한 분석도 중요하지만 경영과학/산업공학의 여러 전문가들이 소셜 네트워크의 다양한 분석 기법도 제안하고 또한 분석 이후에 어떻게 운용하고 통제할 것인가에 대한 의사결정의 기틀을 마련해 주는 역할도 매우 중요할 것이다. 본 연구에서 진행된 컨택 네트워크의 구성 및 분석 등을 바탕으로 네트워크의 추가적인 특성 파악을 위한 연구, 더 나아가 네트워크상 의사결정을 위한 최적화에 대한 연구 등은 향후 연구과제로 남겨 놓고자 한다.

참고문헌

Adamic, L. A. and Huberman, B. A. (2000), Power-law distribution of the World Wide Web, *Science*, **287**, 2115.

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1994), *Network Flows : Theory, Algorithms, and Applications*, Prentice Hall, New Jersey, USA.

Alderson, D. L. (2008), Catching the “network science” bug : Insight and opportunity for the operations researcher, *Operations Research*, **56**, 1047-1065.

Alvarez-Hamelin, J. I., Dall’Asta, L., Barrat, A., and Vespignani, A. (2008), K-core decomposition of internet graphs : Hierarchies, self-similarity and measurement biases, *Network and Heterogeneous Media*, **3**, 371-393.

Balthrop, J., Forrest, S., Newman, M. E. J., and Williamson, M. M. (2004), Technological networks and the spread of computer viruses, *Science*, **304**, 527-529.

Barabasi, A. L. and Albert, R. (1999), Emergence of scaling in random networks, *Science*, **286**, 509-512.

Chen, W., Wang, Y., and Yang, S. (2009), Efficient influence maximization in social networks, *Proc. 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Paris, France.

Chung, F. R. K. (1997), *Spectral Graph Theory*, American Mathematical Society, Rhode Island, USA.

Dimitrov, N. B. and Meyers, L. A. (2010), Mathematical approaches to infectious disease prediction and control, *Tutorials in Operations Research*, *INFORMS*, 1-25.

Fleizach, C., Liljenstam, M., Johansson, P., Voelker, G. M. and Mehas, A. (2007), Can you infect me now? Malware propagation in mobile phone networks, *Proc. 2007 ACM Workshop on Recurring Malcode*.

Hochbaum, D. (2010), Replacing spectral techniques for expander ratio, normalized cut and conductance by combinatorial flow algorithms, *arXiv*, 1010.4535.

Kempe, D., Kleinberg, J., and Tardos, E. (2003), Maximizing the spread of influence through a social network, *Proc. 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Washington, DC, USA.

Lee, H.-J. (2013), Breakthrough with SNS marketing for the recession of the advertisement market, The Kyunghyang Newspaper, July 19, Korea.

Lim, S.-S. (2012), The US presidential election campaign using big data with details, Yonhap News, October 22, Korea.

Seidman, S. B. (1983), Network structure and minimum degree, *Social Networks*, **5**, 269-287.