Journal of The Korea Society of Computer and Information
Vol. 23 No. 12, pp. 57-64, December 2018
www.ksci.re.kr
https://doi.org/10.9708/jksci.2018.23.12.057

# Measurement of graphs similarity using graph centralities

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

## Abstract

In this paper, a method to measure similarity between two graphs is proposed, which is based on centralities of the graphs. The similarity between two graphs $G_1$ and $G_2$ is defined by the difference of distance($G_1$, $G_{R_1}$) and distance($G_2$, $G_{R_2}$), where $G_{R_1}$ and $G_{R_2}$ are set of random graphs that have the same number of nodes and edges as $G_1$ and $G_2$, respectively. Each distance ($G_*$, $G_{R_*}$) is obtained by comparing centralities of $G_*$ and $G_{R_*}$. Through the computational experiments, we show that it is possible to compare graphs regardless of the number of vertices or edges of the graphs. Also, it is possible to identify and classify the properties of the graphs by measuring and comparing similarities between two graphs.

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

## I. Introduction

객체 간의 연결을 통해 구성되는 네트워크는 물리학, 생물학, 경제학, 사회학 등의 다양한 분야에 사용되고 있다. 이는 그래프의 정점과 간선의 관계를 통해 객체들의 관계를 쉽게 표현할 수 있기 때문이다. 따라서 그래프를 이용하여 복잡한 네트워크 환경의 특성을 파악하거나 실제 네트워크 환경이 시간흐름에 따라 변화되는지에 대한 연구가 많이 진행되고 있다.

기존에 그래프를 비교하기 위한 방법으로는 Graph Edit Distance(GED)가 사용되었다[1]. GED는 비교할 한 쌍의 그래프를 $G_1$과 $G_2$라고 할 때 $G_1$이 $G_2$로 변환되는 과정의 단계가 최소가 되도록 정점과 간선을 삽입, 삭제함으로써 그래프 편집 비용을 구하게 된다. 그러나 $GED(G_1,G_2)=GED(G_1,G_3)$라고 하더라도 $G_2$와 $G_3$가 유사한 형태가 된다는 보장이 없다. 즉, 동일한 GED 값에 대해 그래프는 전혀 다른 형태들을 생성할 수 있어 그래프 유사성을 판단하기 부적절하다[2]. [2]에서는 GED 대신 정점의 중요도를 판별하는 중심성(Centrality)[3]을 사용하여 그래프의 유사도를 측정하였다. 주어진 그래프 $G$에 대해 시간에 따라 추가, 삭제되는 간선들에 따른 각 정점의 중심성 값과 무작위 그래프들에서의 각 정점의 중심성이 어떻게 달라지는지를 파악하였다. $G$의 각 정점별로 달라지는 중심성을 계산하고 무작위 그래프들에 있는 각 정점들의 중심성 값 평균과 표준편차를 이용하는 통계적 방법으로 비교하였다.

본 논문에서는 중심성을 사용하여 두 그래프 $G_1$과 $G_2$의 유사도를 계산하기 위해 $G_1$과 $G_2$와 각각 같은 정점 수와 간선 수의 무작위 그래프들과의 거리(distance)를 확인하여 두 그래프의 유사도를 정의하는 방법을 제안한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 중심성과 무작위 그래프에 대해 살펴보고, 3장에서는 사용할 중심성과 무작위 그래프가 그래프 비교에 적합한지 알아본 후 제안하는 방법에 대해 설명한다. 4장에서 제안하는 방법을 통한 실험 결과를 보이며, 마지막으로 5장에서 결론을 기술한다.

• 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: 2018. 10. 02, Revised: 2018. 11. 07, Accepted: 2018. 11. 18.
• This study was conducted as a result of the support by SW-centered College and Korea Electric Power Corporation. (Grant number:R18XA02)

## II. Related Works

2014년에는 Closeness Centrality를 이용하여 대상 그래프가 무작위로 진화하는 경로와 실제 진화 경로를 구별하여 유사도를 측정하는 방법이 연구되었으며, 2015년에는 유클리안 거리(Euclidean distance)와 인접행렬(Adjacency matrix)을 이용하여 두 그래프간의 정점과 간선을 비교하여 그래프 유사도를 측정하는 방법이 연구되었다[4][5]. 2017년에는 여러 종류의 Centrality를 이용하여 하나의 그래프가 정점의 개수를 변하지 않고 간선만 변화할 경우의 그래프 유사도를 측정하는 연구가 진행되었다[2].

본 논문에서는 [2]를 발전시켜 정점과 간선 수가 다르더라도 두 그래프를 비교하여 유사도를 측정하는 방법을 제시한다. 이를 위해 이용할 Centrality와 Random Graph에 대해 설명하고자 한다.

### 1. Centrality

중심성은 계산 방법에 따라 다양하게 나뉘며 정점의 중요도를 식별하는데 사용된다. Fig.1은 정점 집합 V와 간선 집합 E로 정의되는 무향, 연결 그래프 G이다. 이를 통해 아래에서 중심성을 설명한다.

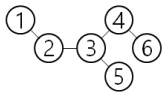


Fig. 1. Centrality example graph

#### 1.1 Betweenness Centrality (BC)

BC는 하나의 정점이 다른 두 정점간의 최단거리에 포함되는 빈도수를 이용하여 정의된다[6]. BC를 기준으로 Fig. 1을 볼 때 다른 두 개의 정점의 최단거리에 속하는 경우가 많은 정점 3이 중요하다고 판단할 수 있다. $\sigma_{yz}$는 정점 $y$와 정점 $z$의 최단 경로 개수라고하고 $\sigma_{yz}(x)$는 $y$와 $z$ 정점의 최단 경로에 정점 $x$가 포함될 경우의 수를 나타낼 때 BC는 (식 1)과 같이 표현된다.

$$C_b(x) = \sum_{x \neq y \neq z} \frac{\sigma_{yz}(x)}{\sigma_{yz}}, x,y,z \in V \text{ (식 1)}$$

#### 1.2 Closeness Centrality (CC)

CC는 정점에서 다른 모든 정점들로 가는 각각의 최단 거리를 이용하여 정의된다[7]. 이는 특정한 정점이 다른 정점들과의 거리가 짧다면 그 정점이 중요하다고 판단하는 것이다. Fig. 1에서 정점 3이 다른 정점들의 최단거리의 합이 최소가 되므로 중심성이 가장 크다. $d(y,x)$는 정점 $x$와 정점 $y$ 간의 거리를 나타내며, (식 2)와 같이 정점 $x$와 다른 정점들의 거리를 구하여 CC를 나타낸다.

$$C_c(x) = \frac{1}{\sum_{y \in V} d(y,x)}, x \in V \text{ (식 2)}$$

#### 1.3 Degree Centrality (DC)

DC는 정점이 가지는 간선 수를 이용하여 정의된다[8]. 하나의 정점이 가지는 간선 수가 많다면 그 정점이 중요하다고 평가한다[7]. Fig. 1에서 가장 많은 간선을 가지는 정점 3이 가장 큰 중심성 DC를 갖는다. $E_x$를 정점 $x$에 연결된 모든 간선의 집합이고, 그 집합의 크기를 $|E_x|$라 하면 정점 $x$의 DC는 (식 3)과 같이 정의된다.

$$C_d(x) = |E_x|, x \in V \text{ (식 3)}$$

#### 1.4 Eigenvector Centrality (EC)

EC는 그래프를 인접행렬로 표현한 행렬의 최대 고윳값(Eigenvalue)에 해당하는 고유벡터(Eigenvector)이다[9]. 정점 $x$의 고유벡터 값을 $\lambda(x)$라고 할 때 (식 4)와 같이 나타낸다.

$$C_e(x) = \lambda(x), x \in V \text{ (식 4)}$$

Table. 1은 Fig. 1 그래프의 각 정점의 중심성 값이다. 각각의 중심성에서 가장 큰 값을 나타낸 정점은 정점 3이다. 따라서 정점 3이 그래프의 정점들 중 가장 중요한 정점임을 알 수 있으며 그에 비해 정점 1과 6은 상대적으로 중요성이 낮은 것을 알 수 있다.

Table 1. Each vertex centrality in Fig. 1

| *Node Number* | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| *Betweenness Centrality* | 0 | 4 | 8 | 4 | 0 | 0 |
| *Closeness Centrality* | 0.38 | 0.56 | 0.71 | 0.56 | 0.45 | 0.38 |
| *Degree Centrality* | 1 | 2 | 3 | 2 | 1 | 1 |
| *Eigenvector Centrality* | 0.37 | 0.71 | 1 | 0.71 | 0.52 | 0.37 |

### 2. Random Graph

유사도를 측정하고자 하는 대상 그래프들을 비교하기 위해서 무작위 그래프를 이용하여 척도를 만든다. 기존에 사용되던 무작위 그래프 생성 방법들의 특징을 살펴본다.

#### 2.1 Erdős-Rényi model

Erdős-Rényi model은 $n$개의 정점들이 각각 독립확률 $p$를 통해 간선을 추가하기 때문에 다른 정점들에 영향을 받지 않는다. 완전 그래프 Kn를 생성한 후 각 간선마다 확률 $p$로 실제 그래프에 존재할지 결정하여 무작위 그래프를 생성한다[10].

#### 2.2 Barabási-Albert model

Barabási-Albert model은 초기 그래프를 생성하고 확률에 따라 새로운 정점을 하나씩 추가하여 무작위 그래프를 만든다. 이 때 확률은 기존 그래프의 정점이 가지고 있는 간선의 수에 비례하여 증가한다. 즉, 많은 간선을 가지는 정점일수록 새로 추가되는 정점과 연결될 확률이 높다[11].

# III. Methods

본 장에서는 중심성이 그래프의 특징을 나타낼 수 있는지 확인하기 위해 세 종류의 특수한 형태를 가지는 그래프들의 중심성을 비교한다. 그리고 무작위 그래프가 그래프 간의 거리 측정에 적합한지 알아보도록 한다. 이를 활용하여 새로운 척도를 정의하고, 제안하는 방법을 자세히 설명한다.

## 1. Goodness of Fit Test

### 1.1 Goodness of Fit Test of Centrality

Fig. 2의 특수한 형태를 가지는 그래프(Path 그래프, Star 그래프, Paley 그래프[12])를 비교하여 중심성이 그래프간의 유사도 평가에 적합한지 살펴보도록 한다.

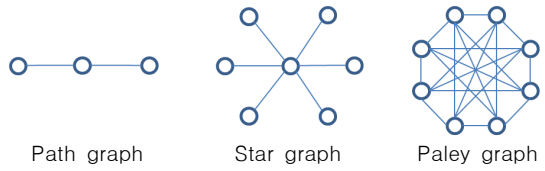


Fig. 2. Graphs with specific properties

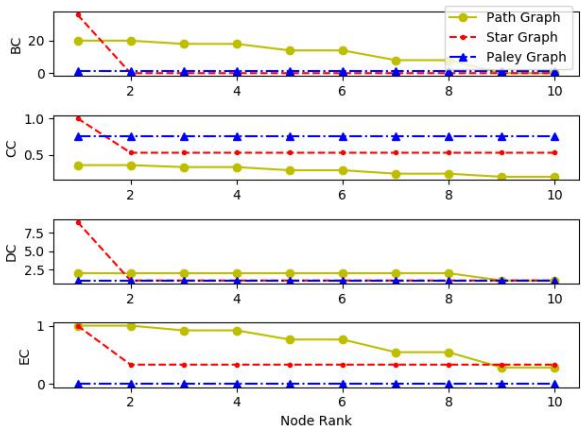


Fig. 3. Centralities of graphs with specific properties

Fig. 3는 10개의 정점을 가지는 Path 그래프, Star 그래프, Paley 그래프의 각 정점에 대한 BC, CC, DC, EC를 비오름차순으로 정렬한 그림이다. Path 그래프의 경우는 모든 중심성에서 대체로 균일한 형태를 보이며 서서히 감소하는 형태를 나타내는 특징이 있다. Star 그래프는 중심성이 가장 큰 값을 가지는 중앙 정점을 제외한 모든 정점에서 동일한 중심성을 보인다. Paley 그래프의 경우, 모든 중심성이 정점에 관계없이 동일한 분포가 나타난다. 이러한 분석을 통해 특별한 형태를 갖는 그래프들은 확연히 다른 중심성의 분포를 보인다는 것을 알 수 있다. 본 논문에서는 그래프의 중심성의 분포 차이를 이용하여 그래프의 거리를 정의하고자 한다. 제안한 방법은 특정 그래프의 중심성 값들을 동일한 크기(정점 수와 간선 수)를 가지는 무작위 그래프들의 중심성과 비교하는데, 이 때 특정 순서(정점 번호가 아닌 비오름차순으로 정렬된 순서)의 대상 그래프의 중심성 값과 여러 개의 무작위 그래프의 중심성의 해당 순서의 평균값과 비교한다. 비교는 통계적인 방법을 통해 평균값과 동일하거나 또는 평균값과 다르다는 검정을 하게 된다. 이들 검정값들을 모아 그래프의 거리를 정의하는데, 자세한 사항은 2.1절과 2.2절에서 설명한다.

### 1.2 Connected Random Graph

중심성을 통한 척도를 추출하기 위해 중심성과의 거리를 비교하는 무작위 그래프는 다음과 같은 조건을 만족해야한다.

1) 어떠한 경우에도 무작위 그래프의 연결성이 보장되도록 생성해야 한다.

2) 특정 정점에 간선이 연결될 확률이 다른 정점에 영향을 받지 않는 균형 있는 무작위 그래프를 생성할 필요가 있다.

지금부터는 Connected 무작위 그래프를 생성하는 방법을 소개한다. 거리를 측정할 대상 그래프가 $G^T$라고 하면 $G^T$의 정점 수를 $n$, 간선 수를 $m$으로 나타내고, 무작위로 생성된 $n-2$개의 원소를 갖는 Prüfer sequence($Ps$)를 생성하여 $G^T$와 같은 정점의 개수를 갖는 트리(tree)를 다음의 방법으로 구성한다[13].

$G^T$의 정점 개수가 6, 간선의 개수를 12, 무작위 $Ps$가 [4,3,4,4]라고 가정한다. Fig. 4의 초기상태와 같이 $G^T$와 동일한 개수의 정점들을 생성하고 node label list = [1,2,3,4,5,6]을 만든다. $Ps$의 첫 번째 값인 4와 node label list에서 $Ps$에 존재하지 않는 값 중 최솟값인 1을 선택한다. 선택된 값인 정점 1과 정점 4를 연결하고 $Ps$와 node label list에서 선택된 값을 제거한다. $Ps$는 [3,4,4]가 되고 node label list는 [2,3,4,5,6]의 값을 가진다. 위의 단계를 거치면 Fig.4의 pass 1과 같은 형태를 보인다.

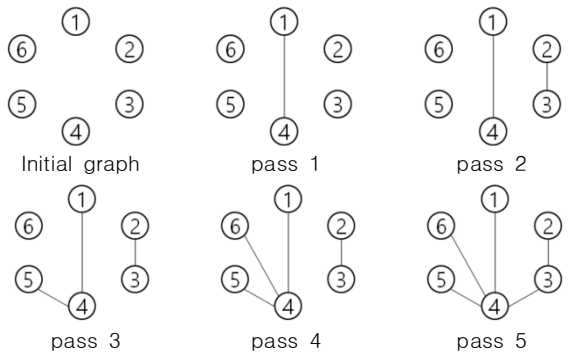


Fig. 4. The process of making a random tree

정점을 선택하고 간선을 연결하는 이러한 과정을 반복하면 pass 2 ~ 4의 형태로 진행된다. pass 4의 단계까지 진행되면 $Ps$는 empty list가 되고 node label list는 [3,4]의 값이 남게 된다. Node label list의 남은 2개의 값을 가지는 정점 3과 정점 4를 연결하면 무작위 트리가 완성된다.

$G^T$와 동일한 간선의 수를 갖도록 $m-(n-1)$개의 간선(트리에 존재하지 않은)을 무작위로 추가하여 connected 무작위

그래프를 생성한다. 트리를 완성하기 위해 5개의 간선이 사용되었으므로 $G^T$와 동일한 간선 수를 가지도록하기 위해서 7개의 간선을 무작위하게 연결하여 connected 무작위 그래프를 완성한다.

### 1.3 Goodness of Fit Test for Random Graph

Barabási-Albert model은 간선을 많이 가지는 정점일수록 연결될 확률이 높아지기 때문에 무작위한 형태의 그래프를 생성하기 어렵다. Erdős-Rényi model은 그래프의 연결성이 보장되지 않는다. Connected 무작위 그래프는 정점이나 간선의 개수에 관계없이 정점이 독립적인 확률을 통해 연결되고, 항상 연결 그래프의 형태를 가지므로 대상 그래프와의 거리 측정 사용에 적합하다.

## 2. Measuring Graph Similarity

### 2.1 Suggested Method

먼저 하나의 중심성에 대해 대상 그래프의 거리를 구하는 방법을 설명한다. 거리 측정 대상 그래프를 $G^T$라고 할 때 $G^T$의 중심성을 계산하여 비오름차순으로 정렬한 값을 $C^t \in R^n$로 정의한다. $G^T$와 동일한 정점 수와 간선 수를 가지는 무작위 그래프를 $p$개만큼 생성한다. $p$개의 무작위 그래프 중심성을 각각 계산하여 비오름차순 정렬한 결과를 $C^r_{\cdot 1}, \ldots, C^r_{\cdot p} \in R^n$로 표시한다. $G^T$의 정렬된 중심성 벡터에서 $i$번째 위치한 값을 $c_i^t$라고 하고, $i$번째 위치한 무작위 그래프들의 중심성을 $c^r_{i,1}, \ldots, c^r_{i,p}$으로 나타낸다. (Fig. 5 참고)

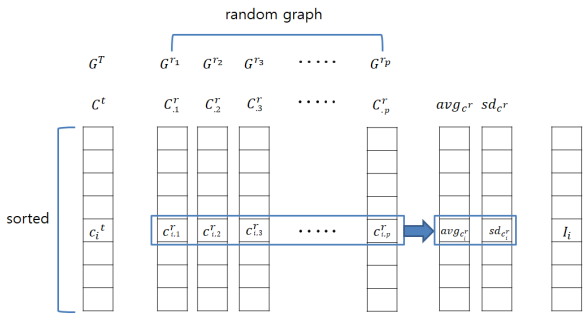


Fig. 5. calculation of distance using centralities

다음으로 $c^r_{i,1}$에서 $c^r_{i,p}$까지 값들의 평균과 표준편차를 계산한다. $avg_{c_i}$를 무작위 그래프들의 $i$번째 중심성의 평균, $sd_{c_i}$를 표준편차라고 할 때 (식 5)와 같이 상수 $\delta$를 기준으로 벗어나는 척도 $I_i$를 구한다.

$$I_i = \begin{cases} 1, & \dfrac{\left| avg_{c_i} - c_i^t \right|}{sd_{c_i}} > \delta \\ 0, & otherwise \end{cases} \quad \text{(식 5)}$$

즉, 대상 그래프의 $i$번째 중심성이 무작위 그래프의 $i$번째 중심성들의 평균값과 비교하여 기준 값 $\delta$범위를 벗어나게 되면, $I_i$값은 1, 그렇지 않은 경우는 0으로 표시되고, 이 값에 가중치 $w_i$를 곱하여 $I_i^w = I_i \times w_i$를 계산하고, 이들 값들을 누적한 값 $\sum_{i=1}^{n} I_i^w$을 구하게 된다. 가중치를 사용하는 이유는 중심성이 큰 값을 갖는 순서에 중요도를 부여하기 위해서이다. 작은 중심성을 갖는 순서는 그래프의 거리를 정의하는데 미미한 영향을 미칠 것이라고 가정한다. $W = \sum_{1}^{n} w_i$이면 하나의 중심성에 대한 대상 그래프의 거리를 (식 6)으로 정의한다.

$$\gamma_\delta = \frac{\sum_{i=1}^{n} I_i^w}{W} \quad \text{(식 6)}$$

이 과정을 각 중심성에 대해 수행하여 BC, CC, DC, EC에 대한 값들을 $\gamma_\delta^{BC}$, $\gamma_\delta^{CC}$, $\gamma_\delta^{DC}$, $\gamma_\delta^{EC}$라고 한다. 그러면, 이 값들을 이용하여 대상 그래프의 거리를 (식 7)로 정의할 수 있다.

$$distance(G^T, G^R) = (\gamma_\delta^{BC}, \gamma_\delta^{CC}, \gamma_\delta^{DC}, \gamma_\delta^{EC}) \quad \text{(식 7)}$$

(식 7)을 이용하여 두 그래프 $G^{T_1}$과 $G^{T_2}$의 유사도를 (식 8)과 같이 정의할 수 있다.

$$simliarity(G^{T_1}, G^{T_2}) = \| distance(G^{T_1}, G^{R_1}) - distance(G^{T_2}, G^{R_2}) \|_2 \quad \text{(식 8)}$$

여기서 $G^{R_1}$과 $G^{R_2}$는 $G^{T_1}$과 $G^{T_2}$와 각각 동일한 정점 수, 간선 수의 무작위 그래프 집합을 의미한다.

### 2.2 Flowchart and Pseudocode for Proposed Method

다음은 그래프 $G^T$와 무작위 그래프 집합 $G^R$의 거리인 distance($G^T$, $G^R$)를 구하는 흐름도이다.

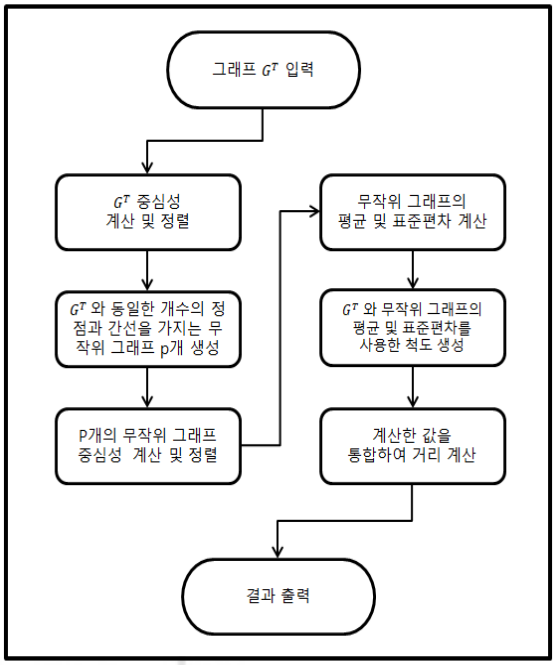


Fig. 6. Flowchart for calculating distances

function compGamma($G^T$, q)

$G^T$ is the target graph $G^T = (V^T, E^T)$

$|V^T| = n,\ |E^T| = m$

$G^R$ is a set of random graphs $G^R = \{G^{r_1}, \dots, G^{r_p}\}$

for $i = 1, \dots, n,\ G^{r_i} = (V^{r_i}, E^{r_i}),\ |V^{r_i}| = n,\ |E^{r_i}| = m$

$\delta \in R$

$C^t = (c_i^t, i = 1, \dots, n) \in R^n$

$C^r = (c_{i,j}^r, i = 1, \dots, n, j = 1, \dots, p) \in R^{n \times p}$

$W = \sum_i^n w_i \in R$

$C^t \leftarrow$ sorted centrality of $G^t$ in nonincreasing order

for $k = 1$ to $p$

  generate $G^{r_k}$

  $C_{.k}^r \leftarrow$ sorted centrality of $G^{r_k}$ in nonincreasing order

for $i = 1$ to $n$

  $avg_{c_i} = average(c_{i,1}^r, \dots, c_{i,p}^r)$

  $sd_{c_i} = standartDeviation(c_{i,1}^r, \dots, c_{i,p}^r)$

for $i = 1$ to $n$

  if $\left| avg_{c_i} - c_i^t \right| / sd_{c_i} > \delta$

    $I_i^w \leftarrow 1 \times w_i$

  else

    $I_i^w \leftarrow 0 \times w_i$

$\gamma_\delta \leftarrow \sum_{i=1}^n I_i^w \ / \ W$

return $\gamma_\delta$=distance between $G^T$ and $G^R$

다음은 distance($G^T$, $G^R$) 벡터를 구성하는 특정 중심성 q에 대한 값 $\gamma_\delta$를 계산하는 함수 compGamma의 알고리즘 의사코드이다.

### 3. Validity Check for Each Centralities

대상 그래프로 실제 존재하는 네트워크 데이터 'Karate' ($n$ = 34, $m$ = 78)를 선정하여 실험한다[14]. 먼저, BC, CC, DC, EC를 계산한다. 무작위 그래프의 개수를 100, 500, 1,000, 1,500, 2,000, 2,500개로 하고 각각의 개수마다 10번씩 수행하였다. 무작위 그래프의 표본이 많을수록 더욱 안정된 수치를 얻을 수 있다. 하지만 표본수가 커질수록 수행시간이 비례하여 증가하게 된다. 따라서 무작위 그래프의 개수를 다르게 실험하여 안정적인 결과를 도출하되 그래프 개수를 최소화하여 수행시간을 줄이도록 한다. 10번의 시행동안 동일한 값이 나온 횟수를 중심성의 척도를 유효하게 얻어냈다고 가정한다.

먼저, BC는 100, 500, 1,000, 1,500개의 무작위 그래프를 각각 생성하여 중심성과 비교하였을 때 최빈값은 동일하였지만, 10번의 시행횟수 중 6번의 시행에서는 0.9899의 값을, 나머지 4번의 시행에서는 0.9832의 값이 측정된 경우가 다수 존재하였다. 1,500개 이하의 무작위 그래프를 생성하여 비교하는 경우에는 안정적인 척도를 얻어냈다고 할 수 없다. 2,000개의 무작위 그래프를 생성하여 척도를 측정한 경우는 1개의 값을 제외한 모든 값이 0.9899로 동일하게 나타났으며 2,500개인 경우는 모든 값이 0.9899로 통일되었다. 비교대상으로 사용한 Karate의 $\gamma_\delta^{BC}$ = 0.9899값을 가진다. 또한, BC의 경우는 2,500개 이상의 무작위 그래프 개수만큼 생성하여 비교할 때 안정된 값을 얻을 수 있었다.

CC와 DC의 경우는 100개의 개수로 무작위 그래프를 생성하여 비교하였을 때는 불안정한 값을 가졌으나 500개 이상의 무작위 그래프 개수를 통해 생성하였을 때는 모두 동일한 결과값을 얻어낼 수 있었다. $\gamma_\delta^{CC}$ = 0.7126의 값을 얻어낼 수 있었으며, $\gamma_\delta^{DC}$ = 0.7613의 값이 나타났다.

EC의 경우는 각각의 시행횟수나 무작위 그래프의 개수에 관계없이 일관성 없는 값이 측정되었으며 그래프의 특징이 되는 값을 얻어낼 수 없었다. 그래서 최종 실험에서는 EC를 제외하였다.

## IV. Experimental Result

실험에서는 무작위 그래프의 개수를 BC에서 2,500개, CC와 DC에서는 500개씩 생성하여 실험하였다. 척도에 사용되는 $\delta$는 무작위 그래프와 비교대상이 되는 그래프의 차이를 판단하는 기준이 된다. $\delta$의 기준에 따라 비교범위가 설정되며 이 범위를 넘어갈 경우 유사하지 않는 값을 가지고 범위 내의 값을 가질 경우는 유사하다고 표현한다. 본 논문에서는 $\delta$값을 2로 설정하였고[2], 척도에 중심성의 차이를 나타내기 위해 계산에서 사용된 가중치 $(w_1, \dots, w_n)$는 1부터 그래프 $G^T$의 정점의 개수인 $n$의 값을 역순으로 부여하였다. 즉, $(w_1, \dots, w_n) = (n, \dots, 1)$로 설정하여 실험을 수행하였다.

[실험 1]은 그래프 크기와 관계없이 특정한 성질을 가지는 Path 그래프, Star 그래프, Paley 그래프, Ring 그래프를 제안하는 방법을 통해 무작위 그래프와의 거리, 즉 $distance(G^T, G^R) = (\gamma_\delta^{BC}, \gamma_\delta^{CC}, \gamma_\delta^{DC})$를 계산하여 비교하였다. 각각의 중심성의 거리를 벡터로 3차원 그래프에 표현함으로써 그래프간의 거리를 보다 쉽게 확인할 수 있다.

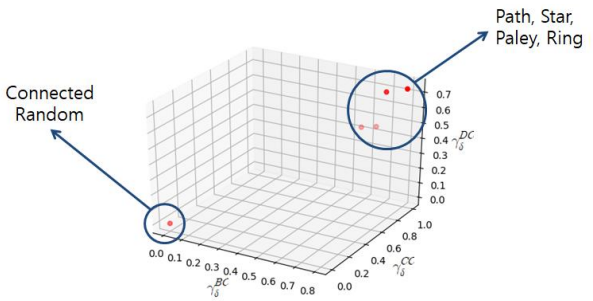


Fig. 7. Distances of [Experiment 1]

먼저 Fig.7은 특성이 다른 Path 그래프, Star 그래프, Paley 그래프, Ring 그래프와 connected 무작위 그래프의 거리를 측정한 결과이다. 제안하는 방법의 척도는 connected 무작위 그래프와의 거리를 기준으로 범위를 벗어나는지 비교하여 측정하기 때문에 하나의 connected 무작위 그래프는 모든 지표가 범위를 벗어나지 않아 모든 값이 0에 가까운 값이다. 나머지 4개의 그래프는 connected 무작위 그래프와 비교하였을 때 거리가 먼 것으로 나타났으며, 4개의 그래프가 상대적으로 가깝게 위치하고 있다는 것을 확인할 수 있다.

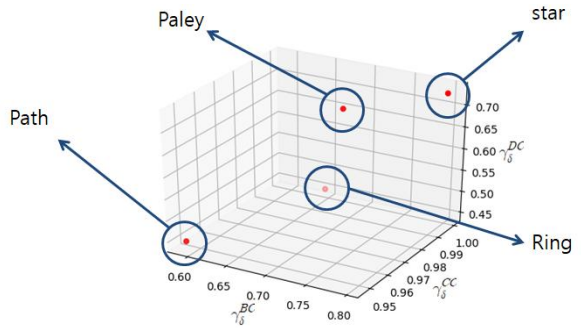


Fig. 8. Distances of each specific property graphs

Fig.8은 특성이 다른 4개의 그래프를 확대한 그림이다. Path 그래프와 Star 그래프는 다른 그래프들에 비해 더 많은 차이를 보이고 있으므로 유사한 특성이 적은 관계임을 확인할 수 있다.

Table. 2의 각 열은 Fig. 8에 표기된 그래프들의 distance 벡터 값이다.

Table 2. Distance of Fig. 8

| | Path | Star | Paley | Ring |
|---|---|---|---|---|
| $\gamma_\delta^{BC}$ | 0.6454 | 0.7983 | 0.6968 | 0.5890 |
| $\gamma_\delta^{CC}$ | 1 | 1 | 0.9871 | 0.9470 |
| $\gamma_\delta^{DC}$ | 0.4487 | 0.7294 | 0.6903 | 0.4489 |

유사함을 확인할 때 비교대상 그래프들의 거리를 정확히 표현하기 위해 Table. 2의 척도 값을 이용하여 Ring 그래프를 기준으로 Path 그래프, Star 그래프, Paley 그래프와의 유사도, 즉 $simliarity(Ring, G^*) = \|distance(Ring, G^{R_1}) - distance(G^*, G^{R_2})\|_2$ 를 계산한 결과이다.

Table 3. Similarity with Ring graph

| | Path | Star | Paley |
|---|---|---|---|
| $simliarity(Ring, G^*)$ | 0.0774 | 0.3540 | 0.2674 |

Ring 그래프와 유사한 순서는 Path 그래프, Paley 그래프, Star 그래프 순이다.

[실험 2]는 실제 네트워크 데이터들을 가지고 거리를 계산하였고, Table. 4는 대상 그래프의 데이터들이다[14][15][16].

Table 4. Target graph data sets

| Data Set | Explanation |
|---|---|
| Karate $(G_K^T)$ | 1970년대 미국 대학교에서 Karate club에 속해있는 34명의 소셜 네트워크 (n = 34, m = 78) |
| Dolphins $(G_D^T)$ | 뉴질랜드의 Doubtful Sound 지역에 살고 있는 62마리의 돌고래들 상이의 소셜 네트워크 (n = 62, m = 159) |
| Neural network $(G_N^T)$ | C. Elegans의 가중치와 방향성을 가지는 신경 네트워크 (n = 297, m = 2359) |
| Football $(G_F^T)$ | 2000년 가을 정규 학기 동안 Division IA 대학 간의 미식축구 게임 네트워크 (n = 115, m = 613) |
| Books about US politics $(G_B^T)$ | 2004년 대통령 선거를 앞두고 온라인 서점 Amazon.com에서 판매 한 미국 정치에 관한 책 네트워크 (n = 105, m = 441) |

Table. 5는 제안하는 방법을 통해 대상 그래프들에 대한 실험을 수행하여 얻은 거리(distance) 벡터를 나타낸다. 각 그래프들은 서로 다른 정점 수, 간선 수를 갖고 있지만 각 그래프에 대해 그에 대응하는 무작위 그래프를 구성할 때, 대상 그래프와 동일한 수의 정점 수, 간선 수를 갖도록 하는 방법을 사용하여 대상 그래프들의 무작위 그래프와의 거리를 계산하였다. 따라서 서로 다른 그래프들의 비교가 가능해진다.

Table 5. The distances between random graph set

| | $G_K^T$ | $G_D^T$ | $G_N^T$ | $G_F^T$ | $G_B^T$ |
|---|---|---|---|---|---|
| $\gamma_\delta^{BC}$ | 0.9899 | 0.6528 | 0.9902 | 0.8033 | 0.9621 |
| $\gamma_\delta^{CC}$ | 0.7126 | 1 | 0.5333 | 0.9999 | 1 |
| $\gamma_\delta^{DC}$ | 0.7613 | 0.5853 | 0.9203 | 0.7954 | 0.8996 |

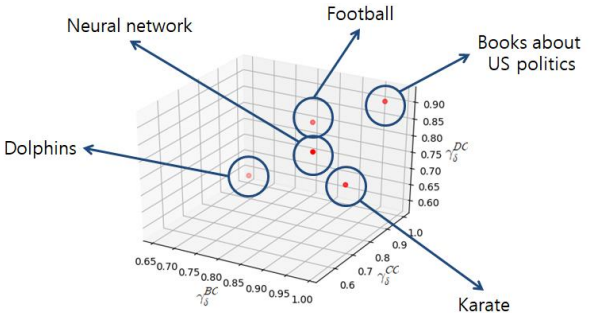


Fig. 9. Distances of [Experiment 2]

이 그림은 비교하는 대상 그래프들의 3가지 지표를 각각의 축으로 하여 3차원 그래프에 표시한 결과이다. Table. 6은 $G_K^T$를 기준으로 유사도를 계산한 결과이다.

Table 6. Similarity with $G_K^T$

| | $G_D^T$ | $G_N^T$ | $G_F^T$ | $G_B^T$ |
|---|---|---|---|---|
| $simliarity(G_K^T, G_*^T)$ | 0.4767 | 0.2396 | 0.3443 | 0.3201 |

$G_N^T$이 가장 유사한 성질을 가졌으며 $G_D^T$가 가장 유사하지 않은 성질을 가진 것을 알 수 있다.

## V. Conclusion

본 논문에서는 비교하고자 하는 그래프의 중심성과 무작위 그래프의 중심성을 이용하여 대상 그래프와 무작위 그래프 간의 거리를 정의하였고, 이 거리 벡터 정보를 이용하여 그래프간의 유사도를 정의할 수 있는 방법을 제안하였다. 또한, 제시한 방법을 통해 서로 다른 정점 수나 간선 수를 가지는 그래프들의 유사도를 측정할 수 있음을 보였다. 이를 통해 그래프의 특성을 파악하고 유사한 특징이나 패턴을 가지는 그래프별로 분류할 수 있다. 본 논문에서 제안한 방법의 경우, 대규모의 네트워크의 유사도를 측정하기 위해서 네트워크와 동일한 정점과 간선을 가지는 무작위 그래프를 일정 개수만큼 생성하여 비교하여야 한다. 따라서 많은 시간이 소요되어 대규모의 네트워크 환경에 적용은 제한적이다.

향후에는 두 개의 그래프의 유사도를 측정하는 것에서 나아가 그래프 집합간의 비교를 통해 군의 특성을 파악, 분석, 비교를 하고자 한다. 이를 통해 비교하고자 하는 대상 그래프들의 집합의 유사도 측정 및 분류가 가능할 것으로 보이며 생성방법이 다른 무작위 그래프들을 집합으로 생성하여 무작위 그래프 간의 특성을 파악하고 성질에 따라 분류할 수 있는 방법에 대해 연구할 계획이다.

## REFERENCES


[1] 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.

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

[3] Newman, Mark. Networks. Oxford university press, 2018.

[4] Roy, Matthieu, Stefan Schmid, and Gilles Tredan. "Modeling and measuring graph similarity: The case for centrality distance." Proceedings of the 10th ACM international workshop on Foundations of mobile computing. ACM, 2014.

[5] Mheich, A., et al. "A novel algorithm for measuring graph similarity: application to brain networks." Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on. IEEE, 2015.

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

[7] Sabidussi, Gert. "The centrality index of a graph." Psychometrika 31.4 (1966): 581-603.

[8] Havel, Václav. "A remark on the existence of finite graphs." Casopis Pest. Mat. 80 (1955): 477-480.

[9] Ronqui, José Ricardo Furlan, and Gonzalo Travieso. "Analyzing complex networks through correlations in centrality measurements." Journal of Statistical Mechanics: Theory and Experiment 2015.5 (2015): P05030.

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

[11] Albert, Réka, and Albert-László Barabási. "Statistical mechanics of complex networks." Reviews of modern physics 74.1 (2002): 47.

[12] Erdős, Paul, and Alfréd Rényi. "Asymmetric graphs." Acta Mathematica Academiae Scientiarum Hungarica 14.3-4 (1963): 295-315.

[13] Prüfer, H. (1918). "Neuer Beweis eines Satzes über Permutationen". Arch. Math. Phys. 27: 742–744.

[14] Newman, Mark EJ, and Michelle Girvan. "Finding and evaluating community structure in networks." Physical review E 69.2 (2004): 026113.

[15] Lusseau, David. "The emergent properties of a dolphin social network." Proceedings of the Royal Society of London B: Biological Sciences 270.Suppl 2 (2003): S186-S188.

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

## 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.