Journal of The Korea Society of Computer and Information
Vol. 21 No. 1, pp. 25-31, January 2016

www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2016.21.1.025

Analysis of Effect of an Additional Edge on Eigenvector Centrality of Graph

Chi-Geun Han*, Sang-Hoon Lee**

Abstract

There are many methods to describe the importance of a node, centrality, in a graph. In this paper, we focus on the eigenvector centrality. In this paper, an analytical method to estimate the difference of centrality with an additional edge in a graph is proposed. In order to validate the analytical method to estimate the centrality, two problems, to decide an additional edge that maximizes the difference of all centralities of all nodes in the graph and to decide an additional edge that maximizes the centrality of a specific node, are solved using three kinds of random graphs and the results of the estimated edge and observed edge are compared. Though the estimated centrality difference is slightly different from the observed real centrality in some cases, it is shown that the proposed method is effective to estimate the centrality difference with a short running time.

▸ Keyword : graph, eigenvector centrality

I. Introduction

그래프는 노드와 에지의 집합으로 표현되는데, 최근 폭발적인 사용을 보이고 있는 사회관계망은 하나의 거대한 그래프로 나타낼 수 있다. 사회관계망의 각 참여자는 하나의 노드로, 두 참여자 간의 관계는 하나의 에지로 그래프 내에 표현된다. 사회관계 망에서 중요한 역할을 하는 주체를 확인하고자 하는 의도는 그래프의 중심성(centrality) 연구로 발전하였고, 중심성을 정의하는 다양한 방법이 제시되었다[1]. 사회관계망 뿐만 아니라 연구논문들의 참고문헌 관계, 영화와 주연 배우 들의 관계들로 그래프로 표현되고, 그래프 내에서 주요한 역할을 하는 주체들을 찾는 다양한 연구들이 있다.

그래프에서 하나의 노드가 차지하는 비중(중요도, 중심성)을 나타내기 위해 상대적인 중요도를 정의하여 중심성이라고 한다. 중심성을 정의하는 방법으로는 연결도(degree) 중심성, 밀접도(closeness) 중심성, 고유 벡터 중심성, betweenness 중심성, Lin's 중심성, Harmonic 중심성, Seeley's 중심성, Katz's 중심성, PageRank, HITS, SALSA 등 다양한 중심성 지표들이 존재한다[1]. 순위(ranking)는 중심성이 큰 순서에 따라 노드에 부여한 자연수 값이다. 중심성이 클수록 순위는 작은 자연수 값을 갖는다.

본 연구에서는 주어진 그래프 $G=(V,E)$, $V=\{v_1,v_2,...,v_n\}$:노드의 집합, E:에지의 집합, 에서 각 노드에 대한 고유 벡터 중심성 $c\in R^n$을 얻은 후, 추가 에지 한 개에 의한 각 노드의 새로운 고유 벡터 중심성 $c'\in R^n$를 계산하는 방법을 연구한다. 단순한 방법으로는 에지 $e\notin E$가 추가된 그래프 $G'=(V,E')$, $E'=E\cup\{e\}$에 대해 c'를 다시 구하는 방법이다. 이 방법은 eigensystem 을 다시 풀어야 하므로, 그래프의 노드 개수 n이 클 경우 계산시간이 많이 소요되는 문제점을 갖고 있다.

추가 에지의 의미는 현재의 관계상으로는 없었던 새로운 관계가 추가되었음을 나타내는데, 웹시스템에서 각 홈페이지의 고유 벡터 중심성을 가장 많이 교란시킬 수 있는 즉, 링크스패밍 (link spamming)에 취약한 추가적인 에지를 파악하기 위한 방법이 될 수 있다[2]. 그리고 하나의 커뮤니티 내에서 하나의 객체가 확보하고 있는 순위(중심성 크기순서)를 최대로 감소(중심성 증가)시키기 위해서 연결해야할 에지를 미리 파악하는 방법으로도 활용할 수 있다[3].

본 연구에서는 연결된 무방향 단순(connected undirected simple) 그래프를 가정한다. 2장에서는 관련 연구를 설명하고, 3장에서는 에지 추가에 따른 모든 노드의 중심성의 변화량 추정방법, 에지 추가에 따른 각 노드의 중심성 변화량 추정방법을

• First Author: Chi-Geun Han, Corresponding Author: Chi-Geun Han
*Chi-Geun Han(cghan@khu.ac.kr), Dept. of Computer Science, Kyung Hee University
**Sang-Hoon Lee (a01b01c01@khu.ac.kr), Dept. of Computer Science, Kyung Hee University
• Received: 2015. 08. 07, Revised: 2015. 09. 17, Accepted: 2016. 01. 06.

설명한다. 4장에서는 전체 노드의 중심성을 가장 많이 변화시키는 추가 에지 결정 문제(문제 1), 자신의 중심성을 최대로 증가시킬 수 있는 노드별 추가 에지 결정 문제(문제 2)에 대해 추정치와 실제값과의 비교를 통해 추정 방법의 유효성을 검증하고, 5장에서 결론을 맺는다.

II. Related works

그래프의 중심에 있다는 것은 여러 가지 성질을 갖게 되는데, (1) 다른 주체들과 연결된 관계가 많다. (2) 다른 주체들로의 최단 거리의 합이 짧다. (3) 주체들 간의 최단경로에 많이 사용 된다. (4) 주체에 직접 연결된 주체들뿐만 아니라, 주변 주체들의 각 중요도가 높다라는 성질이 있다. 이 들 성질이 높은 정도를 수치화 하여 각각 (1) 연결도(degree) 중심성, (2) 밀접도(closeness) 중심성, (3) betweenness 중심성, (4) 고유 벡터 중심성을 정의한다. 이 외에도 다양한 중심성 지표들이 있는데, 일부를 나열해 보면, Lin's 중심성, Harmonic 중심성, Seeley's 중심성, Katz's 중심성, PageRank, HITS, SALSA 등이 있다[1][4].

본 연구에서는 고유 벡터 중심성이 추가 에지에 대해 변화하는 정도를 추정하는 방법에 대한 것이다. 고유 벡터 중심성을 선택한 이유는 고유 벡터 중심성이 eigensystem을 활용하는 spectral measure의 가장 기본적인 중심성이고, 그래프의 변화에 따른 중심성의 민감도(sensitivity)에 관한 연구가 고유 벡터를 구하는 방법에 대해서는 많이 진행되어 있기 때문이다. 가독성을 증가시키고 어휘의 단순화를 위해 지금부터 본 논문에서는 의미의 혼돈이 없는 한 고유 벡터 중심성을 단순히 중심성으로만 표시하기로 한다.

본 연구는 다음의 단순한 문제에서 출발한다. [문제 1] 전체 노드의 중심성을 가장 많이 변화시키는 추가 에지는 무엇인가? [문제 2] 특정 노드의 중심성을 가장 많이 향상시키는 추가 에지는 무엇인가? 예를 들어, 현재의 주체들의 중심성을 교란시키기 위해 추가 에지를 하나 만든다면, 어느 에지를 추가하여야 하는가? 반대로 해석하면, 그 추가 에지는 현재의 안정된 중심성(순위)을 혼란스럽게 할 수 있는 현재 커뮤니티의 가장 취약한 연결고리가 될 것이다. 그리고, 현재 커뮤니티 내의 나의 중심성을 가장 많이 증가시키기 위해 내가 누구와 연결되어야 하는가? 그 추가 에지를 통해 커뮤니티 내의 나의 영향력을 극대화 시킬 수 있을 것이다. [문제 1,2]에 대한 답을 구하기 위해서는 추가 가능한 모든 에지 $e \notin E$에 대해 $e \notin E$가 추가된 그래프 $G' = (V, E')$, $E' = E \cup \{e\}$를 생성한 후, 각 노드의 중심성을 구한 후 각각의 수치를 최대로 하는 에지를 선정하면 된다. 그러나, 이러한 방법을 활용하게 되면, 각 생성 가능한 그래프별로 eigensystem을 풀어야 하므로, 답을 얻기 위해서는 많은 시간이 필요하게 된다. 본 연구는 이들 답을 얻기 위해 완전한 답은 아니지만, 근사적인 해를 쉽게 얻을 수 있는 분석적 방법을 제시한다.

유사한 연구를 살펴보면, 그래프의 일부 데이타가 불확실할 경우(노드/에지의 추가/삭제가 있는 경우) 연결도, 밀접도, 고유 벡터, betweenness 중심성이 어떻게 변화하는지를 분석한 연구가 있다[5]. Segarra와 Ribeiro는 연결도, 밀접도, 고유 벡터, betweenness 중심성을 이용하여 그래프의 일부 구성에서 변화가 있을 때 중심성이 어떻게 변화하는지를 분석하였다[6]. Chartier et al.은 n개의 팀이 모여 리그로 게임을 하여 나온 결과를 이용한 팀 순위의 결정에서 일부 게임의 승패가 달라졌을 때, 팀의 순위에 어떤 영향을 주는지를 선형대수 기반인 Colley, Massey, Markov 방법에 따라 분석했다[7]. Ng et al.은 그래프의 일부 연결정보가 달라졌을 때, 노드의 순위에 어떤 영향을 주는지 HITS와 PageRank 방법에 대해 연구하였다[8]. Avrachenkov와 Litvak는 웹페이지간의 관계에서 새로운 링크가 추가되었을 때 PageRank 방법으로 순위를 정하는 경우 새로운 링크가 전체 순위에 미치는 영향을 분석하였다[3]. Correa와 Ma는 그래프에서 하나의 에지의 가중치(wieght)가 달라짐에 따라 각 노드의 중심성이 변화하는 정도를 시각화하는 방법을 고유 벡터 중심성, 마르코프(Markov) 중심성에 대해 설명하였다[9].

[9]는 한 에지의 가중치의 작은 변화에 따른 각 노드의 중심성 변화를 관찰하였고, 본 연구에서는 하나의 추가 에지에 따른 각 노드의 중심성 변화량을 추정하는 방법을 제시한 점에 차별성이 있다.

III. The Proposed Methods

본 장에서는 먼저 고유 벡터 중심성을 예를 들어 설명하고, 본 논문에서 해결하고자 하는 문제를 정의한 후 그 문제들을 해결할 수 있는 분석적 방법을 제시한다. 그리고 분석적 방법에 내재하고 있는 오차에 대해 설명한다.

3.1 Eigenvector centrality

그래프를 인접행렬(adjacency matrix)로 표현하고, 그 행렬의 최대 고유값(eigenvalue)을 구한 후, 그 특성값에 해당하는 고유 벡터(eigenvector)를 찾아 해당하는 노드에 대응시켜서 구할 수 있다. [그림 1]의 각 노드의 고유 벡터 중심성은 다음 [표 1]과 같다.

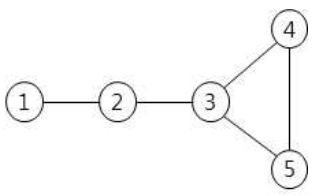

Fig. 1. A graph example

Table 1. Eigenvector centrality of each node in [Fig. 1]

node	1	2	3	4	5
centrality	0.15	0.34	0.60	0.50	0.50

[표 1]에서 노드 3은 5개의 노드 중 가장 큰 중심성을 갖고 있고, 이는 그래프 상에서 가장 중심에 있다는 것을 의미한다. 그리고, 노드 4와 노드 5는 그래프 상에서 같은 위치에 있으므로, 같은 중심성 값을 갖고 있다.

3.2 Problems

본 논문은 다음 두 문제에 대한 답을 효과적으로 찾기 위한 분석적 방법을 제안하는 것이 주목표이다. 그래프 예에 대한 두 문제의 의미와 답을 설명한다.

[문제 1] 전체 노드의 중심성을 가장 많이 변화시키는 추가 에지는 무엇인가?

[그림 1] 그래프에서 추가할 수 있는 에지는 총 5종류가 있다. 각 추가 에지별 전체 노드의 중심성 변화량의 제곱값 $f=\sum_{i=1}^{5}(c_i^{'}-c_i)^2$은 다음과 같다. 여기서, $c=\{c_i\}\in \mathrm{R}^n$는 그래프 G에 있는 노드들의 중심성 벡터이고., $c^{'}=\{c^{'}_i\}\in \mathrm{R}^n$는 추가 에지가 포함된 그래프 $G^{'}$의 중심성 벡터이다.

Table 2. Square of eigenvector centrality difference for each additional edge in [Fig. 1]

additional edge	(1,3)	(1,4)	(1,5)	(2,4)	(2,5)
difference f	0.08	0.05	0.05	0.03	0.03

[표 2]에서 에지 (1,3)이 그래프에 추가되었을 때 모든 노드에 주는 영향이 가장 큰 것을 알 수 있고, 에지 (2,4), (2,5)는 중심성의 변화에 가장 작은 영향을 주는 것을 관찰할 수 있다.

[문제 2] 특정 노드의 중심성을 가장 많이 향상시키는 추가 에지는 무엇인가?

노드 4에 대해 [문제 2]의 답을 찾아본다. 노드 4에는 추가 에지 (4,1), (4,2)가 가능하다. (4,1)이 추가되었을 때 노드 4의 중심성은 0.530이 되고, (4,2)가 추가되었을 때, 0.537이 된다. 따라서, 노드 4의 입장에서는 (4,2)가 추가되는 것이 자신의 중심성을 극대화시킬 수 있다.

노드 수가 많은 그래프의 문제에 대해서 추가 가능한 모든 에지들을 생성한 후 중심성이 어떻게 변화하는지를 확인하는 방법은 추가 가능한 에지 한 개를 추가할 때 마다 eigensystem을 풀어야 하므로 많은 시간이 소요된다. 따라서, 다음 두 절에서는 추가 에지 별로 새로운 그래프를 생성하지 않고, [문제 1,2]를 효과적으로 해결할 수 있는 분석적 방법을 제시한다.

3.3 Estimation method for total centrality difference with each additional edge

[문제 1]은 전체 노드의 중심성을 가장 많이 변화시키는 추가 에지를 찾는 것이다. 본 절에서는 이 문제를 효과적으로 해결할 수 있는 분석적 방법을 제시한다.

주어진 그래프 $G=(V,E),|V|=n,|E|=m$에 대해 인접행렬 A를 구할 수 있다. $Ax=\lambda x$를 풀어 얻어진 λ를 A의 고유값 중 가장 큰 값이라 하고, x를 λ에 대한 고유 벡터라 하면 (식 1)을 얻을 수 있다. I는 항등(identity)행렬.

$$(A-\lambda I)x=0 \qquad \text{(식 1)}$$

$Q=A-\lambda I$라 하고, d_i를 노드 i의 연결도(degree), $t_{ij}=\frac{d_i+d_j}{2}$로 정의하고, t_{ij}는 이산변수이지만, 연속변수라 가정하면, (식 1)로 부터

$$\frac{\partial Q}{\partial t_{ij}}x+Q\frac{\partial x}{\partial t_{ij}}=0, \qquad \text{(식 2)}$$

여기서 $\frac{\partial Q}{\partial t_{ij}}=\frac{\partial A}{\partial t_{ij}}-\frac{\partial \lambda}{\partial t_{ij}}I$. (식 2)로 부터

$$Q\frac{\partial x}{\partial t_{ij}}=-\frac{\partial Q}{\partial t_{ij}}x \qquad \text{(식 3)}$$

(식 3)을 풀면 t_{ij}값의 변화에 따른 고유 벡터 값의 변화량, 즉 $\frac{\partial x}{\partial t_{ij}}$를 구할 수 있다. Q는 이미 알고 있는 행렬, $\frac{\partial x}{\partial t_{ij}}$는 미지수 벡터, $-\frac{\partial Q}{\partial t_{ij}}x$는 알고 있는 벡터이므로, $By=d, B\in R^{n\times n}$, $y,d\in R^n$ 형태의 선형연립방정식이 된다. 여기서 Q는 nonsingular 일 수 있으므로 역행렬을 직접 구할 수 없으므로, (식 3)을 풀기 위해 singular value decomposition (SVD) 방법을 사용한다. 즉, Q의 SVD를 통해 $Q=U\Sigma V^T$, $U,V\in R^{n\times n}$ orthonormal 행렬, $\Sigma=diag(\sigma_1,....,\sigma_n)$, $\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_n$를 얻는다. (식 3)의 해는 다음 (식 4)로 표시될 수 있다.

$$\frac{\partial x}{\partial t_{ij}}= -Q^{+}\frac{\partial Q}{\partial t_{ij}}x \qquad \text{(식 4)}$$

$Q^{+}=V\Sigma^{+}U^T$, $\Sigma^{+}=diag(\sigma_1^{-1},....,\sigma_r^{-1},0,...,0)\in R^{n\times n}$, $r=rank(Q)$.

$f_{ij}=\|\frac{\partial x}{\partial t_{ij}}\|_2^2$ 라 하고, f_{ij}값을 최대로 하는 $(i,j)\notin E$를 구하면, 현재의 중심성을 가장 크게 변화시키는 추가 에지를 찾을 수 있다. 즉, [문제 1]은 $F=MAX_{(i,j)\notin E}f_{ij}$인 에지 (i,j)를 찾는 문제가 된다.

(식 4)에 의해 $f_{ij}=\|\frac{\partial x}{\partial t_{ij}}\|_2^2=\|Q^{+}\frac{\partial Q}{\partial t_{ij}}x\|_2^2$

$=(Q^{+}\frac{\partial Q}{\partial t_{ij}}x)^TQ^{+}\frac{\partial Q}{\partial t_{ij}}x=x^T\frac{\partial Q}{\partial t_{ij}}(Q^{+})^2\frac{\partial Q}{\partial t_{ij}}x$

여기에 $Q^{+}=V\Sigma^{+}U^T$를 대입하고 정리하면,

$$f_{ij}=x^T\frac{\partial Q}{\partial t_{ij}}U(\Sigma^{+})^2U^T\frac{\partial Q}{\partial t_{ij}}x. \qquad \text{(식 5)}$$

f_{ij}를 계산하기 위해서는 $\frac{\partial Q}{\partial t_{ij}}=\frac{\partial A}{\partial t_{ij}}-\frac{\partial \lambda}{\partial t_{ij}}I$를 알아야 하고, $\frac{\partial \lambda}{\partial t_{ij}}=x^T\frac{\partial A}{\partial t_{ij}}x$이므로[10], $\frac{\partial A}{\partial t_{ij}}$의 값을 알아야 한다. 그런데, 인접행렬 A가 노드의 연결도에 따라 이산적으로 바뀌므로, $\frac{\partial A}{\partial t_{ij}}$를 근사적으로 구할 수밖에 없다. 추가 에지 (i,j)에 의해 새로운 인접행렬을 A'라 하면, $A'=A+H$, $H=e_ie_j^T+e_je_i^T$, $e_i\in R^n$,i번째 component가 1인 단위벡터. 따라서, $\frac{\partial A}{\partial t_{ij}}\approx H$로 근사적으로 표시할 수 있다. $\frac{\partial \lambda}{\partial t_{ij}}=x^T\frac{\partial A}{\partial t_{ij}}x \approx x^THx$이므로,

$$\frac{\partial Q}{\partial t_{ij}}=\frac{\partial A}{\partial t_{ij}}-\frac{\partial \lambda}{\partial t_{ij}}I \approx H-x^THxI \qquad \text{(식 6)}$$

로 표시할 수 있다. 따라서, (식 5)와 (식 6)을 합쳐서 추정된 f_{ij}를 $\widetilde{f_{ij}}$로 표시하면,

$$\widetilde{f_{ij}}=x^T(H-x^THxI)U(\Sigma^+)^2U^T(H-x^THxI)x \qquad \text{(식 7)}$$

이 식을 정리하면 $\widetilde{f_{ij}}=x^THU(\Sigma^+)^2U^THx$가 된다. $P=U(\Sigma^+)^2U^T$로 표시하면, $H=e_ie_j^T+e_je_i^T$ 이므로,

$\widetilde{f_{ij}}=x_ip_{ji}x_j+x_ip_{jj}x_i+x_jp_{ii}x_j+x_jp_{ij}x_i$, $P=\{p_{ij}\}$. P는 대칭행렬이므로, 정리하면, (식 8)의 단순한 형태가 된다.

$$\widetilde{f_{ij}}=2x_ix_jp_{ij}+x_i^2p_{jj}+x_j^2p_{ii} \qquad \text{(식 8)}$$

주어진 그래프의 고유 벡터(중심성) x와 Q에 대한 SVD를 한 번 구하면, 추가 에지 (i,j)에 따른 각 노드의 중심성의 변화량 제곱의 합 $\widetilde{f_{ij}}$을 쉽게 구할 수 있다. (식 8)은 각 $(i,j)\notin E$에 대해 7회의 실수 곱셈이 필요하고, $(i,j)\notin E$의 개수는 $O(n^2)$이다. 그리고, P를 구하기 위해서는 Q에 대한 SVD를 한 번 구해야 하고, P의 연산을 위해 $n\times n$ 크기의 행렬 곱셈이 필요하다. 이 단계는 $O(n^3)$의 시간복잡도를 갖는다. 따라서, [문제 1]을 해결하기 위해 추가 가능한 에지 별로 추정값 $\widetilde{f_{ij}}$를 (식 8)을 이용해 계산하는 것은 총 $O(n^3+n^2)$의 시간복잡도를 갖는다. 반면, 실제값 f_{ij}^*를 구한다면 각 추가 에지에 대해 매번 eigensystem를 실제로 해결하여 이 문제를 해결하여야 한다. 한번의 eigensystem 해결에 $O(n^3)$의 시간이 필요하고, 추가 가능한 에지의 수가 $O(n^2)$이므로, 실제값을 구하는 방법의 시간복잡도는 $O(n^2\times n^3)=O(n^5)$이 된다. 이는 [문제 1]을 추정값을 구하는 방법으로 해결하는 제안한 방법이 실제값을 계산하여 해결하는 것 보다 문제를 매우 빠르게 해결한다는 것을 나타낸다.

3.4 Estimation for centrality difference of particular node with each additional edge

본 절에서는 [문제 2] 특정 노드의 중심성을 가장 많이 향상시키는 추가 에지는 무엇인가? 에 대한 답을 효과적으로 찾기 위한 분석적 방법을 설명한다.

에지 (i,j)가 추가되었을 때 노드 i의 중심성 증가량을 k_{ij}라 하면, (식 4)에 의해

$$k_{ij}=(\frac{\partial x}{\partial t_{ij}})^Te_i=-(Q^+\frac{\partial Q}{\partial t_{ij}}x)^Te_i \qquad \text{(식 9)}$$

$Q^+=V\Sigma^+U^T$이고 (식 6)에 의해 $\frac{\partial Q}{\partial t_{ij}}=\frac{\partial A}{\partial t_{ij}}-\frac{\partial \lambda}{\partial t_{ij}}I\approx H-x^THxI$ 이다. 이들을 (식 9)에 대입하고, U_i와 V_i를 U와 V의 i번째 행벡터를 나타낸다고 하면, 에지 (i,j)가 추가되었을 때 노드 i의 중심성 증가량에 대한 추정치 $\widetilde{k_{ij}}$는

$$\begin{aligned}\widetilde{k_{ij}} &= -(V\Sigma^+U^T(e_ie_j^T+e_je_i^T)x)^Te_i\\ &= -x^T(e_ie_j^T+e_je_i^T)U\Sigma^+V^Te_i\\ &= (-x_ie_j^T-x_je_i^T)U\Sigma^+V_i^Te_i\\ &= (-x_iU_j-x_jU_i)\Sigma^+V_i^Te_i \end{aligned} \qquad \text{(식 10)}$$

$\widetilde{k_{ij}}$는 Q의 SVD, $Q=U\Sigma V^T$ 정보와 현재 알고 있는 중심성 x의 정보를 갖고 쉽게 계산할 수 있다. 노드 i에 추가 에지 한 개를 연결하여, 노드 i의 중심성을 최대로 증가시키기 위해서는 $\widetilde{k_{ij}}$값을 최대로 하는 j를 선택하면 된다. 즉, $\widetilde{K_i}=MAX_j\ \{\widetilde{k_{ij}}\}$인 j를 결정하여, 에지 (i,j)를 추가하면, 노드 i의 중심성을 최대로 증가시킬 수 있다. [문제 2]를 제안한 방법으로 해결하는 알고리즘의 시간복잡도는 (식 10)의 $\widetilde{k_{ij}}$을 구하는데 필요한 시간복잡도가 되며, 이는 3.3절의 시간복잡도 분석과 유사하게 $O(n^3+n^2)$가 됨을 알 수 있다. 반면, 추가 가능한 에지에 대해 실제값 k_{ij}^*을 구한다면, 매번 eigensystem를 실제로 해결하여야 하고 ($O(n^3)$ 소요), 각 노드 당 $O(n)$개수의 연결 가능한 에지가 있으므로, 시간복잡도는 $O(n\times n^3)=O(n^4)$가 된다. 따라서, [문제 2]를 제안한 방법으로 해결하는 것이 실제값을 계산하여 해결하는 것 보다 시간적으로 매우 효과적인 것을 알 수 있다.

3.5 Error analysis

3.3절, 3.4절에서 설명한 분석적 방법은 두 종류의 오차를 내재하고 있다.

3.5.1 Error of differential equation

(식 1)으로부터 구한 (식 2)는 미분의 정의에 의해 영에 근접한 $t_{ij}=\frac{d_i+d_j}{2}$의 값에 대해서 유효하다. 즉, 연결도의 변화량이 작을 때에만 (식 2)는 영의 값을 갖게 된다. 연결도는 이산적인 성질을 갖고 있으므로, 작은 변화량을 만들 수 없다. 따라서, (식 6)에서 t_{ij}에 대한 A와 λ의 변화량을 근사적으로 처리하였으므로, (식 2)는 근사적으로 영의 값을 갖게 될 수밖에 없다.

3.5.2 Error of Least Squares problem

(식 4)에서 Q의 rank가 full rank를 갖는다면, (식 4)의 선

형방정식은 오차 없이 풀 수 있게 된다. 그러나, 그래프에 따라 $Q=A-\lambda I$는 full rank를 갖지 않는 경우가 존재하고, 이 경우, 선형방정식은 내재적인 오차를 갖게 된다. 그 오차는 $\sum_{i=r+1}^{n}(U_i^T b/\sigma_i)$가 된다. 여기서 $b=-\frac{\partial Q}{\partial t_{ij}}x$, r=rank(Q).

따라서, 3절에서 소개한 분석적인 방법은 피할 수 없는 내재적인 오차를 갖게 된다. 그러므로, 근사적인 값을 구하는 분석적 방법이 실제적으로 어느 정도의 정확성을 갖는지를 확인하는 절차가 필요하게 된다. 다음 장에서는 실험계산을 설명한다.

IV. Expermental Results

본 연구에서는 3장의 분석방법의 유효성을 검증하기 위해 실험계산을 수행하였다. 실험계산을 위해 3종류의 무작위 그래프 생성 방법을 사용하였는데, 세 종류 무작위 생성 그래프를 Erdös와 Rényi의 그래프 (E그래프)[11], Barabási와 Albert의 그래프 (B그래프)[12], Watts와 Strogatz의 그래프 (W그래프)[13]라 칭한다. 모든 그래프의 노드 개수는 30이고, 에지의 개수는 다음과 같다. 각 그래프, 에지 개수 별로 50개의 그래프를 생성하여 결과를 얻었다. 다음 [표 3]은 세 종류 무작위 생성 그래프에 사용된 파라미터와 에지의 개수를 보여 주고 있다.

Table 3. Graphs of experiments

	E graph		B graph		W graph	
	E1	E2	B1	B2	W1	W2
parameter	pr =20%	pr =26.89%	m_b= 3	m_b= 4	m_w =3	m_w =4
# of Edges	87	117	84	114	90	120

E그래프는 각 에지의 생성 확률 pr을 설정하여 생성한 그래프이다. B그래프는 생성되는 노드에 m_b개의 에지를 이미 생성된 노드에 무작위로 연결하는 방법으로 그래프를 생성하는데, 생성된 그래프는 scale-free 그래프가 된다. W그래프는 small-world 성질을 갖고 있는 노드의 평균 연결도가 m_w인 무작위 그래프이다. 생성 방법에 따라 생성된 에지의 개수가 비슷하도록 최대한 생성 파라미터를 조정하였다.

3장에서 설명한 추가 에지에 따른 중심성의 변화 추정값이 어느 정도 정확한지를 파악하기 위해 실제로 추가한 에지를 갖고 있는 그래프의 중심성을 구한다. 이렇게 구한 경우의 값에는 *를 표시하여 ~로 표시된 추정값과 구분한다.

3장에서 설명한 두 가지 문제를 포함하여, 총 4 종류의 평가항목을 정의한다. 각 평가항목은 실제로 측정한 값과 분석방법을 통한 추정치를 비교하는 방식으로 분석방법의 타당성을 검증한다. 4 종류의 평가항목과 그 결과는 다음과 같다.

(1) [문제 1] 추가 가능한 에지를 하나씩 모두 생성하여, 각 추가 에지 (i,j)별 f_{ij}^*를 구한 후, f_{ij}^*를 최대로 만드는 추가 에지 (u,v)를 찾을 수 있다. 실제로 에지 (i,j)를 그래프 G에 추가하여 만들어진 그래프 G'에 대해 구한 중심성 $c(G')$와 원래 그래프 G의 중심성 $c(G)$의 차의 제곱, 즉 $f_{ij}^*=\|c(G)-c(G')\|_2^2$, $G'=(V,E')$, $E'=E\cup\{(i,j)\}$, $(i,j)\notin E$를 구한 후, 그 값을 최대로 하는 F^*를 구한다. 즉, $F^*=MAX_{(i,j)\notin E}\{f_{ij}^*\}$. 그리고, 3.3절의 결과를 이용하여 추정치 $\widetilde{f_{ab}}$를 계산할 수 있다. $\widetilde{f_{ab}}=MAX_{(i,j)\notin E}\{\widetilde{f_{ij}}\}$이면, f_{ab}^*는 $\{f_{ij}^*,\ (i,j)\notin E\}$ 집합의 원소들을 내림차순으로 정렬하였을 때 몇 등에 해당하는지를 찾는다. 즉, 최대 변화를 가져올 것으로 추정된 추가 에지가 실제적으로 몇 번째의 큰 변화량을 가져오는지 확인하여, 분석적 방법의 정확성을 확인하고자 한다. 다음 [표 4]는 50개의 그래프에 대해 추정된 추가 에지가 실제적으로 몇 번째 내에 위치하고 있는지에 대한 누적결과를 보여 주고 있다.

Table 4. Results of [problem 1] for each graph

graph \ rank	1	2	3	4	5	6
E1	42	49	49	50	50	50
E2	45	49	50	50	50	50
B1	42	48	50	50	50	50
B2	44	47	48	49	50	50
W1	36	43	48	49	49	50
W2	46	50	50	50	50	50

[표 4]에 표시된 수자는 누적 그래프 수를 나타내는데, E2인 경우, 45개의 그래프에서 $\tilde{f}$의 값을 가장 크게 만드는 추가 에지 (a,b)가 실제 f^*의 값을 가장 크게 한 에지와 일치하였고, 4개의 그래프에서 두 번째로 큰 변화를 가져왔고, 1개의 그래프에서 세 번째 큰 변화를 주는 것을 나타낸다. 에지의 개수가 클수록 추정한 에지가 1등이 될 가능성이 높음을 알 수 있다. 각 그래프에서 추가 가능한 에지의 개수는 315개~351개이다. 모든 그래프에 대해, 추정된 변화량을 최대로 하는 에지는 실제로 변화량을 최대로 하는 에지 순서로 6등 이내의 에지인 것으로 확인되었으므로, 추정된 에지의 신뢰성이 있음을 보여 주고 있다.

(2) 실제로 에지를 하나 추가하여 구한 중심성의 변화량 제곱값과 추정한 중심성 변화량 제곱값 간의 비교를 위해 $Error_F=|F^*-\tilde{F}|/F^*$를 계산한다. $F^*=MAX_{(i,j)\notin E}\{f_{ij}^*\}$, $\tilde{F}=MAX_{(i,j)\notin E}\{\widetilde{f_{ij}}\}$. 50문제에 대한 평균은 다음 [표 5]와 같다.

Table 5. Error for each graph(%)

graph	E1	E2	B1	B2	W1	W2
$Error_F$	0.162	0.141	0.033	0.040	0.179	0.153

이 값은 실제로 가장 큰 변동을 가져오는 추가 에지 (u,v)에 대한 변화량 값과 가장 큰 변동을 가져 올 것으로 추정된 추가 에지 (a,b)에 대한 변동량과의 비교이다. 실제 같은 쌍의 에지에 대한 비교인 경우는 더 큰 오차를 보일 것이다. [표 5]에서 scale-free 성질을 갖는 B그래프에서 작은 오차를 보였지만, E그래프와 W그래프에서는 값의 오차가 큰 것을 확인할 수 있다. 이는 단순한 무작위 그래프보다 scale-free 그래프인 B그래프가 중심성이 뚜렷한 노드들을 갖고 있기 때문으로 설명할 수 있다.

(3) [문제 2] 노드 i에 대해 $k_{iq}^{*} = MAX_{j:(i,j)\notin E}\{k_{ij}^{*}\}$, $\widetilde{k_{ir}} = MAX_{j:(i,j)\notin E}\{\widetilde{k_{ij}}\}$일 때 prob(q=r)의 값을 계산한다. 즉, 실제 계산에 의한 노드 i의 중심성을 가장 많이 향상시키는 추가 에지 (i,q)와 가장 많이 향상시킬 것으로 추정된 에지 (i,r)의 일치 확률을 계산한다. 50개의 그래프에서 각 그래프는 30개의 노드를 갖고 있다. 따라서 총 1,500개의 노드에 대해 추정에지와 실제 에지가 일치한 확률을 다음 [표 6]에 표시했다.

Table 6. Probability of [problem 2] solution

graph	E1	E2	B1	B2	W1	W2
probability	0.965	0.969	0.992	0.997	0.947	0.958

[표 6]은 제안한 방법이 모든 그래프에서 높은 일치율을 보였고, 특히 B그래프인 경우 일치하는 확률이 가장 높다는 것을 보여 주고 있다. 이는 [표 5]에서 관찰한 바와 같이, B그래프의 중심성이 뚜렷한 특징으로부터 기인한다. 에지의 개수가 클수록 일치하는 확률이 근소한 차이로 높다는 것은 에지가 많은 그래프가 상대적으로 변화에 민감하지 않다는 것을 보여 주고 있다.

(4) 각 노드 u에 대해 실제 최대 증분을 가져오는 q를 찾는다. 그리고, 분석적 방법의 추정에서 최대 증분 노드 q를 찾지 못하는 경우와 맞추는 경우를 구분하여 오차 $Error_k = |k_{uq}^{*} - \widetilde{k_{uq}}| / k_{uq}^{*}$를 구한다. 오차에 대한 평균은 다음 [표 7]과 같다.

Table 7. Error of [problem 2] solution(%)

	E1	E2	B1	B2	W1	W2
$Error_k$ in finding successfully	4.63	4.48	4.02	3.89	5.54	5.23
$Error_k$ in finding unsuccessfully	7.92	5.91	23.53	17.16	8.91	6.74

[표 7]에서 그래프의 에지의 개수가 클수록 오차는 작아지는 것을 알 수 있다. 이는 [표 5, 6]에서 관찰한 바와 같이 에지가 많을수록 그래프가 안정화된다는 것을 의미한다. 그리고, B그래프에 대한 일치할 경우의 오차가 다른 그래프보다 작은 것은 [표 5]에서 관찰한 바와 같이 B그래프와 다른 그래프간의 중심성이 뚜렷한 정도로 설명가능하다. 그런데, B그래프가 불일치할 경우 오차가 매우 커지는 것은 매우 드문 경우에서 수치적 오류에 의한 잘못된 추정이 중심성이 뚜렷한 경우 그 오차가 커진다는 것을 보여 주고 있다.

[표 4]~[표 7]을 통해 scale-free 그래프인 B그래프에서 본 논문에서 제안한 분석적 방법이 작은 오차를 갖고, [문제 1,2]를 대부분의 경우 해결한다는 것을 알 수 있다.

V. Conclusions

본 논문에서는 추가 에지가 있는 경우 고유 벡터 중심성을 효과적으로 계산하는 분석적 방법을 제시하였다. 본 논문에서 제시한 방법을 이용하여 [문제 1,2]의 근사해를 효과적으로 찾을 수 있다는 것을 실험계산을 통해 보였다.

제안한 방법은 지속적으로 변화하는 커뮤니티의 구조에서 가장 취약한 연결고리를 실시간으로 파악하거나, 자신의 중심성을 극대화시킬 방법을 찾는 현실적인 방법이 될 수 있다. 이후 연구로 Eigensystem을 기반으로 하는 PageRank 방법[14] 등에 본 방법을 확장할 수 있는 방안을 모색하도록 한다. 또한 Bandwidth Minimization 문제[15]에 대해 추가 에지에 따른 노드 중심성 변화와 해의 관계를 연구할 계획이다.

REFERENCE

[1] P. Boldi and S. Vigna, “Axioms for Centrality”, Social and Information Networks, Nov. 2013.

[2] R. Lempel and S. Moran, “Rank-Stability and Rank-Similarity of Link-Based Web Ranking Algorithms in Authority-Connected Graphs”, Information Retrieval, Vol. 8, No. 2, pp 245-264, 2005.

[3] K. Avrachenkov and N. Litvak, “The Effect of New Links on Google Pagerank, Stochastic Models”, 22 (2), pp. 319-331, 2006.

[4] J. Ronqui and T. Gonzalo, “Analyzing Complex Networks Through Correlations in Centrality Measurements”, Social and Information Networks, June 2014.

[5] S. Borgatti, K. Carley and D. Krackhardt, “On the Robustness of Centrality Measures under Conditions of Imperfect Data”, Social Networks 28.2, pp.124-136, 2006.

[6] S. Segarra and A. Ribeiro, “Stability and Continuity of Centrality Measures in Weighted Graphs”, Social and

Information Networks, Oct. 2014.
[7] T. Chartier, E. Kreutzer, A. Langville, and K. Pedings, “Sensitivity and Stability of Ranking Vectors”, SIAM J. Sci. Comput., 33(3), pp. 1077-1102, 2011.
[8] A. Ng, A. Zheng and M. Jordan, “Stable Algorithms for Link Analysis”, SIGIR '01 Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 258-266, 2001.
[9] C. Correa and K. Ma, “Visual Reasoning about Social Networks using Centrality Sensitivity”, IEEE Transactions on Visualization and Computer Graphics, Vol. 18, No. 1, pp. 106-120, 2012.
[10] G. Golub and C. Loan, “*Matrix Computations*”, Johns Hopkins, 1988.
[11] P. Erdös and A. Rényi, “On Random Graphs, I”, Publicationes Mathematicae 6, pp. 290-297, 1959.
[12] A.-L. Barabási and R. Albert, “Emergence of Scaling in Random Networks”, Science, 286, pp. 509-512, 1997.
[13] D. Watts and S. Strogatz, "Collective Dynamics of 'Small-World' Networks", Nature 393 (6684), pp. 440-442. 1998.
[14] Woo-Key Lee, “Detecting Intentionally Biased Web Pages In terms of Hypertext Information”, Journal of The Korea Society of Computer and Information, Vol. 10, No. 1, pp.59-66, 2005.
[15] Sang-Un Lee, “Maximum Degree Vertex Central Located Algorithm for Bandwidth Minimization Problem”, Journal of The Korea Society of Computer and Information, Vol. 20, No. 7, pp.41-47, 2015.

Authors

Han Chi Geun received the B.S. and M.S. degrees in Industrial Engineering from Seoul National University and Ph.D. degree in Computer Science from Pennsylvania State University in 1983, 1985 and 1991, respectively.

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 at Kyung Hee University. He is interested in Graph Theory, Optimization, and Community Detection.

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.