Journal of The Korea Society of Computer and Information
Vol. 24 No. 7, pp. 61-69, July 2019
www.ksci.re.kr
https://doi.org/10.9708/jksci.2019.24.07.061

# The Classification of random graph models using graph centralities

Tae-Soo Cho*, Chi-Geun Han*, Sang-Hoon Lee*

## Abstract

In this paper, a classification method of random graph models is proposed and it is based on centralities of the random graphs. Similarity between two random graphs is measured for the classification of random graph models. The similarity between two random graph models $G^{R_1}$ and $G^{R_2}$ is defined by the distance of $G^{R_1}$ and $G^{R_2}$, where $G^{R_2}$ is a set of random graph $G^{R_2} = \{G_1^{R_2}, \ldots, G_p^{R_2}\}$ that have the same number of nodes and edges as random graph $G^{R_1}$. The distance($G^{R_1}, G^{R_2}$) is obtained by comparing centralities of $G^{R_1}$ and $G^{R_2}$. Through the computational experiments, we show that it is possible to compare random graph models regardless of the number of vertices or edges of the random graphs. Also, it is possible to identify and classify the properties of the random graph models by measuring and comparing similarities between random graph models.

▸Keyword: graph, centrality, random graph, similarity measure, random graph model classification

## I. Introduction

그래프는 객체뿐만 아니라 객체 간의 관계를 정의하기 위한 목적으로 정점과 간선으로 구성된 수학적 구조이다[1]. 랜덤 그래프는 특정 조건을 만족시키는 실제 데이터들을 찾아 그래프를 구성하지 않아도 그 존재성 및 실효성을 파악하기 위한 방법을 제시하는 수학적 모델이다[2]. 그래프의 유사도를 측정하는 중요한 이유는 그래프를 통해 패턴을 인식하고 분석함으로써 유용한 정보를 파악할 수 있기 때문이다. 그래프를 파악하여 유전자 알고리즘, 자연어 처리, 사회 연결망 분석 등 다양한 분야에 활용하기 위해 기계학습이나 데이터 마이닝 등을 통해 분석하는 연구가 활발히 진행되고 있다.

기존에는 그래프 간의 비교를 위해 Graph Edit Distance(GED)가 사용되었다. GED는 두 그래프를 비교하기 위해 하나의 그래프가 다른 그래프로 변환하기 위해 정점과 간선을 삽입, 삭제, 대체하는 비용을 계산한다[3]. GED를 발전 및 입증하기 위해 다양한 연구가 최근까지 이루어지고 있다[4]. 하지만 GED는 복잡한 네트워크 환경의 특성을 파악하기 어려우며 시간이 지남에 따라 변화되는 네트워크를 파악하기에 적합하지 않다. 이러한 문제를 보완하기 위해 그래프의 특성을 파악하기에 적합한 중심성(Centrality)을 이용하여 그래프 비교 방법이 연구되고 있다. 최근에는 네트워크가 시간이 지남에 따른 변화를 파악하기 위해 그래프의 정점을 고정하고 간선의 변화에 따른 중심성을 계산 및 분석하는 연구가 진행되었다[5]. [5]는 정점과 간선의 수가 동일한 형태의 랜덤 그래프들을 생성하여 실제 간선이 달라진 네트워크와의 비교하였다. [6]은 정점과 간선이 다른 두 그래프의 유사함을 비교하기 위해 척도를 만들어 거리를 계산한 뒤 유사도를 측정하였다. [6]에서는 특성을 가지는 그래프들의 비교를 통해 중심성이 유사도 측정에 적합함을 보였으며 실제 네트워크 환경의 그래프들을 비교하였다.

본 논문에서는 그래프 중심성을 이용하여 서로 다른 정점과

• First Author: Tae-Soo Cho, Corresponding Author: Chi-Geun Han
*Tae-Soo Cho (taesoocho@naver.com), Dept. of Computer Engineering, Kyung Hee University
*Chi-Geun Han (cghan@khu.ac.kr), Dept. of Computer Engineering, Kyung Hee University
*Sang-Hoon Lee (a01b01c01@khu.ac.kr), Dept. of Computer Engineering, Kyung Hee University
• Received: 2019. 04. 02, Revised: 2019. 06. 26, Accepted: 2019. 06. 27.
• This study was conducted as a result of the support by SW-centered College and Korea Electric Power Corporation. (Grant number:R18XA02)

간선의 수를 가지는 랜덤 그래프 모델들의 특성을 파악하고 수치화 및 가시화를 통해 유사성을 비교 및 분류하고자 한다. 이를 통해 각각의 필요 분야에 어떠한 성질의 랜덤 그래프가 필요한지 판단할 수 있는 객관적인 자료로 사용되는 것을 목표로 한다. 또한, 실제 네트워크 환경 등에 더욱 적합한 랜덤 그래프 모델을 개발하기 위한 토대가 될 수 있는 척도를 제시하고자 한다.

먼저, 척도를 구하기 위해 사용한 중심성과 랜덤 그래프 모델들을 2장 Related Works에서 설명하고, 3장 Methods에서 중심성이 그래프의 특성을 설명할 수 있는지와 척도에 사용될 랜덤 그래프 모델들이 적합한지 파악하고 제안하는 방법의 진행 절차를 그림, 흐름도 및 의사코드를 기반으로 설명한다. 4장에서는 척도를 사용하여 유사성 비교 및 분류 실험에 사용될 랜덤 그래프 모델을 간단히 소개한 후 실험 결과를 분석한다. 마지막으로 5장에서는 실험을 통한 결론을 서술한다.

## II. Related Works

2장에서는 그래프의 특성을 파악하기 위해 사용된 중심성과 제안하는 방법에 사용되는 랜덤 그래프 모델에 대해 설명한다.

### 1. Centrality

중심성은 정점과 간선의 관계를 통해 중요한 정점을 판별하는 방법이다. 제안하는 방법에서 사용될 중심성은 다음과 같다.

#### 1.1 Betweenness Centrality (BC)

BC는 두 정점 사이의 최단 경로에 다른 하나의 정점이 bridge 역할을 하는 횟수를 정량화하는 방법이다[7]. $n_{y,z}$가 정점 $y$와 정점 $z$의 최단 경로의 개수이고 $n_{y,z}(x)$가 정점 $x$를 지나는 정점 $y$와 정점 $z$의 최단 경로의 수를 나타낼 때 BC는 (식 1)과 같이 표현된다.

$$c_x^{BC} = \sum_{\substack{y,z \in V \\ x \neq y, \neq z}} \frac{n_{y,z}(x)}{n_{y,z}}, x \in V \text{ (식 1)}$$

#### 1.2 Closeness Centrality (CC)

CC는 하나의 정점에서 다른 모든 정점들로 가는 각각의 최단 경로를 이용하여 정의된다[8]. $\delta(x,y)$는 정점 $x$와 정점 $y$ 간의 최단 경로 길이이며 다른 정점들이 정점 $x$와의 거리를 계산하여 CC는 (식 2)와 같이 나타낸다.

$$c_x^{CC} = \frac{1}{\sum_{\substack{y \in V \\ x \neq y}} \delta(x,y)}, x \in V \text{ (식 2)}$$

#### 1.3 Degree Centrality (DC)

DC는 하나의 정점에 연결된 간선의 개수로 정의된다[9]. $\deg(x)$가 정점 $x$에 연결된 간선의 수라고 할 때 DC는 (식 3)과 같이 나타낸다.

$$c_x^{DC} = \deg(x), x \in V \text{ (식 3)}$$

#### 1.4 Eigenvector Centrality (EC)

EC는 정점의 중요도를 측정하기 위해 power iteration method를 사용하여, 그래프를 인접행렬로 표현한 고유벡터(Eigenvector)이다[10]. 각 정점들의 상대적인 값을 가진다. 정점 $x$의 고유벡터 값을 $\lambda x$라고 할 때 (식 4)와 같이 표현된다.

$$c_x^{EC} = \lambda x, x \in V \text{ (식 4)}$$

### 2. Random Graphs

랜덤 그래프는 1959년에 E.N. Gilbert가 베르누이 확률을 통해 간선들이 정점에 연결되는 모델을 제시하며 시작되었으며, 비슷한 시기에 현재까지 대표적으로 사용되는 독립 확률을 통해 랜덤 그래프를 생성하는 Erdos-Renyi(ER) 모델이 연구되었다[11][12]. 랜덤 그래프의 필요성이 대두되면서 ER 모델에서 삼자 폐쇄를 형성하여 랜덤 그래프를 생성하는 Watts-Strogatz(WS) 모델, 실제 네트워크 환경에서 scale이 없으며 중요한 허브가 존재한다는 점을 통해 종속적인 확률로 각 정점에 간선을 연결하여 랜덤 그래프를 생성하는 Barabasi-Albert(BA) 모델, WWW(World Wide Web)의 네트워크를 구조화하기 위해 BA모델에 fitness 변수를 적용하여 랜덤 그래프를 생성하는 Bianconi-Barabasi(BB) 모델 등 네트워크의 구조를 파악하고 구성하기 위해 랜덤 그래프 모델들을 발전시켜왔다[13][14]. 이외에도 랜덤 그래프를 활용하여 각 분야에 적용하기 위해 다양한 모델이 제시되고 있으며 활발히 연구가 진행되고 있다.

#### 2.1 Erdős-Rényi model (ER)

ER은 정점들에 독립적인 확률을 통해 간선을 연결하는 방법이다. 따라서 모든 정점에서 간선을 가질 확률이 동일한 랜덤 그래프를 생성한다[12].

#### 2.2 Barabási-Albert model (BA)

BA은 실제 네트워크가 시간이 지날수록 정점이 증가하고, 새로운 정점이 추가될 때 기존 그래프의 높은 degree를 가지는 정점에 연결될 확률이 높다는 특징을 통해 랜덤 그래프를 생성하는 방법이다. 즉, 간선들과 많이 연결된 정점일수록 더욱 높은 확률을 통해 새로운 정점과 연결하여 랜덤 그래프를 생성한다[15].

#### 2.3 Random Graph using $\mathrm{Pr}\ddot{u}fer\,Sequence$ (PS)

PS는 구축하고자 하는 그래프의 정점의 개수 n과 간선의 개수 m이 주어졌을 때, 먼저 $\mathrm{Pr}\ddot{u}fer\,Sequence$를 이용하여 n개

의 정점이 모두 연결되도록 tree를 구성한다. Tree를 구성하기 위해 사용된 간선의 개수인 n-1개를 뺀 m-(n-1)개의 간선을 독립적인 확률을 통해 다른 정점들과 연결하여 랜덤 그래프를 생성하는 방법이다[6].

### 2.4 Random Geometric Graph (RGG)

RGG는 인간 사회 관계망을 표현하기 위해 연구되었으며, radius 범위를 설정하고 각 정점마다 설정한 범위보다 가까운 영역에 다른 정점이 존재한다면 두 정점을 간선으로 연결하는 방법이다[16]. 아래 그림은 3개의 정점이 있을 때 설정한 radius가 각각의 정점마다 점선으로 표현되어있다. 2개의 정점은 서로 radius 범위에 포함되므로 간선을 연결되며 다른 하나의 정점은 radius 범위에 포함되지 않으므로 간선이 연결되지 않는 형태로 랜덤 그래프를 형성한다.

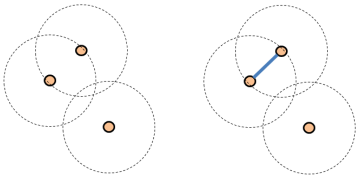

Fig. 1. Example RGG graph

### 2.5 Random Regular Graph (RRG)

RRG는 정점이 가질 수 있는 degree를 설정하여 랜덤 그래프를 생성하는 모델이다[2]. 아래의 그림을 예로 살펴보면 두 그래프의 형태는 다르지만 각 그래프의 모든 정점의 degree가 3인 3-Regular Graph이다.

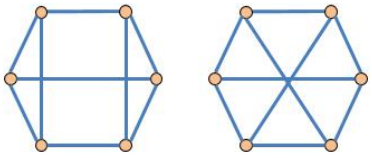

Fig. 2. Example RRG graph

### 2.6 Watts-Strogatz (WS)

WS는 ER모델에서 삼자폐쇄를 생성하도록 하여 ER 모델의 짧은 평균 경로 길이를 유지하면서 클러스터링을 형성할 수 있는 형태의 랜덤 그래프 모델이다[13]. 아래의 예제 그림과 같이 동일한 개수의 정점과 간선을 가지더라도 클러스터링 변수($\beta$)에 따라 다른 형태의 그래프가 형성된다. 좌측의 그래프는 $\beta$가 0일 때의 그래프이고, 우측 그래프는 $\beta$가 1인 경우의 그래프이다.

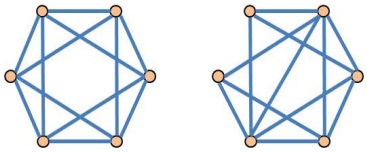

Fig. 3. Example WS graph

### 2.7 Bianconi-Barabasi (BB)

BB는 BA모델에 Fitness라는 매개변수를 추가하여 정점마다 적합도에 따라 연결되는 간선의 수를 다르게 생성하는 랜덤 그래프 모델이다[14].

## III. Methods

본 장에서는 특성을 가지는 그래프들의 중심성 비교를 통해 중심성이 그래프의 성질 비교에 적합함을 보이고, 랜덤 그래프를 비교하기에 적합한 랜덤 그래프 모델을 선정한다. 중심성과 선정한 랜덤 그래프 모델을 이용하여 랜덤 그래프들의 성질을 파악하고 유사도에 따른 분류하는 방법을 제안하고자 한다.

### 1. Goodness of Fit Test

#### 1.1 Centrality

본 절에서는 중심성이 각각의 그래프마다 특성이 다름을 나타낼 수 있는지를 확인한다. 중심성 비교를 위해 특성을 가지는 Path Graph, Cycle Graph, Complete Graph를 선정하여 비교하였다.

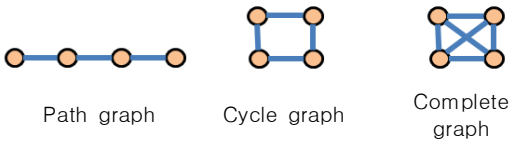


Fig. 4. Graphs with specific properties

Fig. 4는 10개의 정점을 가지는 3개의 그래프의 중심성(BC, CC, DC, EC)을 계산하여 비 오름차순으로 정렬한 그림이다. 그래프간의 중심성을 비교하여 그래프들의 특징을 설명한다.

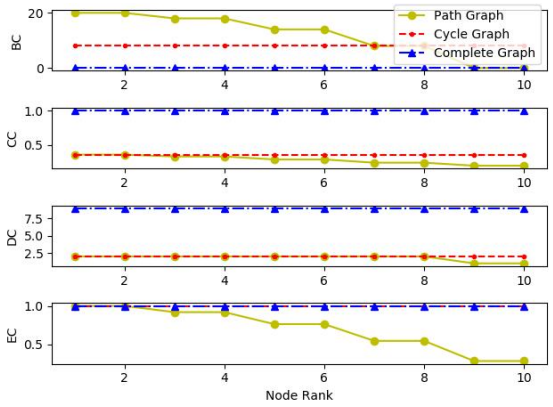


Fig. 5. Centralities of graphs with specific properties

Path Graph는 BC, CC, EC에 대해서 서서히 감소하는 형태를 가지고 있으며 DC의 경우, 처음과 끝 정점의 degree는 1이며 다른 정점들의 degree는 2의 값을 가지는 특성을 알 수 있다. Cycle Graph는 모든 중심성에서 균일한 형태의 중심성을 가지는 것을 알 수 있다. Complete Graph는 모든 정점이 최단 경로로 이루어져있으므로 BC는 0의 값을, CC는 1의 값을 가지며 DC를 통해 모든 정점의 degree가 다른 그래프보다 많다는 것을 알 수 있다. 이처럼 특성을 가지는 그래프들마다 다른 중심성을 가지며 이는 그래프 비교에 적합하다는 것을 알 수 있다.

### 1.2 Random Graph Models

본 절에서는 랜덤 그래프 모델들이 각각의 특성이 다름을 보이고, 그래프 비교에 적합한 랜덤 그래프 모델을 선정하기 위해 PS, ER, BA을 비교한다. 중심성은 정점과 간선의 수에 민감하기 때문에 중심성을 이용한 비교를 위해서는 두 가지의 제한사항이 존재한다. 첫 번째로 비교할 두 그래프가 동일한 정점과 간선의 수를 가져야 한다. 특정 랜덤 그래프 모델의 경우, 간선의 수가 랜덤하게 생성되므로 이러한 랜덤 그래프 모델을 비교하기 위해서는 간선의 수를 지정하여 랜덤 그래프를 생성할 수 있는 모델이 필요하다. 두 번째로는 그래프가 항상 연결 그래프 형태로 존재하여야 한다. 3개의 모델을 비교하여 척도에 사용될 랜덤 그래프 모델이 가지는 특성을 관찰한다. 두 랜덤 그래프 모델 $R_1$과 $R_2$를 비교한다고 할 때 $R_1$ 모델의 그래프 $G^{R_1}$를 생성하고 동일한 개수의 정점과 간선의 수를 가지는 $R_2$ 모델의 그래프를 p개 생성하여 ($G^{R_2}=\{G_1^{R_2},\dots,G_p^{R_2}\}$) 각 그래프의 중심성을 계산한 후 비 오름차순(nonincreasing order)으로 정렬한다. $G^{R_2}$집합에 속한 그래프마다 정렬된 중심성의 해당 순서별 평균과 표준편차 값을 구한다. $G^{R_1}$의 중심성과 $G^{R_2}$의 평균과 표준편차를 이용하여 $distance(G^{R_1},G^{R_2})=(\overline{Y^{BC}},\overline{Y^{CC}},\overline{Y^{DC}},\overline{Y^{EC}})$를 구한다. 거리를 구하는 세부적인 내용은 다음 절에서 자세히 설명한다.

세 개의 모델(PS, ER, BA)은 정점이 100, 200, 300, 400개, 그리고 각 랜덤 그래프들의 특성을 잘 나타낼 수 있도록 $n(n-1)/2\times 0.35$개의 간선을 가지도록 하여 비교하였다. 또한, p는 100개로 설정하였다. 거리는 0과 가까울수록 유사한 형태의 그래프를 의미하고 1과 가까울수록 서로 다른 형태의 그래프를 나타내도록 정의하였다.

Table 1. $distance(G^{PS},G^{ER})$

| $distance(G^{PS},G^{ER})$ | $\overline{Y^{BC}}$ | $\overline{Y^{CC}}$ | $\overline{Y^{DC}}$ | $\overline{Y^{EC}}$ |
|---|---|---|---|---|
| n = 100, m = 1,732 | 0.0380 | 0.0533 | 0.0543 | 0.0547 |
| n = 200, m = 6,965 | 0.0373 | 0.0525 | 0.0535 | 0.0232 |
| n = 300, m = 15,697 | 0.0591 | 0.0670 | 0.0797 | 0.0332 |
| n = 400, m = 27,930 | 0.0465 | 0.0565 | 0.0585 | 0.0005 |

$distance(G^{PS},G^{ER})$의 경우, 두 모델의 거리가 매우 가깝다는 것을 알 수 있으며 특성이 매우 유사하다는 것을 알 수 있다.

Table 2. $distance(G^{PS},G^{BA})$

| $distance(G^{PS},G^{BA})$ | $\overline{Y^{BC}}$ | $\overline{Y^{CC}}$ | $\overline{Y^{DC}}$ | $\overline{Y^{EC}}$ |
|---|---|---|---|---|
| n = 100, m = 1,732 | 0.9937 | 0.9950 | 0.9950 | 0.8313 |
| n = 200, m = 6,965 | 0.9990 | 0.9982 | 0.9982 | 0.8508 |
| n = 300, m = 15,697 | 0.9993 | 0.9987 | 0.9987 | 0.8688 |
| n = 400, m = 27,930 | 0.9990 | 0.9990 | 0.9990 | 0.8755 |

Table 3. $distance(G^{ER},G^{BA})$

| $distance(G^{ER},G^{BA})$ | $\overline{Y^{BC}}$ | $\overline{Y^{CC}}$ | $\overline{Y^{DC}}$ | $\overline{Y^{EC}}$ |
|---|---|---|---|---|
| n = 100, m = 1,732 | 0.9977 | 0.9923 | 0.9923 | 0.8302 |
| n = 200, m = 6,965 | 0.9982 | 0.9960 | 0.9960 | 0.8505 |
| n = 300, m = 156,97 | 0.9996 | 0.9980 | 0.9980 | 0.8912 |
| n = 400, m = 27,930 | 0.9993 | 0.9990 | 0.9990 | 0.8867 |

$distance(G^{PS},G^{BA})$와 $distance(G^{ER},G^{BA})$의 경우는 두 거리 모두 먼 것을 알 수 있다. 따라서 BA는 PS와 ER에 대해 특성이 유사하지 않은 것을 알 수 있다.

세 개의 모델 중 PS와 ER은 유사한 특성을 가지는 것을, BA는 두 모델과 유사한 특성을 가지지 않음을 알 수 있었다. 제안하는 방법에서는 정점과 간선 개수를 정하여 랜덤 그래프를 생성할 수 있고, 정점이나 간선의 개수에 상관없이 연결 그래프를 생성할 수 있는 PS, BA를 이용하여 다른 랜덤 그래프 모델들을 비교하는 척도를 제안한다.

## 2. Measuring Graph Similarity

### 2.1 Suggested Method

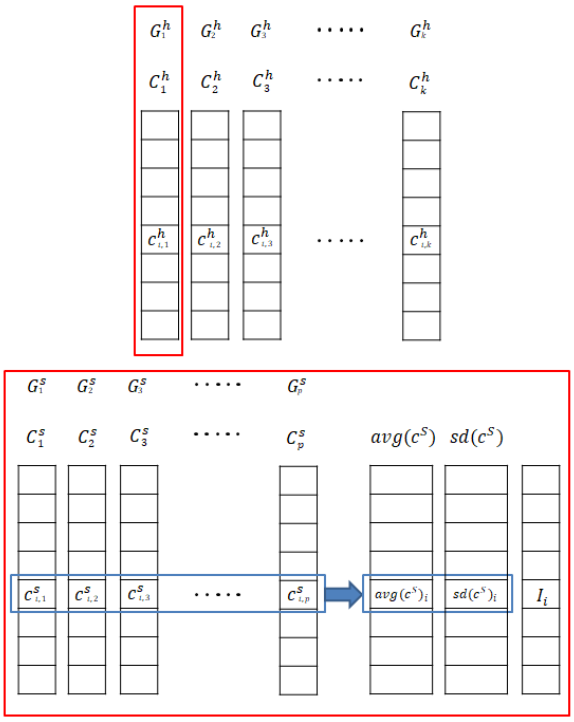


Fig. 6. Calculation of distance using centralities

먼저 특정 중심성 q에 대해 그래프의 거리를 구하는 방법을 설명한다. 비교하고자하는 랜덤 그래프($G^h$)를 생성한 뒤, 중심성 벡터 $c^h \in R^n$를 계산하고 비 오름차순으로 정렬한다. 척도로 사용될 그래프 $G^S = \{G_1^s, ..., G_p^s\}$ (여기서 $G^S$는 $G^{PS}$ 또는 $G^{BA}$이다.)를 생성하고 $G^S$ 그래프들의 각각의 중심성 $(C_1^s, ..., C_p^s)$를 계산하고 비 오름차순으로 정렬한다. 생성된 p개의 중심성의 정렬된 순서 $i$에 따른(그림에서 행으로 표시, $\{c_{i,1}^s, ..., c_{i,p}^s\}$) 평균$(avg(c^S)_i)$과 표준편차$(sd(c^S)_i)$를 구한다. $G^h$의 중심성과 $G^S$의 $avg(c^S)$와 $sd(c^S)$를 이용하여 척도를 계산한다.

$$I_i = \begin{cases} 1, & \text{if } \left| \dfrac{c_i^h - avg(c^S)_i}{sd(c^S)_i} \right| > \delta, i = 1, ..., n \\ 0, & otherwise \end{cases} \quad (식\ 5)$$

척도 $I_i$는 두 랜덤 그래프의 유사도를 판단하여 유사할 경우 0의 값을 가지고 유사하지 않은 경우는 1의 값을 갖는다. 따라서 중심성이 큰 (그래프에서 중요하다고 판단한 정점) 경우를 파악할 수 없기 때문에 가중치 $I_i^w = I_i \times w_i (w_i = n - i + 1)$를 부여한다.

$$Y^q = \sum_{i=1}^{n} I_i^w / \{n(n+1)/2\} \quad (식\ 6)$$

이 방법을 이용하여 다른 중심성들에 대해서도 $Y^q$를 계산하여 $(Y^{BC}, Y^{CC}, Y^{DC}, Y^{EC})$를 얻게 된다. 지금까지의 과정은 Fig. 5.에 해당하며 하나의 $G^h$에 대해서 나타낸 것이다. 전체 과정을 k번 반복하여 (즉, p개의 $G^s$를 생성하여 $Y^q$를 계산하는 과정을 $G^h$가 k개 생성될 때까지의 반복) 표준 랜덤 그래프들 간의 거리를 파악한다. 거리 $(Y^{BC}, Y^{CC}, Y^{DC}, Y^{EC})$는 최종적으로 k개가 얻어지게 되고, 이들 값들의 평균 계산하여 각각의 중심성에 대한 표준 랜덤 그래프의 거리를 구한다. 즉, 각 중심성 q에 대해 $\overline{Y^q} = \sum_{i=1}^{k} Y_i^q / k$이며 $distance(G^h, G^S) = (\overline{Y^{BC}}, \overline{Y^{CC}}, \overline{Y^{DC}}, \overline{Y^{EC}})$이다. $G^h$가 일정 개수가 되기까지 반복하는 이유는 하나의 $G^h$는 랜덤 그래프 모델의 특정한 하나의 예이므로 랜덤 그래프 모델의 특성을 잘 대변한다고 볼 수 없다. 따라서 일정개수가 될 경우까지 반복하여 랜덤 그래프 모델들의 특성이 보편적으로 나타날 수 있도록 하여 비교하였다.

### 2.2 Flowchart and Pseudocode for Proposed Method

제안하는 방법의 흐름도와 의사코드는 다음과 같다. 먼저 비교하고자 하는 랜덤 그래프 $G^h$를 생성한다. $G^h$를 각각의 중심성 (BC, CC, DC, EC)에 대해 계산하고 비오름차순으로 정렬한다. $G^h$와 동일한 정점과 간선의 수를 가지는 $G^s$를 p개만큼 생성하고 각각의 중심성에 대해 계산하고 비오름차순으로 정렬한다. Fig. 6과 같이 정렬한 p개에 $G^s$ 그래프의 행별로 평균과 표준편차를 구한다. (식 6)을 통해 척도를 계산하여 거리를 측정하게 된다. $G^h$가 k개가 될 때까지 일련의 과정을 반복하고 각각의 수행을 통해 얻어낸 거리들을 계산하여 유사도를 측정하게 된다. 이 때 $G^s$는 독립적인 확률을 통해 정점과 간선을 연결하는 PS 모델과 종속적인 확률을 통해 정점과 간선을 연결하는 BA 모델을 각각 수행하여 비교대상 랜덤 그래프 모델이 어떤 모델과 가까운지 혹은 두 모델과 달리 특이한 형태로 그래프를 생성하는지 비교한다.

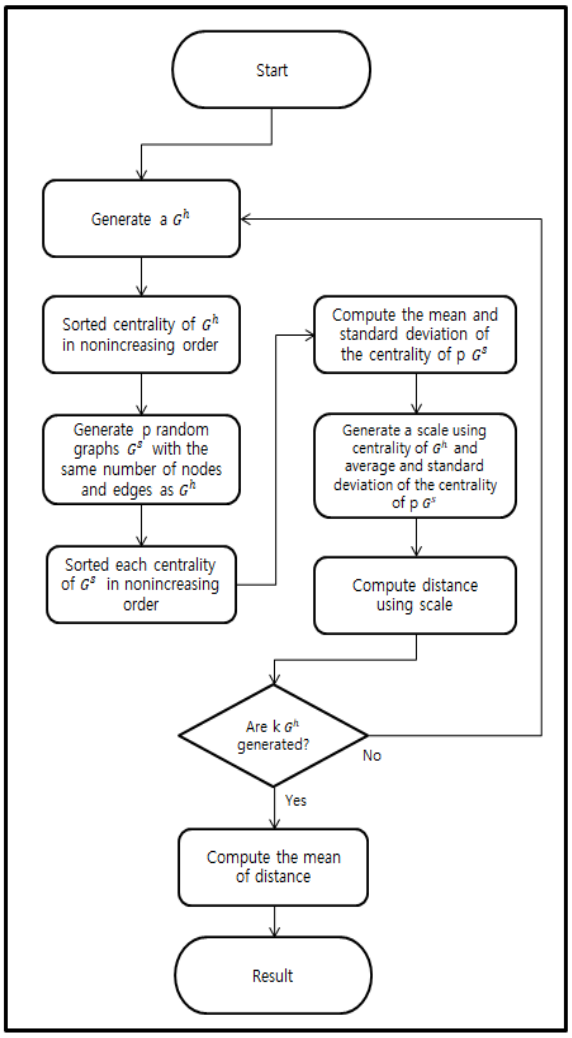


Fig. 7. Flowchart for calculating distances

CENT = {BC, CC, DC, EC}

$G^S$ = { $G^{PS}$ or $G^{BA}$ }

for $i = 1,...,k$

generate a $G_i^h$ of model $G^H$

$G_i^h = (V,E), |V| = n, |E| = m$

undirected and connected graph

generate a set of random graphs

$G^S = \{G_1^s,...,G_p^s\}$

// compute $Y_i = (Y_i^{BC}, Y_i^{CC}, Y_i^{DC}, Y_i^{EC})$

for each centrality q in CENT

$Y_i^q$ = compCentMG( $G^h$, $G^S$, $q$)

$$\overline{Y^q} = \sum_{i=1}^{k} Y_i^q / k$$

$$distance(G^H, G^S) = (\overline{Y^{BC}}, \overline{Y^{CC}}, \overline{Y^{DC}}, \overline{Y^{EC}})$$

compCentMG( $G^h$, $G^S$, $q$)

$c^h$ = sorted centrality q of $G^h$ in nonincreasing order

$avg(c^S), sd(c^S) =$ compAvgSd $(G^S, q)$

$avg(c^S), sd(c^S) \in R^n$

for $i = 1,...,n$

$$I_i = \begin{cases} 1, & \text{if } |\frac{c_i^h - avg(c^S)_i}{sd(c^S)_i}| > \delta \\ 0, & otherwise \end{cases}$$

for $i = 1,...,n$

$I_i^w = I_i \times w_i$

val = $\sum_{i=1}^{n} I_i^w / \{n(n+1)/2\}$

return val

compAvgSd $(G^S, q)$

for each $G_j^s \in G^S, j = 1,...,p$

Let $a_q \in R^n$ be the centrality q of $G_j^s$ and sort $a_q$ in nonincreasing order

$c_{.j} = a_q$

$C = \{c_{ij}\}, i = 1,...,n, \ j = 1,...,p$

for $i = 1,...,n$

$avg_{c_i} = average(c_{i,1},...,c_{i,p})$

$sd_{c_i} = standartDeviation(c_{i,1},...,c_{i,p})$

return

$c_{avg} = (avg_{c_1},...,avg_{c_n}), c_{sd} = (sd_{c_1},...,sd_{c_n})$

## IV. Experimental Results

실험에 사용된 랜덤 그래프 모델은 다음과 같다.

Table 4. Random graph model sets

| Model Set | Explanation |
|---|---|
| Random Geometric ($G_G^H$) | Radius를 사용하여 랜덤 그래프를 생성하는 모델 |
| Random Regular ($G_R^H$) | Regular를 사용하여 랜덤 그래프를 생성하는 모델 |
| Watts-Strogatz ($G_W^H$) | ER모델에서 클러스터링 변수를 추가하여 랜덤 그래프를 생성하는 모델 |
| Bianconi-Barabasi ($G_B^H$) | BA모델에서 Fitness라는 매개변수를 추가하여 생성하는 모델 |

Table. 4에 나타난 4개의 랜덤 그래프 모델은 각각 100, 200, 300, 400개의 정점을 가지도록 실험하였다. 각 정점의 간선 비율을 유사하게 비교하기 위해 정점이 n개일 때 그래프는 최대 n(n-1)/2개의 간선 수를 가질 수 있으므로, 랜덤 모델을 잘 표현 할 수 있게 최대 간선 수의 30~35%를 가지도록 그래프를 생성하였다. 또한, WS 모델의 경우에는 정점과 간선에 상관없이 클러스터링 변수($\beta$)에 따라 랜덤 그래프가 생성된다. $\beta$를 0과 1(최솟값과 최댓값)로 나누어 실험하였다.

중심성(BC, CC, DC, EC)을 이용하여 그래프들의 유사정도를 파악하기 위해 $distance(G^H, G^S) = (\overline{Y^{BC}}, \overline{Y^{CC}}, \overline{Y^{DC}}, \overline{Y^{EC}})$의 값을 비교한다. 또한, 종합적인 거리를 평가하기 위해 $\overline{\overline{Y_{H,S}}} = \sum_{q \in BC,CC,DC,EC} \overline{Y^q}/4$, $H \in \{G,R,D,W,B\}$, $S \in \{PS,BA\}$를 계산하였다.

Table 5. $distance(G_G^H, G^{PS})$

| $distance(G_G^H, G^{PS})$ | $\overline{Y^{BC}}$ | $\overline{Y^{CC}}$ | $\overline{Y^{DC}}$ | $\overline{Y^{EC}}$ | $\overline{\overline{Y_{G,PS}}}$ |
|---|---|---|---|---|---|
| n = 100 | 0.9507 | 0.9342 | 0.8975 | 0.8000 | 0.8956 |
| n = 200 | 0.9753 | 0.9690 | 0.9552 | 0.8781 | 0.9444 |
| n = 300 | 0.9855 | 0.9810 | 0.9745 | 0.9066 | 0.9619 |
| n = 400 | 0.9930 | 0.9914 | 0.9881 | 0.9340 | 0.9766 |

Table 6. $distance(G_G^H, G^{BA})$

| $distance(G_G^H, G^{BA})$ | $\overline{Y^{BC}}$ | $\overline{Y^{CC}}$ | $\overline{Y^{DC}}$ | $\overline{Y^{EC}}$ | $\overline{\overline{Y_{G,BA}}}$ |
|---|---|---|---|---|---|
| n = 100 | 0.8443 | 0.7878 | 0.8169 | 0.7725 | 0.8054 |
| n = 200 | 0.9302 | 0.8978 | 0.9084 | 0.8588 | 0.8988 |
| n = 300 | 0.9484 | 0.9298 | 0.9357 | 0.9062 | 0.9300 |
| n = 400 | 0.9603 | 0.9453 | 0.9492 | 0.9249 | 0.9449 |

RGG의 경우, $distance(G_G^H, G^{PS})$와 $distance(G_G^H, G^{BA})$가 모든 거리에서 차이를 보였으며 PS와 BA 모두 유사하지 않은 형태의 그래프를 생성한다는 것을 파악할 수 있다.

Table 7. $distance(G_R^H, G^{PS})$

| $distance(G_R^H, G^{PS})$ | $\overline{Y^{BC}}$ | $\overline{Y^{CC}}$ | $\overline{Y^{DC}}$ | $\overline{Y^{EC}}$ | $\overline{\overline{Y_{R,PS}}}$ |
|---|---|---|---|---|---|
| n = 100 | 0.8594 | 0.8582 | 0.8580 | 0.8757 | 0.8628 |
| n = 200 | 0.9088 | 0.9023 | 0.9015 | 0.9158 | 0.9071 |
| n = 300 | 0.9235 | 0.9202 | 0.9192 | 0.9327 | 0.9239 |
| n = 400 | 0.9356 | 0.9321 | 0.9310 | 0.9433 | 0.9355 |

Table 8. $distance(G_R^H, G^{BA})$

| $distance(G_R^H, G^{BA})$ | $\overline{Y^{BC}}$ | $\overline{Y^{CC}}$ | $\overline{Y^{DC}}$ | $\overline{Y^{EC}}$ | $\overline{\overline{Y_{R,BA}}}$ |
|---|---|---|---|---|---|
| n = 100 | 0.8610 | 0.8604 | 0.8603 | 0.8781 | 0.8650 |
| n = 200 | 0.9527 | 0.9492 | 0.9487 | 0.9536 | 0.9511 |
| n = 300 | 0.9658 | 0.9627 | 0.9598 | 0.9655 | 0.9635 |
| n = 400 | 0.9715 | 0.9693 | 0.9687 | 0.9721 | 0.9704 |

RRG는 $distance(G_R^H, G^{PS})$의 $\overline{\overline{Y_{R,PS}}}$에서 0.86~0.93 사이의 값을 가지고, $distance(G_R^H, G^{BA})$의 $\overline{\overline{Y_{R,BA}}}$는 0.86~0.97 사이의 값을 가지는 것을 알 수 있다. 이는 PS와 BA에 대해 유사하지 않은 형태를 가지는 것을 알 수 있다.

WS는 클러스터링 비율을 결정하는 특정 변수에 따라 그래프의 형태가 달라지므로 변수의 확률을 최대와 최소로 하였을 때를 구분하여 실험하였다. 변수의 최댓값과 최솟값인 0과 1일 때의 그래프 형태를 실험하였으며 각각 $G_{W_0}^H$, $G_{W_1}^H$라고 표현한다.

Table 9. $distance(G_{W_0}^H, G^{PS})$

| $distance(G_{W_0}^H, G^{PS})$ | $\overline{Y^{BC}}$ | $\overline{Y^{CC}}$ | $\overline{Y^{DC}}$ | $\overline{Y^{EC}}$ | $\overline{\overline{Y_{W_0,PS}}}$ |
|---|---|---|---|---|---|
| n = 100 | 0.8406 | 0.9186 | 0.8976 | 0.9061 | 0.8907 |
| n = 200 | 0.9413 | 0.9707 | 0.9465 | 0.9504 | 0.9522 |
| n = 300 | 0.9762 | 0.9881 | 0.9648 | 0.9672 | 0.9741 |
| n = 400 | 0.9902 | 0.9951 | 0.9738 | 0.9752 | 0.9836 |

Table 10. $distance(G_{W_0}^H, G^{BA})$

| $distance(G_{W_0}^H, G^{BA})$ | $\overline{Y^{BC}}$ | $\overline{Y^{CC}}$ | $\overline{Y^{DC}}$ | $\overline{Y^{EC}}$ | $\overline{\overline{Y_{W_0,BA}}}$ |
|---|---|---|---|---|---|
| n = 100 | 0.9263 | 0.9627 | 0.9458 | 0.9447 | 0.9449 |
| n = 200 | 0.9579 | 0.9789 | 0.9686 | 0.9685 | 0.9685 |
| n = 300 | 0.9680 | 0.9840 | 0.9757 | 0.9760 | 0.9759 |
| n = 400 | 0.9739 | 0.9870 | 0.9805 | 0.9808 | 0.9806 |

Table 11. $distance(G_{W_1}^H, G^{PS})$

| $distance(G_{W_1}^H, G^{PS})$ | $\overline{Y^{BC}}$ | $\overline{Y^{CC}}$ | $\overline{Y^{DC}}$ | $\overline{Y^{EC}}$ | $\overline{\overline{Y_{W_1,PS}}}$ |
|---|---|---|---|---|---|
| n = 100 | 0.1008 | 0.1352 | 0.1492 | 0.1119 | 0.1243 |
| n = 200 | 0.3231 | 0.4077 | 0.4357 | 0.3368 | 0.3758 |
| n = 300 | 0.8059 | 0.8290 | 0.8366 | 0.6744 | 0.7865 |
| n = 400 | 0.8572 | 0.8738 | 0.8793 | 0.6960 | 0.8266 |

Table 12. $distance(G^H_{W_1}, G^{BA})$

| $distance(G^H_{W_1}, G^{BA})$ | $\overline{Y^{BC}}$ | $\overline{Y^{CC}}$ | $\overline{Y^{DC}}$ | $\overline{Y^{EC}}$ | $\overline{\overline{Y_{W_1,BA}}}$ |
|---|---|---|---|---|---|
| n = 100 | 0.9220 | 0.9778 | 0.9168 | 0.8670 | 0.9209 |
| n = 200 | 0.9532 | 0.9500 | 0.9486 | 0.8960 | 0.9370 |
| n = 300 | 0.9656 | 0.9632 | 0.9625 | 0.9574 | 0.9622 |
| n = 400 | 0.9729 | 0.9702 | 0.9695 | 0.9659 | 0.9696 |

WS는 $distance(G^H_{W_0}, G^{PS})$에서는 PS 모델과 비교하였을 때 유사하지 않은 형태를 가지고 $distance(G^H_{W_1}, G^{PS})$의 경우에는 정점의 수가 적은 경우에서 유사한 형태를 가지지만 정점의 수가 증가할수록 다른 형태를 보이는 것을 알 수 있다. 또한, $distance(G^H_{W_0}, G^{BA})$와 $distance(G^H_{W_1}, G^{BA})$에서는 유사하지 않은 형태를 가지는 것을 알 수 있다.

Table 13. $distance(G^H_B, G^{PS})$

| $distance(G^H_W, G^{PS})$ | $\overline{Y^{BC}}$ | $\overline{Y^{CC}}$ | $\overline{Y^{DC}}$ | $\overline{Y^{EC}}$ | $\overline{\overline{Y_{B,PS}}}$ |
|---|---|---|---|---|---|
| n = 100 | 0.0427 | 0.0493 | 0.0503 | 0.0487 | 0.0478 |
| n = 200 | 0.0638 | 0.0732 | 0.0737 | 0.0755 | 0.0716 |
| n = 300 | 0.0735 | 0.0741 | 0.0747 | 0.0740 | 0.0741 |
| n = 400 | 0.0385 | 0.0555 | 0.0560 | 0 | 0.0375 |

Table 14. $distance(G^H_B, G^{BA})$

| $distance(G^H_W, G^{BA})$ | $\overline{Y^{BC}}$ | $\overline{Y^{CC}}$ | $\overline{Y^{DC}}$ | $\overline{Y^{EC}}$ | $\overline{\overline{Y_{B,BA}}}$ |
|---|---|---|---|---|---|
| n = 100 | 0.9510 | 0.9440 | 0.9420 | 0.8780 | 0.9288 |
| n = 200 | 0.9715 | 0.9645 | 0.9650 | 0.9380 | 0.9598 |
| n = 300 | 0.9797 | 0.9757 | 0.9753 | 0.9347 | 0.9664 |
| n = 400 | 0.9835 | 0.9790 | 0.9790 | 0.9625 | 0.9760 |

BB의 경우는 $distance(G^H_B, G^{PS})$에서 PS와 거의 유사한 그래프 형태를 가지는 것을 알 수 있으며 $distance(G^H_B, G^{BA})$에서 BA와 매우 다른 그래프 형태를 가지도록 생성되는 것을 알 수 있었다.

Table. 15와 Table. 16은 각각의 랜덤 그래프 모델의 유사도를 비교하기 위해 계산한 Table. 5-14의 마지막 행의 값들이다. 이는 각각의 랜덤 그래프 모델들의 중심성들을 통해 얻어낸 종합적인 표준거리($\overline{\overline{Y_{H,S}}}$)들의 값이다.

Table 15. $distance(G^H_*, G^{PS})$

| $distance(G^H_*, G^{PS})$ | $\overline{\overline{Y_{G,PS}}}$ | $\overline{\overline{Y_{R,PS}}}$ | $\overline{\overline{Y_{D,PS}}}$ | $\overline{\overline{Y_{W,PS}}}$ | $\overline{\overline{Y_{B,PS}}}$ |
|---|---|---|---|---|---|
| n = 100 | 0.8063 | 0.7401 | 0.2969 | 0.7974 | 0.0478 |
| n = 200 | 0.9139 | 0.7491 | 0.6796 | 0.8069 | 0.0716 |
| n = 300 | 0.9285 | 0.7495 | 0.7507 | 0.8231 | 0.0741 |
| n = 400 | 0.8998 | 0.7493 | 0.8555 | 0.8080 | 0.0375 |

Table 16. $distance(G^H_*, G^{BA})$

| $distance(G^H_*, G^{BA})$ | $\overline{\overline{Y_{G,BA}}}$ | $\overline{\overline{Y_{R,BA}}}$ | $\overline{\overline{Y_{D,BA}}}$ | $\overline{\overline{Y_{W,BA}}}$ | $\overline{\overline{Y_{B,BA}}}$ |
|---|---|---|---|---|---|
| n = 100 | 0.8571 | 0.7378 | 0.9234 | 0.9458 | 0.9288 |
| n = 200 | 0.9460 | 0.7314 | 0.9592 | 0.9509 | 0.9598 |
| n = 300 | 0.9658 | 0.7264 | 0.9717 | 0.9295 | 0.9664 |
| n = 400 | 0.9702 | 0.7217 | 0.9759 | 0.9204 | 0.9760 |

이 값들을 통해 랜덤 그래프 모델 간의 분류를 위해 가로축을 $\overline{\overline{Y_{H,PS}}}/\overline{\overline{Y_{H,BA}}}$인 PS와 BA를 통한 비율을 사용하고 세로축을 $\overline{\overline{Y_{H,PS}}}$값으로 하여 이차원 유클리드 공간에서 표현하였다.

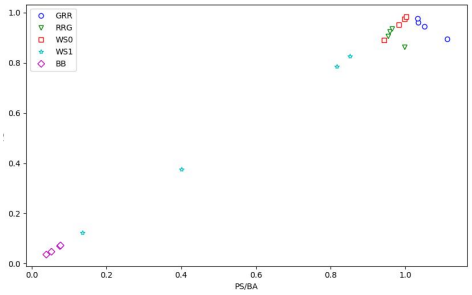


Fig. 8. Distances of experiment

2차원 유클리드 공간에 표현하였을 때 모델들의 분포를 통해 각 랜덤 그래프 모델의 특성별 분류가 가능하다. BB의 경우, BA와는 유사하지 않고 PS와 유사하므로 가로축과 세로축 모두 0과 가까운 공간에 군집되는 것을 알 수 있다. RGG와 RRG는 PS와 BA 모두 유사하지 않은 그래프 형태를 가지므로 가로축과 세로축 모두 1과 가까운 공간에 위치하였다. WS는 $\beta$값이 0인 경우는 PS와 BA에 유사하지 않은 형태로 군집되었으며 $\beta$값이 1인 경우에는 BA에서는 유사하지 않고 PS의 경우에는 점점 유사하지 않는 형태를 보여, 선형형태로 표현되는 것을 알 수 있다, 각각의 모델별로 군집되어 있는 형태를 가지는 것을 알 수 있다.

## V. Conclusion

본 논문에서는 중심성을 이용하여 랜덤 그래프 모델들의 거리를 측정하고 분류하는 방법에 대해 제안하였다. 간선이 독립적인 확률을 통해 생성되는 그래프 모델 PS와 간선이 종속적인 확률에 따라 생성되는 그래프 모델 BA를 기준으로 랜덤 그래프 모델들의 그래프 형태가 어느 쪽에 더 유사한지, 혹은 두 그래프 형태와 전혀 다른 형태로 그래프가 형성하는지를 가시화하여 구분할 수 있도록 하였다. 이를 통해 여러 분야에서 사용되는 랜덤 그래프 모델들의 특성을 비교할 수 있으며 특정 분야에 더욱 효율적인 랜덤 그래프 모델을 적용하거나 새로운 랜덤 그래프 모델을 연구될 경우, 객관적인 척도로 사용되는 것을 목표로 하고 있다.

본 논문은 하나의 랜덤 그래프 모델을 비교하기 위해 랜덤 그래프를 일정 개수 생성, 비교하여 그래프간의 유사도를 측정하는 방법을 반복하여 랜덤 그래프 모델의 정모델을 파악하였다. 이 방법으로 그래프간의 유사도를 비교하는 경우, 많은 수행이 반복되므로 수행시간이 길어지는 단점이 있다. 또한, PS 모델과 BA 모델을 통해 유사도를 비교하기 때문에 두 모델과 유사하지 않는 랜덤 그래프 모델의 특성을 파악하기 어렵다.

향후에는 본 논문에서 많은 수행을 반복하는 문제와 랜덤 그래프 모델간의 간접적인 비교에서 나타나는 제한사항을 개선하기 위해서 모델간의 직접적인 비교를 통해 전체적인 그래프 모델의 특성을 정확하고 신속하게 파악하기 위한 연구를 진행할 계획이다.

## REFERENCES


[1] West, Douglas Brent. Introduction to graph theory. Vol. 2. Upper Saddle River: Prentice hall, 2001.

[2] Bollobás, Béla. "Random graphs." Modern graph theory. Springer, New York, NY, 1998. 215-252.

[3] Sanfeliu, Alberto, and King-Sun Fu. "A distance measure between attributed relational graphs for pattern recognition." IEEE transactions on systems, man, and cybernetics 3 (1983): 353-362.

[4] Akoglu, Leman, Hanghang Tong, and Danai Koutra. "Graph based anomaly detection and description: a survey." Data mining and knowledge discovery 29.3 (2015): 626-688.

[5] Pignolet, Yvonne Anne, et al. "The many faces of graph dynamics." Journal of Statistical Mechanics: Theory and Experiment 2017.6 (2017): 063401.

[6] Tae-Soo Cho, Chi-Geun Han, and Sang-Hoon Lee. "Measurement of graphs similarity using graph centralities." JKCSI 23.12 (2018): 57-64.

[7] Freeman, Linton C. "A set of measures of centrality based on betweenness." Sociometry (1977): 35-41.

[8] Okamoto, Kazuya, Wei Chen, and Xiang-Yang Li. "Ranking of closeness centrality for large-scale social networks." International Workshop on Frontiers in Algorithmics. Springer, Berlin, Heidelberg, 2008.

[9] Freeman, Linton C. "Centrality in social networks conceptual clarification." Social networks 1.3 (1978): 215-239.

[10] Bonacich, Phillip. "Some unique properties of eigenvector centrality." Social networks 29.4 (2007): 555-564.

[11] Gilbert, Edgar N. "Random graphs." The Annals of Mathematical Statistics 30.4 (1959): 1141-1144.

[12] ERDdS, P., and A. R&WI. "On random graphs I." Publ. Math. Debrecen 6 (1959): 290-297.

[13] Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of 'small-world'networks." nature 393.6684 (1998): 440.

[14] Bianconi, Ginestra, and A-L. Barabási. "Competition and multiscaling in evolving networks." EPL (Europhysics Letters) 54.4 (2001): 436.

[15] Barabási, Albert-László, and Réka Albert. "Emergence of scaling in random networks." science 286.5439 (1999): 509-512.

[16] Penrose, Mathew D. "On k-connectivity for a geometric random graph." Random Structures & Algorithms 15.2 (1999): 145-164


## Authors


Tae-Soo Cho received the B.S. in Medical IT Marketing from Eulji University, Korea, in 2018, respectively. He is currently in master's course in the Department of Computer Engineering, Kyung Hee University. He is interested in Graph Theory, Genetic Algorithm and Network Analysis.

Chi-Geun Han received the B.E. and M.E. degrees in Industrial Engineering from Seoul National University and Ph.D. degree in Computer Science from the Pennsylvania State University, USA 1991. Dr. Han joined the faculty of the Department of Computer Engineering at Kyung Hee University, Korea, in 1992. He is currently a Professor in the Department of Computer Engineering, Kyung Hee University. He is interested in Graph Theory and Network Analysis.

Sang-Hoon Lee received the B.S., M.S. in Computer Engineering from Kyung Hee University, Korea, in 2010, 2012, respectively. Sang Hoon Lee went on for a doctorate of the Department of Computer Engineering at Kyung Hee University, Suwon, Korea, in 2012. He is currently in doctorate course in the Department of Computer Engineering, Kyung Hee University. He is interested in community detection, Genetic Algorithm and graph theory, and metaheuristic.