Journal of the Korean Society of Agricultural Engineers
Vol. 56, No. 1, pp. 11~20, January, 2014
DOI:http://dx.doi.org/10.5389/KSAE.2014.56.1.011

ISSN 1738-3692
eISSN 2093-7709

동적 확률 재규격화를 이용한 네트워크 연쇄 관계 해석

Analysis of Network Chain using Dynamic Convolution Model

이형진* · 김태곤* · 이정재** · 서교***,†
Lee, Hyungjin · Kim, Taegon · Lee, JeongJae · Suh, Kyo

ABSTRACT

Many classification studies for the community of densely-connected nodes are limited to the comprehensive analysis for detecting the communities in probabilistic networks with nodes and edge of the probabilistic distribution because of the difficulties of the probabilistic operation. This study aims to use convolution method for operating nodes and edge of probabilistic distribution. For the probabilistic hierarchy network with nodes and edges of the probabilistic distribution, the model of this study detects the communities of nodes to make the new probabilistic distribution with two distribution. The results of our model was verified through comparing with Monte-carlo Simulation and other community-detecting methods.

Keywords: Clustering; convolution; probability; network theory

I. 서 론

네트워크에서 공통된 특성을 가진 절점들이 군집되는 것은 자연스러운 일이다. 네트워크로 해석되는 자연의 다양한 문제들의 주체인 절점들은 특정 집단에 속하려는 성질을 가지고 있으며 다른 절점들과 관계, 전체 구조에서 위치 등의 자료를 이용해서 그 성질을 표현한다. 어떤 절점이 전체 네트워크에서 어떤 소집단에 포함되어있는지에 대한 문제에 대한 해석은 전체 네트워크의 프로세스와 구조에 대한 이해를 도울 수 있을 뿐만 아니라 특정 절점이 속한 집단의 다른 절점을 통해서 그 절점의 기능적인 특성을 예측할 수 있다. 절점은 자신과 밀접하게 연결된 절점들과 소집단을 형성하고 그 소집단의 정체성을 대변한다. 소집단은 자신만의 프로세스, 구조, 방향성 등을 가지는데, 이 특성들은 소집단 내의 절점들의 특성들을 포괄적으로 표현한 것이다. 유사하게 소집단들이 모이면 전체 네트워크를 구성하기 때문에 전체 네트워크 해석은 소집단들의 정체성에 대한 분석에서 시작한다고 볼 수 있다. 군집화 방법을 적용하여 네트워크에서 소집단들을 구분하는 것은 네트워크를 해석하는데 주요한 문제이다.

네트워크 구조에서 소집단을 찾아내는 방법은 그래프 이론에서 큰 그래프를 작은 단위의 그래프로 분할하는 문제와 동일하다. 이 방법은 구조적 문제에서 하중 분할이나 네트워크 설계의 회로 분할 등에 다양하게 적용된다 (Pothen et al., 1990; Roy et al., 2003; Zhou, 2003). 전체 구조의 특성이나 절점들을 연결하는 연결선의 특성에 따라 다양한 알고리즘을 이용하여 소집단을 구하는 방법론에 대한 연구가 활발히 진행되고 있다 (Girvan and Newman, 2002; Gudkov J. et al., 2002; Guimera et al., 2003; Leskovec et al., 2007, 2010; Marcos et al., 2008; Reichardt and Bornoholdt 2004, 2006; Wu and Huberman, 2004). 그래프 이론을 기초로 네트워크 구조에서 소집단을 찾아내는 방법은 매우 유용하지만 절점과 연결선의 특성이 결정론적인 수치를 가지는 네트워크 문제에서 적용할 때만 유효했다. 절점과 연결선을 대표하는 수치들이 확률론적 범위를 가질 때 복잡계 네트워크를 기반으로 한 네트워크 구조 및 주요한 절점들의 위상에 대한 해석을 시도한 연구 (Fefferman, 1970; Guimera et al., 2003; Leskovec et al., 2010; Peter et al., 2010)나 절점을 확률변수로 하고 연결된 두 절점의 확률적 의존성을 표현하여 군집화를 시도하는 연구 (Dechter and Pearl, 1989; Friedman et al., 1997; Friedman, 2004)가 있긴 하지만 연결선에 관한 확률 정보와 함께 포괄적으로 분석하는데 한계가 있다.

* 서울대학교 농업생명과학대학 생태 조경·지역시스템공학부
** 서울대학교 농업생명과학대학 조경·지역시스템공학부, 농업생명과학연구원
*** 서울대학교 농업생명과학대학 조경·지역시스템공학부, 농업생명과학연구원, 그린바이오과학기술 연구원
† Corresponding author Tel.: +82-2-880-4591
Fax: +82-2-873-2087
E-mail: kyosuh@snu.ac.kr
2013년 11월 11일 투고
2013년 12월 10일 심사완료
2013년 12월 10일 게재확정

네트워크로 모델링 되는 다양한 현상들은 비선형적이며 복잡한 시스템으로 정확히 예측하는 것이 불가능하다. 네트워크 모델링에는 물리적 현상의 수치적인 반영에서 뿐만 아니라 초기 조건에 존재하는 한계에 의한 어려움이 존재한다. 임의의 절점에서 발생한 작은 오차는 직접 연결된 절점들에 전달되어 전체 구조까지 진행되면서 점차 증폭되며 예측하지 못한 문제를 유발하기도 한다. 문제를 예측할 수는 없지만 예측이 가능한 범위를 결정하기 위해서 확률적 해석은 필연적이다. 불확실성이 높은 현상들에 대해서 정확하게 네트워크로 모델링하고 예측하는 것은 근본적으로 불가능하다. 하지만 결정론적 방법에서 생기는 오류를 지적할 수 있는 등 많은 상황에 대한 해석의 시도가 가능하다는 점에서 확률 기반의 네트워크 해석은 의의가 있다.

확률론적 네트워크 해석의 보다 정확한 모의를 위해서 절점의 특성을 고려할 필요가 있다. 네트워크 이론의 기본 근간이 되는 그래프 이론에서는 절점은 연결선에 대한 정보만을 가지고 있으며 연결선은 시작 절점과 도착 절점 및 전달하는 값에 대한 정보를 가지고 있다 (Newman and Girvan, 2003). 주로 네트워크 구조는 한 가지 물질로 지배되는 단일 네트워크를 표현하고 절점을 통해서 운반되는 그 유일한 물질의 크기에 대한 정보는 연결선이 가지고 있으며 절점은 방향성만 결정한다. 하지만 절점은 운반되는 그 물질에 대한 저항적 특성을 가질 수 있다. 그리고 절점이 가지는 저항과 연결선이 가지는 값이 각각 확률적 분포를 가지고 있다면 확률적 연산에 이용되는 컨볼루션을 이용하여 이 둘을 연산하고 새로운 하나의 확률적 분포로 재규격화 할 수 있다.

최근 구제역 및 조류인플루엔자와 같은 매우 강한 전염성을 가진 가축질병이 급격히 증가하고 있다. 가축전염병은 농가의 경제적인 피해뿐만 아니라 인간의 생명에게도 직접적인 위협을 주고 있을 정도로 심각한 문제가 되고 있다. 또한 질병 바이러스의 변이도 다양해지면서 백신을 통한 근본적인 방제 대책이 어려운 실정이다. 발병 이후 발병 농가의 일정 반경내 농가를 모두 폐쇄하는 대책을 수행하기는 하나 바이러스의 전달 매개체에 대한 분석이나 각 농가별 특성을 고려하지 않는 대책이기 때문에 효율적으로 확산을 차단하기에는 미흡하다. 네트워크에서 군집화는 매개체를 통한 확산 경로 분석에 매우 유용하게 적용될 수 있다. 절점 간의 연결성을 분석하고 상호 연관도가 높은 절점을 군집화하면 확산 경로를 예측할 수 있고 효율적인 방제 대책으로 활용할 수 있다. 그리고 가축질병을 전파하는 농가 관련 업체 및 관계자의 방문 빈도, 농가의 위생상황 및 야생조수류 출현빈도 등과 같은 환경적 요인을 연결선과 절점에 확률적으로 반영할 수 있다면 가축질병에 대한 대응책의 활용도와 해석의 정확도가 향상될 수 있다.

본 연구에서는 절점들의 관계 정보들을 기반으로 네트워크 군집화에 대해서 해석하되 그 관계가 확률론적으로 구성되어 있을 때 수학 연산자인 컨볼루션을 활용하는 방법론을 제시하였고 이를 가상의 확산 네트워크 구조에 반영하고 기존의 방법과 비교 검증하였다. 이 방법론은 절점과 연결선이 가지는 확률 분포 등의 정보를 포괄적으로 해석하여 그 발현가능성에 따라서 절점들의 특성을 구분할 수 있고 확률 기반 네트워크를 군집화 할 수 있을 것이며 농가의 가축질병 등의 확산 대응에 유용하게 활용할 수 있을 것으로 기대된다.

II. 연구방법

1. 동적 확률 재규격화

네트워크의 연결선이 가지는 값은 출발 절점에서 주어지는 값으로만 결정되어지지 않는다 (Newman and Girvan, 2003). 연결선의 값이 절점의 특징적인 저항력에 의해서 영향을 받기 때문에 절점으로 들어오는 외력과 절점이 가지고 있는 고유의 저항력, 즉 두 사상을 포괄적으로 해석할 수 있는 연산이 필요하다. 임의의 입력신호에 따라 출력이 결정되는 시스템에서 두 개의 독립된 신호에 관한 확률을 연산하기 위한 방법으로 컨볼루션(convolution)이라는 수학적 연산자를 이용한다 (O'Neil, 1963). 컨볼루션 기법을 통하여 두 개의 독립된 확률적 사상을 통합하여 새로운 사상으로 재해석하고 하나의 확률 사상으로 재규격화 할 수 있다. 그 기작은 입력값을 y축으로 대칭이동한 후에 단계별로 출력의 확률값을 곱해나가는 방식으로 연산을 수행해 나가는 방법이다 (Ringer, 1971; Burt and Garman, 1971). 시스템 입력신호의 확률분포를 $I(t)$라 하면 출력에 대한 확률적 연산에 의한 값은 특정시점 t에서의 입력신호를 기준으로 그보다 작은 $-\infty$까지 입력신호 값의 합과 출력신호, $O(t)$의 곱으로 식 (1)과 같다.

$$y(x)=\left(\int_{-\infty}^{t} I(\tau)d\tau\right)\times O(t) \quad (1)$$

$I(t)$: Input Signal
$O(t)$: Output Signal

본 연구에서는 절점에서 주어지는 외력에 대한 저항력의 두 확률적 사상을 합성하여 그 연결선의 값을 구하는데 시스템에서 얻어지는 값을 이전의 입력에 대한 출력으로 해석하는 컨볼루션 기법을 적용하고자 한다. Fig. 1과 같이 기존의 시스템의 입력과 출력에 대한 확률분포가 $y(x)$라는 새로운 확률 분포로 주어진다면 네트워크 모델에서 절점으로 들어오는 연결선이 확률분포

를 가지는 외력 (Lord)을 가지고 있고 절점이 저항 (Resistance)을 가지고 있다면 컨볼루션 연산을 통해서 절점을 나가는 연결선은 새로운 값 ($W(x)$)를 가지게 된다.

네트워크에서 L은 절점으로 들어오는 연결선이 가지는 외력이고 R은 절점이 가지는 저항이다. 그리고 W는 절점에서 나가는 연결선이 가지는 외력을 의미한다. L, R의 확률분포를 정규분포로 가정할 때 평균값을 대푯값으로 이용하면 W의 평균값은 $L-R$과 같고 컨볼루션 해법은 식 (2)와 같다.

$$\begin{aligned}\int_{-\infty}^{\infty} W(x)dx &\Leftrightarrow \int_{-\infty}^{\infty} L(x)\times R(x)dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{x} L(\tau)d\tau R(x)dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{0} L(\tau+t)\,R(\tau)dtdx \\ &= \int_{-\infty}^{0} L(t)\otimes R(-t)dt\end{aligned} \tag{2}$$

L: Lord in link
R: Resistance in node
W: Weight in link

2. 동적 확률 재규격화를 이용한 군집화

가. 적용 범위

네트워크 구조의 군집화를 위해서는 각 절점들의 관계 정보인 연결선과 절점의 특성을 포괄적으로 고려하여 절점들을 대표할 수 있는 확률 분포를 구해야한다. 방법론을 적용하기 위해 본 연구에서 사용된 네트워크의 범위는 다음과 같다.

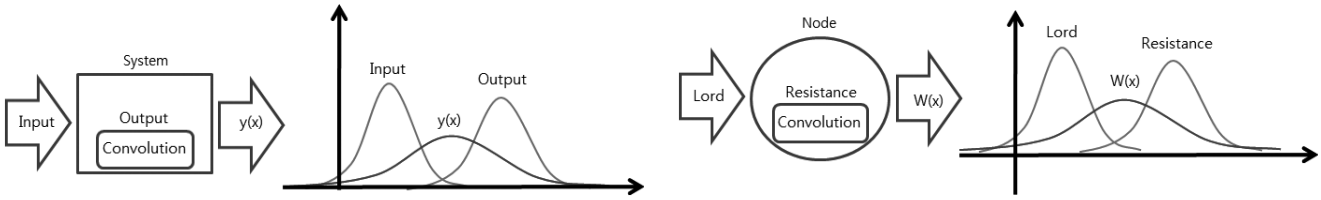
본 연구에서 적용할 네트워크의 형태는 계층형 네트워크 (Hierarchy network) 구조이다.

절점은 특징적인 R (저항)의 확률분포를 가진다.

연결선은 L (하중)의 확률분포를 가지며 방향성을 가진다.

R과 L은 서로 대응하며 컨볼루션 연산 후에 연결선은 W의 확률분포를 가진다.

Fig. 2는 본 연구에서 제안한 모델에서 적용 가능한 범위에 있는 네트워크 구조이다. 최상위 0번 절점에서 흐름이 시작되며 각 절점은 흐름에 저항을 가지고 있다. 최상위 0번 절점에서 물질의 흐름이 시작된다면 동적 확률 재규격화를 이용한 군집화 (Clustering with Probabilistic Convolution, CPC)를 이용해서 각 절점이 가지는 저항에 따라서 그 흐름을 아래 절점으로 전달시켜주는 것이 가능한지 여부를 각 절점별로 확률 값을 표현할



(a) System (b) Network

Fig. 1 Convolution in network

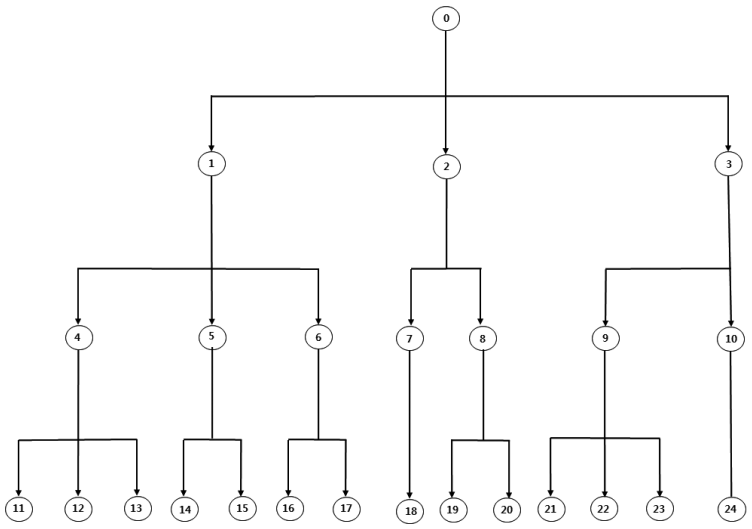


Fig. 2 Hierarchy Network

수 있고 그 값에 따라서 최상위 절점과의 연결 정도를 정량화할 수 있다.

나. CPC 모델의 개발

Fig. 3과 같이 i, j, k 절점이 순서대로 방향성을 가지고 이어져있는 경우 i 절점에서 j 절점에 전달하는 외력 L은 j 절점의 저항력 R과 연산되어 k 절점으로 전달되는 W가 산정된다.

W는 k 절점에서 다른 절점으로의 연결을 연산할 때 L로 치환될 수 있다. 해석은 단계별로 축차적으로 해석된다.

$$W = \max(L-R, 0) \tag{3}$$

식 (3)와 같이 연결선이 가지는 값은 절점에 작용하는 L에서 R을 뺀 값이며 음의 값이면 연결선이 발현되지 않은 것을 의미한다. 이 때 임의의 하중, x에 따라 절점에 작용하는 외력의 확률을 $L(x)$, 저항의 확률을 $R(x)$라 하고 $W(x)$라 하면,

$$\begin{aligned} W(x) &= W(L-R) \\ &= \left(\int_{-\infty}^{x} R(\tau) d\tau \right) \times L(x) \end{aligned} \tag{4}$$

식 (4)와 같다. $W(x)$는 하중 x가 나타날 수 있는 모든 경우를 합해야 한다. $L(x)$와 $R(x)$가 가질 수 있는 모든 확률적 경우를 합산할 경우 $W(x)$가 된다. 따라서 이를 전체적인 확률분포로 표현하기 위해 임의의 하중을 $+\infty$에서 $-\infty$까지 정리하면 연결선의 확률밀도함수는 다음과 같다.

$$PDF_{W(x)} = conv(PDF_{L(x)}, PDF_{R(x)}) \tag{5}$$

$$\begin{aligned} \int_{-\infty}^{\infty} L(x) \times R(x) dx &= \int_{-\infty}^{\infty} \int_{-\infty}^{x} L(\tau) d\tau R(x) dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{0} L(\tau + t)\, R(\tau) dt dx \\ &= \int_{-\infty}^{0} L(t) \otimes R(-t) dt \end{aligned}$$

임의의 절점에서의 네트워크의 임의의 절점이 가지는 수식으로 확장하면 식 (6)과 같다.

$$PDF_{W(x_n)} = conv(PDF_{L(x_{start})}, PDF_{R(x_n)}) \tag{6}$$

x_n: node n

x_{start}: start node lined with node n

x_{start}는 임의의 절점과 연결되어있는 시작점을 의미한다. 식 (6)의 연산이 시작절점에서부터 단계별로 이루어지기 때문에 전체 네트워크 구조에서 이루어지는 연산 과정은 식 (7)과 같다.

$$PDF_{W(x_k)} = conv(conv(PDF_{L(x_i)}, PDF_{R(x_j)}), PDF_{R(x_k)}) \tag{7}$$

PDF: Probability Density Function

$conv(a,b)$: convolution a and b

Fig. 4는 시작 절점에서 컨볼루션 연산의 반복된 수행이 어떻게 진행되는지에 대한 순서도를 나타낸 것이다.

복잡한 계층형 네트워크를 축차적으로 컨볼루션을 수행함에 의해서 각 절점이 가지는 연결선에 대한 확률적 분포를 얻을 수 있

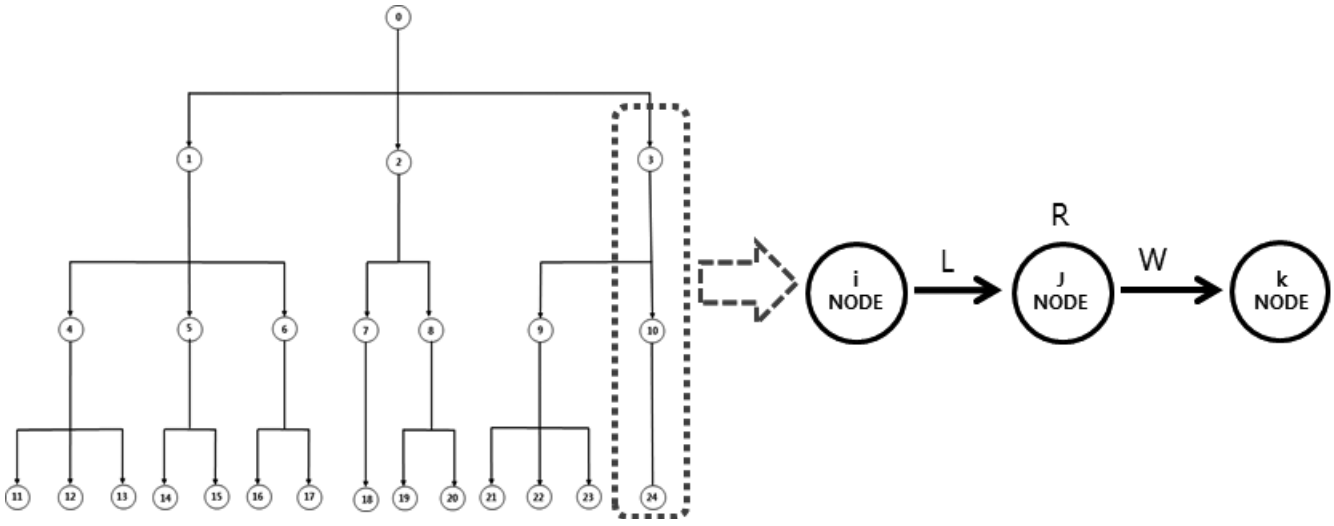


Fig. 3 Convolution in Hierarchy network

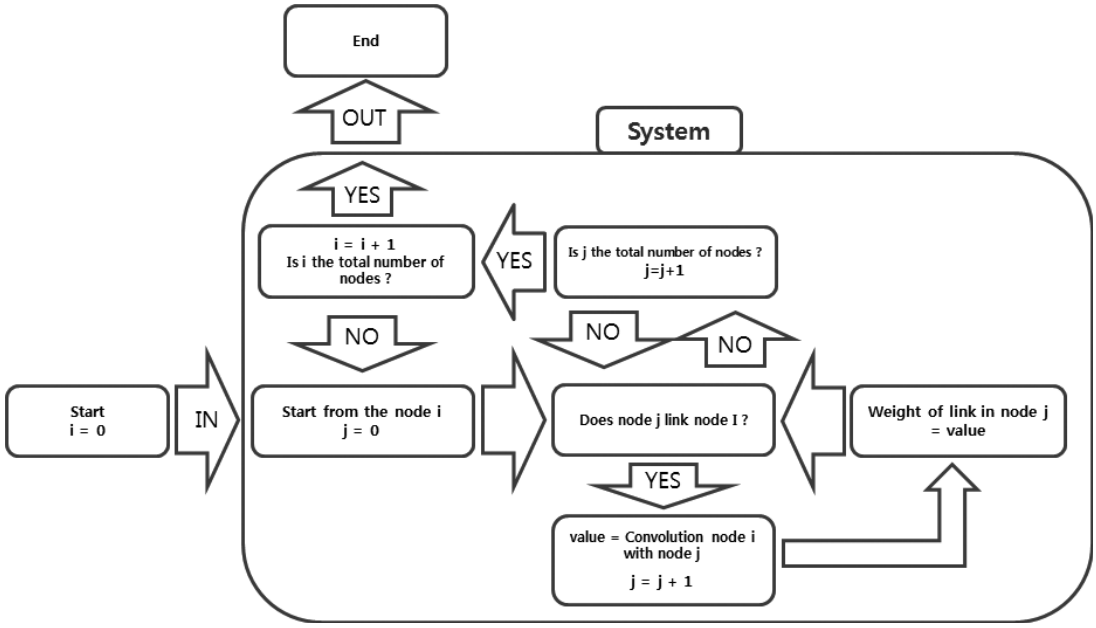


Fig. 4 System flow of CPC model

다. Fig. 3의 k 절점은 i 절점에서 시작된 흐름이 j 절점에서의 저항과 k 절점의 저항을 거치기 때문에 k 절점의 W 분포의 평균값은 j 절점의 W 분포의 평균값과 비교할 때 더 작다. 그리고 발현의 기준이 되는 W값이 0보다 작을 확률이 j 절점보다 작다. 각 절점마다 저항의 차이가 있겠지만 최상위 절점에서 하위 절점으로 단계가 내려갈수록 발현되지 않을 확률이 점점 커지게 된다. Fig. 4의 순서도에 따라 연산이 수행되면 절점이 가지는 연결선이 발현될 확률, $W_{positive}$을 각각 산정할 수 있고 이를 이용해서 최상위 절점과 밀접한 연관이 있는 절점들의 집단을 구분할 수 있다.

$$W_{positive} = \int_0^{\infty} PDF_{node} \tag{8}$$

절점의 W가 0보다 큰 확률을 식 (8)과 같이 쓸 수 있으며,

$$Group_{node_x} = Clustering\begin{pmatrix} node_1, node_2, node_3, node_4, \\ ..., node_n,\ W_{positive} > C.C. \end{pmatrix} \tag{9}$$

$C.C.$: Criteria of Clustering

식 (9)와 같이 군집화의 기준 (Clustering of Criteria, $C.C.$)이 마련되면 1부터 n까지 절점의 $W_{positive}$를 분석하여 x 절점과 가장 밀접하게 연관된, 그리고 특성이 유사한 절점들을 구분할 수 있다. 군집화의 기준은 절점들의 정보와 적용하고자 하는 문제에 따라 유연하게 설정할 수 있으며 그 기준에 따라 적합한 군집화 결과를 얻을 수 있다.

III. CPC 모델의 적용 및 고찰

1. 계층형 네트워크에 대한 CPC 모델의 적용

임의의 확산 네트워크 구조를 가정하여 컨볼루션을 이용한 군집화 방법을 적용하고 그 결과를 난수를 이용한 수치해법인 몬테카를로 방법과 비교, 검증하며 다른 군집화 방법과의 차이를 분석해보고자 한다. Fig. 5의 네트워크는 계층형이며 방향성을 가지는 절점 11개와 연결선 10개로 구성된 네트워크이다.

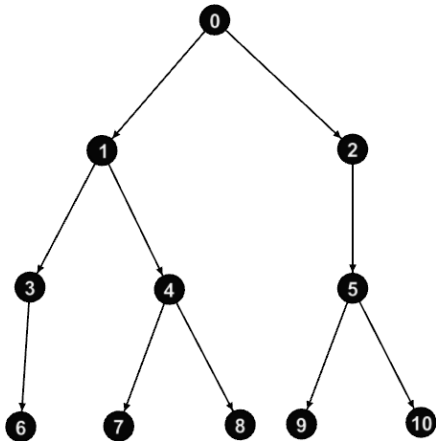


Fig. 5 Example of the hierarchy network

최상위 0번 절점에서 바이러스가 발현되어 확산이 시작되는 구조라면 분기점이 되는 0번 절점에서 1, 2번 절점, 1번 절점에서 3, 4번 절점, 4번 절점에서 7, 8번 절점, 5번 절점에서 9, 10번 절점으로 바이러스의 확산이 각각 갈라지는데 그 비는 각각 Table 1과 같다.

11개 절점의 바이러스에 대한 저항력 R의 확률 분포는 Table 2와 같고 바이러스를 전달하는 전달력 L의 확률 분포는 Table 1에서 가정한 비에 따르며 최초 0번에서 주어지는 연결선의 바이러스의 전달력은 10으로 하고 표준편차는 1로 하고 모든 분포는 정규분포 (Normal Distribution)의 형태를 가진다고 가정했다. 가축질병에 적용한다면 업체나 관계자의 방문 경로에 따라서 방향이 결정될 것이고 연결선 값의 평균 및 분산은 빈도에 따라서 결정된다. 농가의 위생상황이나 야생조수류 출현 빈도와 같은 환경적 요인은 절점의 평균과 분산을 결정할 수 있다.

Table 1 Distribution of the node

start node	0		1		4		5	
end node	1	2	3	4	7	8	9	10
portion	0.7	0.3	0.6	0.4	0.7	0.3	0.4	0.6

Table 2 Distribution of the Resistance in node

Node	Mean	Std.
0	2	1
1	2	$\sqrt{3}$
2	3	1
3	2	1
4	3	$\sqrt{3}$
5	2	1
6	3	$\sqrt{6}$
7	2	$\sqrt{3}$
8	3	1
9	2	$\sqrt{5}$
10	2	$\sqrt{2}$

가정된 자료를 이용하여 계산하면 1번 절점의 바이러스 감염확률은 다음과 같다.

$$PDF_{W_1(x)} = conv(\frac{7}{10} PDF_{L_0(x)}, PDF_{R_1(x)}) \tag{10}$$

이때 $L_0(x)$ 분포는 평균이 7 (0번에서 주어지는 하중 10에 0.7을 곱한 값), 표준편차는 1인 정규분포이다. 식 (10)에서 구한 값을 이용한 3번 절점의 감염확률은

$$\begin{aligned} PDF_{W_3(x)} &= conv(\frac{6}{10} conv(\frac{7}{10} PDF_{L_0(x)}, PDF_{R_1(x)}), PDF_{R_3(x)}) \\ &= conv(\frac{6}{10} PDF_{L_1(x)}, PDF_{R_3(x)}) \end{aligned} \tag{11}$$

식 (11)와 같다. 같은 방법으로 각 절점의 감염확률을 구하는 방법은 Table 3과 같고 단계별 계산 방법을 Fig. 6과 같이 도식화 하였다.

식 (11)에 따르면 절점 0에 영향을 받는 절점 1의 감염 확률 ($W_{positive}$)은

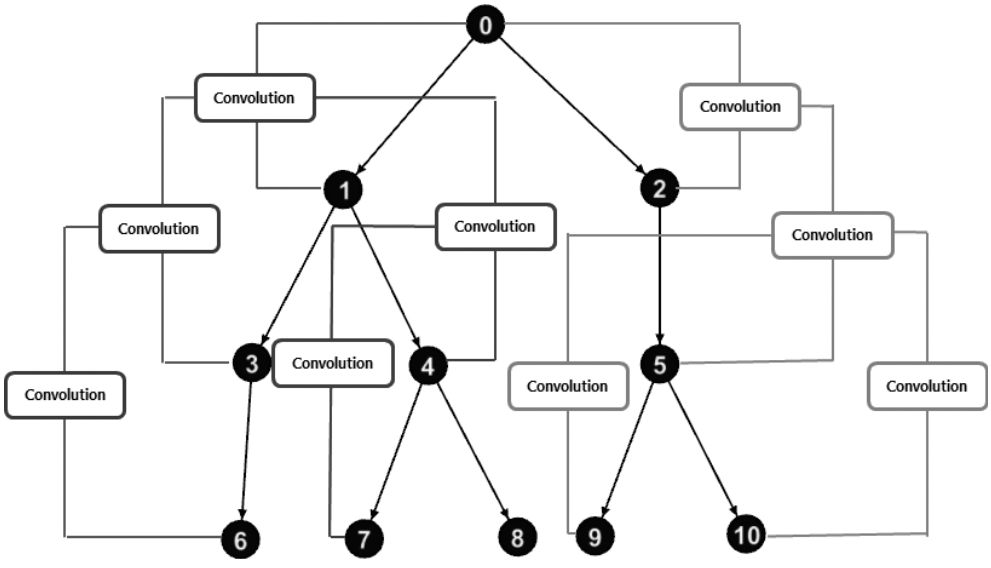


Fig. 6 Analysis process of the example

$$W_{positive} = \int_0^{\infty} PDF_{w_1} = 99.76\% \quad (12)$$

식 (12)가 된다. 식 (12)의 결과는 0번 절점에서 바이러스의 확산이 시작되면 1번 절점이 영향을 받을 확률이 99 % 이상이라는 것을 의미하며 0번 절점과 아주 밀접한 연관을 가지고 있는 절점이라는 것을 의미한다. 앞선 수식과 Fig. 6의 프로세스를 통하여 산정된 각 절점의 감염 확률은 Table 4와 같다.

1, 2, 3, 5번 절점은 감염 확률이 50 % 이상으로 0번에서 시작된 바이러스의 확산에 영향을 받으나 4번과 6번 절점은 40 %를 조금 넘기며 나머지, 7, 8, 9, 10번 절점은 10～20 % 정도의 감염확률을 가진다. 최초절점과의 단계가 커질수록 감염확률은 각 절점의 저항에 영향을 받아 작아졌다. 8번 절점은 가장 감염확률이 낮았다. 평균적으로 0번 절점에서 10번의 확산이 시행되면 1번 정도 영향을 받을 수 있을 만큼 0번 절점과의 연결도가 낮다는 것을 의미한다. 반면에 1번 절점은 0번 절점과의 직접적으로 연결되어있는 만큼 연결도가 가장 높으며 군집화 할 수 있다. 이 결과를 바탕으로 군집화의 기준에 따라서 다음과 같이 집단화할 수 있다. 군집화의 기준이 90 %면 0, 1, 2번 절점이 군집화 되고 50 %면 0, 1, 2, 3, 5번 절점이 군집화 되며 40 %면 0, 1, 2, 3, 4, 5, 6번 절점이 서로 간의 연결도가 높은 소집단으로 구분된다.

$$\begin{aligned} &W_{negative} < 90\%, Group = [node_0, node_1, node_2] \\ &W_{negative} < 50\%, Group = [node_0, node_1, node_2, node_3, node_5] \\ &W_{negative} < 40\%, Group = [node_0, node_1, node_2, node_3, node_4, node_5, node_6] \end{aligned} \quad (14)$$

Table 3 Explanation to the probability of infection in node

Node	Explanation
1	$PDF_{W_1(x)} = conv(\frac{7}{10}PDF_{L_0(x)}, PDF_{R_1(x)})$
2	$PDF_{W_2(x)} = conv(\frac{3}{10}PDF_{L_0(x)}, PDF_{R_2(x)})$
3	$PDF_{W_3(x)} = conv(\frac{6}{10}PDF_{L_1(x)}, PDF_{R_3(x)})$
4	$PDF_{W_4(x)} = conv(\frac{4}{10}PDF_{L_1(x)}, PDF_{R_4(x)})$
5	$PDF_{W_5(x)} = conv(PDF_{L_2(x)}, PDF_{R_5(x)})$
6	$PDF_{W_6(x)} = conv(PDF_{L_3(x)}, PDF_{R_6(x)})$
7	$PDF_{W_7(x)} = conv(\frac{7}{10}PDF_{L_4(x)}, PDF_{R_7(x)})$
8	$PDF_{W_8(x)} = conv(\frac{3}{10}PDF_{L_4(x)}, PDF_{R_8(x)})$
9	$PDF_{W_9(x)} = conv(\frac{4}{10}PDF_{L_5(x)}, PDF_{R_9(x)})$
10	$PDF_{W_{10}(x)} = conv(\frac{6}{10}PDF_{L_5(x)}, PDF_{R_{10}(x)})$

Table 4 Probability of Infection in node

Node	Probability of Infection
1	99.76 %
2	92.14 %
3	71.92 %
4	42.21 %
5	50.12 %
6	41.42 %
7	21.16 %
8	10.12 %
9	15.32 %
10	22.36 %

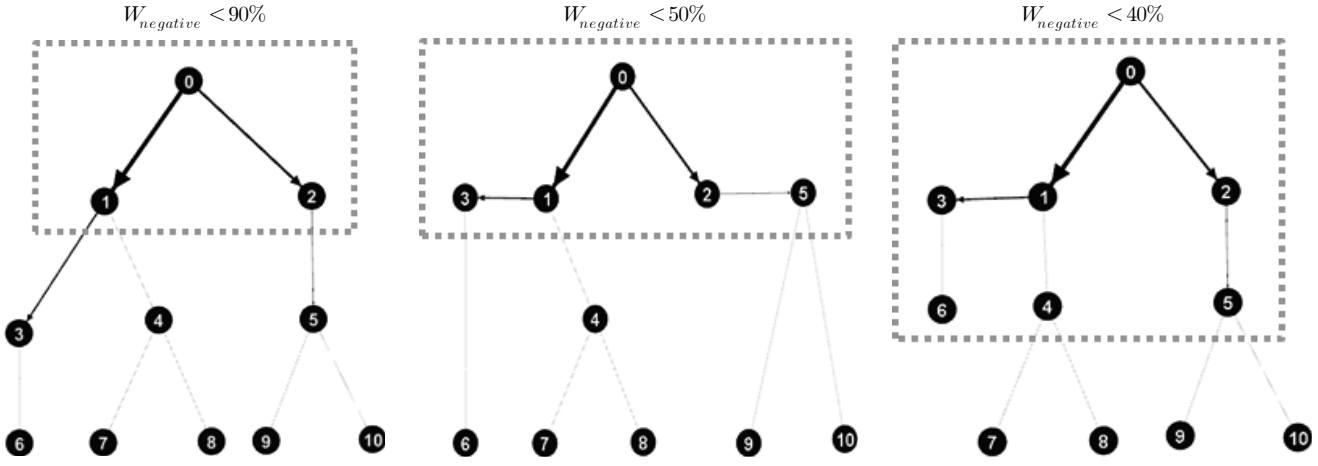


Fig. 7 Result with the criteria of clustering

기준이 완화될수록 0번 절점과 같은 집단에 속하는 절점들이 많아진다. 1번, 2번 절점은 0번 절점의 영향 아래 90 % 이상 감염되기 때문에 $W_{Negative}$가 90 %보다 낮으면 항상 같은 집단으로 포함되며 4번, 6번 절점은 40 %를 조금 상회하는 발현 확률 값에 따라 40 % 이하가 되기 전까지는 포함되지 않는다.

본 방법을 통해서 적용하고자 할 때 관련된 문제 다수의 자료 고찰을 통해서 문제에 대한 해법을 찾기 위한 적절한 군집화의 기준을 설정할 수 있을 것으로 판단된다. 질병의 확산 사례에 적용이 된다면 바이러스의 특성에 따라서 기준을 설정하고 확산의 양상을 분석하고 방역을 위해 차단해야할 거점을 설정할 수 있다.

2. 모델의 검증

결정론적 모델과 비교할 때 변수의 범위가 확률적 분포를 가지는 확률 모델에서는 분석적 해를 구하는 것이 어렵다. 본 연구에서 제안한 CPC 모델 역시 각 절점의 연결선 발현에 대한 값의 확률적 분포만을 제공하기 때문에 분석적인 해를 찾는 것이 불가능하다. 따라서 분포가 주어진 난수를 반복적으로 발생시켜 수치적으로 모의하여 답을 찾는 몬테카를로 시뮬레이션 (Monte-Carlo Simulation)을 통해서 CPC 모델의 결과를 비교 검증할 수 있다. 일반적으로 분석적 해를 통한 검증이 어려울 경우 난수에 가까운 몬테카를로 시뮬레이션의 수치적 해법을 통한 검증은 신뢰할 수 있는 방법이다 (Burt and Garman, 1971; Romualdo and Alessandro, 2000). 본 연구에서 수행한 몬테카를로 시뮬레이션을 통해서 발생한 난수는 1,000,000회 계산이었으며 컨볼루션을 이용한 각 절점의 연결선 발현확률인 Table 4의 결과와 비교하였다. 발현확률을 구하는 방법은 다음 식과 같다.

$$W_{positive} = \frac{n_p}{1{,}000{,}000} \tag{15}$$

n_p : the number of generating nodes

여기서, n_p는 난수를 통한 해석시 절점의 연결선이 발현 되는 횟수를 의미하고 총 1,000,000번의 시행가운데 그 횟수를 통해서 각 절점의 발현 확률을 Table 5와 같이 구할 수 있다.

본 연구에서 개발한 컨볼루션을 이용한 군집화 방법 (Clustering with Probabilistic Convolution)과 몬테카를로 시뮬레이션을 이용한 결과를 유효숫자 5자리까지 비교하였으며 몬테카를로 시뮬레이션의 상대오차의 방법에 따라 오차를 구하면 다음과 같다 (Ringer, 1971; Burt and Garman, 1971).

Table 5 Comparison with MCS

Node	Probability of Infection	
	CPC	MCS
1	0.99761	0.99389
2	0.92143	0.91935
3	0.71922	0.71711
4	0.42211	0.42116
5	0.50125	0.49709
6	0.41424	0.41382
7	0.21164	0.20956
8	0.10121	0.10038
9	0.15322	0.15266
10	0.22363	0.22148

CPC: Clustering with Probabilstic Convolution
MCS: Mote-Carlo Simulation

$$error = \sqrt{\sum_{k=1}^{10}\left(\frac{CPC_k - MCS_k}{MCS_k}\right)^2} \times 100 \tag{16}$$

$$= 0.5208\%$$

식 (16)을 이용해 Table 5의 10개 절점의 연결선 발현 확률을 각각 비교하면 99.4 % 이상의 신뢰도를 보여준다.

3. 기존 모델과 비교

본 연구에서 소개된 방법 (Clustering with Probabilistic Convolution)을 방향성이 있는 가중 네트워크를 소집단으로 군집하는 다른 연구들과 비교해보면 Table 6과 같다. Fig. 5의 예제 네트워크에서 최상위 0번 절점과 연결되어있는 연결선의 합을 10으로 하고 Table 1의 비율과 같이 연결선의 L을 가정하였고 Table 2의 각 절점의 R 평균값을 이용하여 $L-R$로 최종적으로 산정된 연결선의 값이 적용된 결정론적 가중 네트워크를 기준으로 분석하였다.

Clauset-Newman-Moore 모델과 Girvan-Newman 모델은 CPC와 동일한 절점을 군집하였다. Wakita-Tsurumi의 모델은 5번 절점을 제외하였는데 5번 절점은 CPC를 이용한 방법에서도 1, 2, 3, 5번 절점 가운데 0번 절점과 가장 연관도가 낮은 절점이다. 따라서 군집화의 기준을 5번 절점의 감염 확률인 50.12 %보다 높게 설정하면 Wakita-Tsurumi의 모델과 같은 군집 결과

Table 6 Comparison with the existing models

method	CPC (50 %)	Clauset-Newman-Moore	Wakita-Tsurumi	Girvan-Newman
Group	0, 1, 2, 3, 5	0, 1, 2, 3, 5	0, 1, 2, 3	0, 1, 2, 3, 5

를 얻을 수 있다. 5번 절점과 5개의 절점 가운데 2번째로 낮은 3번 절점의 감염 확률 차이는 약 21 % 이상으로 나머지 4개 절점과 같은 집단으로 구분하기에는 범위가 크며 5번째로 낮은 4번 절점과의 차이가 약 7.9 %로 오히려 더 높다.

기존 모델은 소집단을 추출하기 위해 전체 네트워크의 부하량을 산정해서 절점의 감염 확률을 구할 때마다 반복적으로 산정해야하기 때문에 절점수가 많은 네트워크에 있어서는 효율적인 방법이 아닐 수도 있다 (Guimera et al., 2003; Peter et al., 2010). 하지만 CPC의 경우 축차적으로 절점의 감염 확률을 구하며 전체 네트워크 구조와 관계없이 연계된 절점간의 확률분포만 합산하는 분산처리가 가능하기 때문에 전체 연산의 효율도를 높일 수 있다는 장점을 가진다.

$$\frac{number\ of\ eration\ in\ CPC}{number\ of\ eration\ in\ the\ existing\ method} = \frac{=\ number\ of\ links}{=\ {}_nC_2 = \frac{n(n-1)}{2}} \quad (17)$$

n: the number of nodes

식 (17)은 각 모델들의 연산 반복 횟수를 나타낸 것이다. CPC 모델에서 연산 반복 횟수는 절점간의 분산처리에 따라 연결선의 수만큼 시행하면 되지만 기존 모델의 경우 각각의 절점간의 연결 관계를 모두 산정해야하기 때문에 절점 수가 많아질수록 반복 횟수가 기하급수적으로 증가한다. CPC 방법과 기존의 방법을 동일 계층형 구조에서 단계별로 비례적으로 절점 수만 늘려가면서 반복 횟수를 비교해보면 Fig. 8과 같다. 절점 수가 많아질수록 연산 반복 횟수의 차이가 점점 커지는 것을 확인할 수 있다.

기존의 방법과 비교할 때 CPC 방법은 경계조건에 따라 절점과 연결선의 확률론적 값을 모두 반영할 수 있고 기준에 따라서

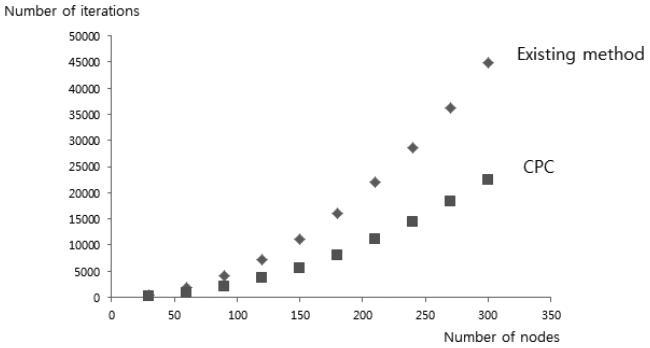


Fig. 8 Comparison of Number of iteration with the existing method

정확한 군집 결과를 산정할 수 있을 뿐만 아니라 연산의 효율도에 있어서도 장점을 가진다.

Ⅳ. 결 론

네트워크 연구 대상이 확대되면서 구현되는 모델과 그 실제 대상의 간극을 줄이려는 시도가 끊임없이 진행되고 있다. 그 차이를 줄이기 위해 선행되어야 할 부분 중 하나는 모델의 입력 자료에 대한 확률적 고찰이다. 불확실성을 내포하는 결정론적 네트워크 연구와 비교할 때 확률론적 네트워크는 결정론적 해답을 도출하기는 어렵지만 확률적 연산이 수행된다면 실측값에 대한 정확한 범위와 확률분포를 제시할 수 있다.

본 연구에서는 네트워크의 절점과 연결선에 저항, 외력과 같은 변수를 적용한 모델을 만들었다. 그리고 저항, 하중의 확률분포를 가지는 절점 및 연결선의 정보에 대한 컨볼루션 연산을 통하여 새로운 연결선의 정보를 확률분포로 규격화하였다. 이 모델을 적용하여 임의의 확산 네트워크 구조를 대상으로 확률분포를 분석하고 그 절점의 감염 확률을 해석하여 유사한 절점들을 군집화 하는데 이용하였다. 이 방법은 신뢰할 수 있는 수치적 해법 가운데 하나인 분포가 동일한 난수를 통한 몬테카를로 시뮬레이션 방법과 비교했을 때 99.4 % 이상의 신뢰도를 확보할 수 있었다. 그리고 기존의 다른 결정론적 네트워크의 군집화 방법과 비교했을 때 군집화의 기준에 따라 유연한 결과를 얻을 수 있으며 정확한 군집화가 가능함을 보였다.

농가의 가축질병에 대한 피해가 확대되면서 그 대응에 관한 연구가 활발해지고 있다. 가축질병의 매개체는 농가를 방문하는 업체 및 관계자이기 때문에 방문빈도에 대한 자료를 수집하면 매개체에 대한 확률적 자료로 구축될 수 있다. 그리고 가축질병에 영향을 미치는 농가의 위생상황과 야생조수류 출현 빈도와 같은 농가별 특징적인 상황도 확률적 자료로 구축할 수 있다. 영국에서는 가축질병에 대응하기 위하여 축산 및 가금 농가를 전산에 등록하고 농가를 방문하는 매개체와 농가의 환경에 대한 자료를 수집하고 있다. 추후 국내에서도 농가에 대한 자료의 전산화와 수집이 본격화되면 이를 확률적 자료로 구축할 수 있을 것이고 본 연구의 활용방안이 높아질 것이다.

확률론적 네트워크에서 이종의 확률적 연산을 통한 새로운 확률의 규격화는 네트워크의 구조 해석을 위한 중요한 도구가 될 수 있음을 확인하였다. 그리고 가축전염병과 관련된 확산을 주제로 하는 방향성을 가지는 네트워크 모델에서 확산 예측 및 차단 등의 해법에 대하여 본 연구에서 개발한 군집화 방법론을 통하여 확률론적으로 해석하고 적절한 방안을 제안할 수 있을 것으로 기대된다.

이 논문은 농촌진흥청 연구사업 (과제번호: PJ009134)의 지원에 의해 이루어진 것임

REFERENCES

1. Burt, J. M. and M. B. Garman, 1971. Conditional Monte Carlo: A simulation technique for stochastic network analysis. *Management Science* 18(3): 207-217.
2. Dechter, R., and J. Pearl, 1989. Tree clustering for constraint networks. *Artificial Intelligence* 38(3): 353-366.
3. Fefferman, C., 1970. Inequalities for strongly singular convolution operators. *Acta Mathematica* 124(1): 9-36.
4. Friedman, N.. 2004. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science* 303.5659: 799-805.
5. Friedman, N., D. Geiger, and M. Goldszmidt, 1997. Bayesian Network Classifiers. *Machine learning* 29 (2-3): 131-163.
6. Girvan, M. and M. E. J Newman, 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99(12): 7821-7826.
7. Gudkov, V., J. E. Johnson, and S. Nussinov, 2002. Graph equivalence and characterization via a continuous evolution of a physical analog. arXiv preprint cond-mat/0209112.
8. Leskovec, J., K. J. Lang, and A. Dasgupta, 2007. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics* 6(1): 29-123.
9. Leskovec, J., K. J. Lang, and M. Mahoney, 2010. Empirical comparison of algorithms for network community detection. Proceedings of the 19th international conference on World wide web, 631-640. New York, USA.
10. Marcos, G. Q., L. Zhao, L. Ronaldo, and A. F. Roseli, 2008. Particle competition for complex network community detection. Chaos: An Interdisciplinary. *Journal of Nonlinear Science* 18(3): 033107-033107.
11. Newman, M. E. J. and M. Girvan, 2004. Finding and evaluating community structure in networks. *Physical Review E.* 69(2): 026113.
12. O'Neil, R., 1963. Convolution operators and $L(p,q)$ spaces. *Duke Mathematical Journal* 30(1): 129-142.
13. Peter, J. M., T. Richardson, K. Macon, A.P. Mason, and J. Onnela, 2010. Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science* 328(5980): 876-878.
14. Pothen, A., H. D. Simon, and K. P. Liou, 1990. Partitioning sparse matrices with eigenvectors of graphs. *Journal on Matrix Analysis and Applications* 11(3): 430-452.
15. Reichardt, J. and S. Bornoholdt, 2004. Detecting fuzzy community structures in complex networks with a Potts model. *Physical Review Letters* 93(21): 218701.
16. Reichardt, J. and S. Bornoholdt, 2006. Statistical mechanics of community detection. *Physical Review Letters* 74(1): 016110.
17. Ringer, L. J., 1971. A statistical theory for PERT in which completion times of activities are inter-dependent. *Management Science* 17(11): 717-723.
18. Romualdo, P. and V. Alessandro, 2001. Epidemic spreading in scale-free networks. *Statistical Mechanics* 86(14): 3200-3203.
19. Roy, S., D. Saha, D. Bandyopadhyay, T. Ueda, and S. Tanaka, 2003. A network-aware MAC and routing protocol for effective load balancing in ad hoc wireless networks with directional antenna. *Proceedings of the 4th ACM international symposium on Mobile ad hoc networking & computing* 88-97. New York, USA.
20. Wu, F. and B.A. Huberman, 2004. Finding communities in linear time: a physics approach. *The European Physical Journal* B 38(2): 331-338.
21. Zhou, H., 2003. Distance, dissimilarity index, and network community structure. *Physical Review E.* 67(6): 061901.