

네트워크의 중첩 클러스터링 구조를 고려한 정보 확산 모델

(Cascading Behavior and Information Diffusion in
Overlapping Clusters)

이 우 중 [†]

(Woojung Lee)

황 지 영 ⁺⁺

(Joyce Jiyoung Whang)

요약 정보 전파(information diffusion) 연구란 네트워크에서 특정 노드(initial adopters)가 새로운 정보를 습득할 것으로 가정하고, 그로부터 다른 노드로 퍼지는 정보의 전파 양상(cascading behavior)에 대한 메커니즘을 모델링하는 것이다. 대부분의 기존 정보 전파 연구들은 노드가 하나의 클러스터(cluster)에만 속한다고 가정하며, 이 가정에 의해 클러스터가 정보 전파를 방해함을 보여 왔다. 하지만, 실제 네트워크에는 노드가 여러 클러스터에 속할 수 있다. 즉, 클러스터는 중첩될 수 있다. 본 연구는 네트워크에 중첩 클러스터(overlapping cluster)가 존재할 때의 정보 전파 양상에 대해 소개한다. 새로운 정보를 습득한 노드가 네트워크의 중첩 클러스터에 속하거나 네트워크의 각 노드에 대해 정보 공존성(compatibility)을 허용하였을 때 클러스터는 정보 전파의 방해요소가 아님을 보인다. 네 개의 실제 네트워크 데이터 셋(dataset)에서 본 연구에서 제시한 모델과 이론을 검증한다.

키워드: 정보 전파 양상, 정보 전파, 중첩 클러스터링, 공존성, 그래프 마이닝

Abstract Information diffusion models formulate and explain cascading behavior in networks where a small set of initial adopters is assumed to acquire new information and the new information is propagated to the other nodes in the network. Most existing information diffusion models assume that a node in a network belongs to only one cluster, and based on this assumption, it has been shown that clusters are obstacles to cascades. However, in many real-world networks, a node can belong to multiple clusters, i.e., clusters can overlap. In this paper, we study cascading behavior in a network when clusters overlap. We show that clusters are not obstacles to cascades if the initial adopters are placed in the overlapped region between the clusters or if we allow compatibility. We verify our theorems and models on four real-world datasets.

Keywords: cascading behavior, information diffusion, overlapping clustering, compatibility, graph mining

· 본 연구는 정부(과학기술정보통신부)의 지원으로 한국연구재단의 지원을 받아 수행된 사업임(NRF-2019R1C1C1008956, NRF-2018R1A5A1059921)

[†] 학생회원 : 성균관대학교 소프트웨어학과 학생

wj.lee@skku.edu

⁺⁺ 종신회원 : 성균관대학교 소프트웨어학과 교수
(Sungkyunkwan Univ.)

jjwhang@skku.edu

(Corresponding author)

논문접수 : 2019년 12월 23일

(Received 23 December 2019)

논문수정 : 2020년 2월 21일

(Revised 21 February 2020)

심사완료 : 2020년 2월 26일

(Accepted 26 February 2020)

Copyright©2020 한국정보과학회 : 개인 목적으로나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지 제47권 제4호(2020. 4)

1. 서 론

정보 전파(information diffusion) 연구는 노드(node)와 간선(edge)으로 이루어진 그래프 테이터에서 특정 노드(initial adopter)에 새로운 정보가 주어졌을 때 해당 정보의 확산 패턴에 대해 고찰한다. 정보 전파 연구는 간선으로 연결된 노드 간의 영향력을 고려하여 노드 개인이 정보를 전파할 수 있는 능력을 수학적으로 모델링하는 연구[1,2]와 노드가 주변 이웃노드(1-hop neighbor)에 영향을 받아 전파되는 정보의 흐름을 수학적으로 모델링하는 연구[3]로 나뉜다. 위 연구들은 노드가 하나의 클러스터(cluster)에만 속하는 것을 전제로 정보 전파 알고리즘을 개발하였다.

기존 정보 전파 연구의 가정과 다르게 소셜 네트워크와 같은 실제 네트워크에는 노드가 여러 클러스터에 속하는 중첩 클러스터(overlapping cluster)구조가 존재하며 네트워크의 중첩 클러스터 구조를 탐지하기 위해 많은 연구가 진행되었다[4-7]. 또한, 중첩 클러스터 구조에 속한 노드들은 다른 노드들보다 더 밀집(dense)하게 연결되어 있으며 해당 노드들이 그래프의 허브(hub)역할을 한다는 것을 증명한 연구[8]도 존재한다.

부분의 정보 전파 연구는 한 노드(node)에 대해 이웃 노드(1-hop neighbor)들의 영향력을 일정 기준(threshold)을 넘으면 정보가 전파되도록 정보 전파 메커니즘을 모델링하였다. 다시 말해, 새로운 정보가 노드에게 이득이 되면 노드는 그 정보를 습득하고, 이로 인해 정보 전파가 일어난다고 주장한다. 노드 자신에게 이득이 되는 새로운 정보를 습득한다는 관점에서, 실제 그래프에서 노드가 여러 정보를 습득(compatibility)하는 것이 해당 노드에게 가장 이득이 되는 경우가 있을 수 있다.

정보 전파에 관한 선행연구는 그래프의 중첩 클러스터 구조를 고려하지 않고 정보 전파 메커니즘을 모델링하였으며, [3]을 제외하고 정보의 공존성(compatibility)을 배제한 채로 정보 전파 메커니즘을 모델링하였다. 이 점을 보완하고자 본 연구에서는 중첩 클러스터 구조와 정보의 공존성을 모두 고려하여 실제 그래프에 부합한 새로운 정보 전파 모델을 제시한다. 또한, 중첩 클러스터 구조를 갖는 실제 네트워크에서의 정보 전파 양상은 기존 연구들이 제시한 것과 차이가 있음을 보인다.

2. 관련 연구

네트워크상의 정보 전파과정을 모델링하기 위해 많은 연구가 진행되었다. 이 연구들은 네트워크에서 새로운 정보를 습득한 임의의 노드(initial adopter)가 다른 노드에게 미치는 영향력이 각각 다르다는 가정 하에 그 영향력을 확률 분포로 표현하여 정보 전파 메커니즘을

모델링하였다.

대표적으로 소셜 네트워크(social network)에 대한 정보 없이 노드와 노드 사이의 영향력을 수학적으로 모델링함으로써 네트워크 전반에 대한 노드의 영향을 함수(Linear Influence Model)로서 표현한 연구[1]와 새로 정보를 습득한 노드가 다른 노드에 미치는 영향과 정보 전파 정도의 상관관계를 수학적으로 모델링한 연구[2]가 존재한다.

노드의 영향력을 표현하는 방법 중에 하나로 노드의 중요도(centrality)가 있으며, 이를 구하는 방법으로는 네트워크 전체에서 노드 개개인의 중요도를 구하는 방법(PageRank 등)과 클러스터 구조를 고려하여 노드의 지역적인 중요도를 구하는 방법[9]이 존재한다.

또한, 네트워크의 구조를 고려하여 정보 전파과정을 모델링한 연구[10]가 진행됐다. 이 연구는 같은 클러스터 사이에 존재하는 노드의 동질성(homophily)을 고려하여 정보 전파를 최대화하는 모델(community-based fast influence)을 제시했다.

다른 연구와 다르게 [3]에서는 한 노드가 다른 노드에 미치는 영향력을 이용하지 않고, 한 노드가 이웃 노드(1-hop neighbors)들에게 받는 영향력을 이용하여 정보 전파 메커니즘을 모델링하였다. 이 연구는 노드가 이웃 노드들에게 받는 영향력을 종합하여 자신에게 이익이 되는 정보를 선택하고, 이 과정을 통해 정보가 전파된다고 주장한다. 또한, [3]은 정보의 공존성(compatibility)을 고려했다. 이 연구에서 노드는 자신이 얻을 수 있는 이득을 최대로 하기 위해 이미 갖고 있는 정보와 새로운 정보 모두를 습득할 수 있다고 가정하였다. 실제 소셜 네트워크에서 노드가 여러 정보를 습득하고 있다는 점에서 이 연구는 실제 네트워크 상황을 고려했다고 할 수 있다. 또한, [3]은 [10]과 마찬가지로 네트워크의 클러스터 구조와 정보 전파의 상관관계도 분석하였는데, 클러스터 내부에 존재하는 노드들이 밀도 있게(dense) 연결되어 있고 서로 다른 클러스터에 존재하는 노드들은 그보다 희박하게(sparse) 연결되어 있기 때문에 클러스터가 정보 전파를 방해한다고 주장한다.

기존 네트워크에서의 정보 전파(information diffusion) 연구는 노드가 하나의 클러스터(cluster)에만 속한다는 가정 하에 노드가 다른 노드에 미치는 영향과 네트워크의 구조를 고려하여 정보 전파 알고리즘을 개발하였다. 하지만, 소셜 네트워크(social network) 등의 실제 그래프에는 중첩 클러스터(overlapping cluster)구조가 존재한다. 이와 같은 실제 네트워크의 중첩 커뮤니티를 정확하게 탐지하는 것이 대두되었고, 최근에 네트워크의 중첩 커뮤니티를 탐지하기 위한 연구가 활발히 진행되고 있다[4-7].

네트워크의 중첩 커뮤니티에 존재하는 노드와 간선의 주요 특징을 탐색하는 연구도 진행되고 있다. [8]에서는 실제 네트워크의 중첩 커뮤니티에는 다른 노드와 밀도 있게(dense) 연결되어 있는 노드가 많으며 중첩된 커뮤니티를 연결하는 역할을 하는 노드(connector node)가 존재한다고 주장한다.

다른 연구[11]에서는 중첩 커뮤니티에 존재하는 노드가 새로운 정보를 습득하였을 때, 다른 노드가 새로운 정보를 습득하였을 때보다 정보 전파의 범위가 넓어지는 것을 실험적으로 밝혔다. 이 연구는 중첩 커뮤니티에 존재하는 노드는 정보 전파의 핵심이라고 주장한다.

3. 알고리즘

본 연구는 중첩 클러스터 구조를 갖는 네트워크에서 클러스터가 정보 전파의 방해요소가 아님을 증명한다. 네트워크에서 클러스터란 간선으로 밀집(dense)하게 연결되어 있는 노드들의 집합을 일컫는다. 하나의 클러스터에 속한 노드는 다른 클러스터에 속한 노드들과 희박(sparse)하게 연결되어 있다. 본 연구는, 또한, 정보의 공존성(compatibility)도 고려하여 실제 네트워크에 부합하는 정보 전파 메커니즘을 제시한다. 본 연구에서 제시하는 정보 전파 알고리즘은 [3]의 정의 및 가정을 사용하였다. 때문에 같은 맥락을 이어가기 위해 3.1에서 [3]의 정의 및 가정을 소개한다.

3.1 정의 및 가정

정보를 습득한 노드는 해당 정보에 부합한 행동을 취한다고 가정한다. 노드가 취할 수 있는 행동은 영어 대문자(A, B 등)로 표현하며 그 때의 이득(payoff)은 소문자(a, b 등)로 표현한다. 또한, 증명의 용이성을 위해 노드가 취할 수 있는 행동은 A 와 B 두 개 뿐이라고 가정한다. 행동 A 는 새로 취할 수 있는 행동을 의미하며 행동 B 는 이미 네트워크 전반에 만연한 예전 행동을 의미한다. 따라서 행동 A 를 취했을 때의 이득(a)은 B 를 취했을 때의 이득(b)보다 항상 크거나 같다고 가정한다.

행동 B 를 취하고 있는 노드 v 가 새로운 행동 A 를 취하려면 노드 v 의 이웃노드 중 적정 수준 이상의 노드가 행동 A 를 취하고 있어야 한다. 만약 노드 v 가 행동 A 를 취할 때 B 를 취했을 때보다 이득이 더 크다면, 노드 v 는 행동 A 를 취한다. 노드 v 의 주변 노드 중 행동 A 를 취하는 노드의 비율을 p 로, 노드 v 의 degree를 d 라고 하면, v 는 행동 A 를 취하고 있는 $p*d$ 만큼의 노드와 연결되어 있고 행동 B 를 취하고 있는 $(1-p)*d$ 만큼의 노드와 연결되어 있다고 할 수 있다. 노드 v 가 행동 A 를 취하게 될 이득의 기준(threshold)은 식 (1)과 같이 표현할 수 있으며 그 기준을 q 라고 정의한다.

$$\begin{aligned} p^*d^*a &\geq (1-p)^*d^*b \\ \Leftrightarrow p &\geq \frac{b}{a+b} \end{aligned} \quad (1)$$

정의 1. (q value) 노드가 기준에 취하던 행동에서 새로운 행동을 취하기 위한 기준 q 는 다음과 같이 정의 할 수 있다.

$$q = \frac{b}{a+b}$$

즉, 행동 A 를 취하고 있는 이웃 노드의 비율 p 가 q 이상이면 노드 v 는 행동 A 를 취한다. 또한, 앞의 가정에 의해 b 는 a 보다 작거나 같으므로, q 의 값은 항상 $1/2$ 보다 작거나 같다. 식 (1)에서 $p*d^*a$ 는 노드 v 가 행동 A 를 취했을 때의 이득을, $(1-p)*d^*a$ 는 노드 v 가 행동 B 를 취했을 때의 이득을 나타낸다. 식 (1)의 p 와 q 를 이용하면 노드 v 가 행동 A 를 취했을 때의 이득과 행동 B 를 취했을 때의 이득을 아래와 같이 새로 정의할 수 있다.

$$\begin{aligned} p^*d^*a &\geq (1-p)^*d^*b \\ \Leftrightarrow p^*d^*\frac{a}{a+b} &\geq (1-p)^*d^*\frac{b}{a+b} \\ \Leftrightarrow p^*d^*(1-q) &\geq (1-p)^*d^*q \\ \Leftrightarrow p^*(1-q) &\geq (1-p)^*q \end{aligned} \quad (2)$$

따라서 식 (2)에 의해 노드 v 가 행동 A 를 취할 때의 이득과 B 를 취할 때의 이득은 각각, $P_A(v) = (1-q)^*p$ 와 $P_B(v) = q^*(1-p)$ 로 표현 가능하다.

클러스터(cluster)가 정보 전파에 방해요소가 된다는 것을 보이기 위해, 노드 밀도(node density)와 클러스터, 그리고, 클러스터 밀도(cluster of density ϵ)에 대해 정의할 필요가 있다.

정의 2. (노드 밀도) 노드의 집합 S 와 노드 v 가 있을 때, v 의 이웃 노드가 S 에 속하는 비율을 노드 밀도라고 정의하고 다음과 같이 수식으로 나타낼 수 있다.

$$\rho(v, S) := \frac{\text{links}(v, S)}{\deg(v)}$$

이 때, $\text{links}(v, S)$ 는 노드 v 가 어떤 노드의 집합 S 와 연결된 간선(edge)의 개수이며 $\deg(v)$ 는 노드 v 의 degree를 의미한다.

정의 3. (클러스터 밀도) 클러스터에 속한 노드의 해당 클러스터에 대한 노드 밀도의 최소값이 ϵ 일 때 클러스터는 클러스터 밀도 ϵ 을 갖는다고 정의한다. 클러스터를 C 라고 하였을 때, C 의 클러스터 밀도는 다음이 나타낼 수 있다.

$$\text{Cluster of density } \epsilon := \min_{v \in C} \rho(v, C)$$

그림 1과 같은 네트워크의 클러스터가 존재할 때, 노드 d 에 의해 해당 클러스터의 클러스터 밀도는 $2/3$ 가 된다.

[3]에서는 정보 확산과 클러스터의 상관관계를 정리 1을 통해 보였으며, 정리 1은 다음과 같다.

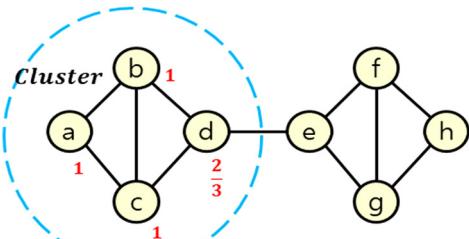


그림 1 클러스터 밀도가 2/3인 클러스터

Fig. 1 Cluster of density 2/3

정리 1. 중첩 클러스터가 존재하지 않는 네트워크에서 다음의 필요충분조건(Claim 1, Claim 2)은 참이다.

- ① Claim 1: 클러스터는 정보 전파를 방해한다. 즉, 새로운 정보를 습득한 노드를 제외한 네트워크에 클러스터밀도가 $1-q$ 이상인 클러스터가 존재할 때, 해당 클러스터에 의해 새로운 정보는 네트워크 전반으로 확산되지 않는다.
- ② Claim 2: 클러스터는 정보 전파를 방해하는 유일한 요소이다. 즉, 새로운 정보를 습득한 노드로부터 해당 정보가 네트워크 전반으로 확산되지 않을 때, 네트워크는 클러스터밀도가 $1-q$ 이상인 클러스터를 포함하여야 한다.

정리 1에 대한 증명은 [3]에서 확인할 수 있다. [3]에서는 Claim 1과 Claim 2의 필요충분조건이 성립함을 보임으로써 정보 전파와 클러스터는 서로 상반된다고 주장하였다.

3.2 중첩 클러스터가 존재하는 네트워크에서의 정보 전파

정리 1은 네트워크에 중첩 클러스터구조가 존재하지 않는다고 가정 한 후 정보 전파와 클러스터가 상반됨을 주장하였다. 하지만, 실제 네트워크에는 중첩 클러스터 구조가 존재한다. 정리 1의 주장이 중첩 클러스터 구조를 갖는 네트워크에도 부합하는지 확인하기 위해 본 연구에서는 중첩된 두 개의 클러스터(X, Y)을 이용하여 다음과 같이 가정한다.

가정 1. 네트워크에는 두 개의 클러스터 X 와 Y 가 존재하고 모든 노드는 X 또는 Y 에 속한다.

가정 2. 클러스터 X 와 Y 모두에 속하는 노드가 존재한다. 즉, X 와 Y 는 중첩된다.

가정 3. 행동 A 를 취하기 위한 기준 q 는 $1/2$ 보다 작거나 같고 클러스터 X 와 Y 의 클러스터 밀도는 $1-q$ 보다 크다.

위와 같은 가정 하에 중첩 클러스터 구조에서의 정보 전파를 확인하기 위해 두 가지를 확인할 필요가 있다. 첫째로 중첩 영역에 속하지 않은 노드로부터 중첩 영역에 속한 노드로의 정보 전파(diffusion into overlapped

area). 둘째로 중첩 영역에 속한 노드로부터 중첩 영역에 속하지 않은 노드로의 정보전파(diffusion from overlapped area)이다.

3.2.1 중첩 영역에 속한 노드로의 정보 전파

중첩 영역에 속한 노드로의 정보 전파는 클러스터 X 에만 속하는 노드로부터 중첩 영역에 속한 노드로의 정보 전파로 간소화할 수 있다. 이 때, 중첩 영역에 속한 노드는 모두 행동 B 를 취하고 있으며 집합 $X \setminus Y$ 에 속한 노드의 이웃 노드 중 중첩 영역에 속한 노드의 집합을 H 라고 한다. H 의 이웃 노드이며 집합 $X \setminus Y$ 에 속한 노드는 모두 행동 A 를 취한다고 가정한다. 다음과 같은 정리(theorem)를 통해 중첩 영역에 속한 노드로 정보 전파가 이뤄지지 않음을 알 수 있다.

정리 2. 중첩 영역으로의 정보 전파 상황에서 클러스터는 중첩 영역으로의 정보 전파를 방해한다. 즉, Claim 1은 참이다.

증명. 노드 h 가 집합 H 에 속하고 시점 t 에서 h 가 행동 A 를 취했다고 가정한다. 이는 시점 $t-1$ 에 h 가 행동 A 를 취했을 때의 이득이 B 를 취했을 때보다 크다고 판단했기 때문이다. 따라서 h 가 행동 A 를 취하기 위한 시점 $t-1$ 에서의 조건은, 행동 A 를 취하는 노드의 집합을 V_A 라고 했을 때, $P_A(h) \geq P_B(h) \Leftrightarrow \rho(h, V_A) \geq q$, for $h \in H$ 이다. 가정 3에 의해 H 에 속한 임의의 노드 h 는 Y 에 대해 $1-q$ 보다 큰 노드 밀도를 갖고 있다. 따라서 다음과 같이 정리할 수 있다.

$$\begin{aligned} & \rho(h, Y) > 1-q \\ & \Leftrightarrow 1 - \rho(h, X \setminus Y) > 1-q \\ & \Leftrightarrow \rho(h, X \setminus Y) < q, \text{ for } \forall h \in H \end{aligned} \quad (3)$$

집합 이론(set theory)에 의해 (3)과 같이 말할 수 있고 (3)은 가정에 의한 조건과 모순된다. 따라서 Claim 1은 참이다. \square

정리 3. 중첩 영역으로의 정보 전파 상황에서 클러스터는 중첩 영역으로의 정보 전파를 방해하는 유일한 요소이다. 즉, Claim 2는 참이다.

증명. 집합 S 는 정보 전파 과정이 끝난 후 행동 B 를 취하는 노드의 집합이고 노드 w 는 S 에 속한다고 가정한다. 본 연구는 S 를 클러스터라고 했을 때, S 의 클러스터 밀도가 $1-q$ 보다 크다는 것을 증명한다. 노드 w 는 끝까지 행동 A 를 취하지 않았으므로, 행동 A, B 를 취하는 노드의 집합을 각각 V_A, V_B 라고 하면, $\rho(w, V_A) = \rho(w, X \setminus Y) < q$ 라고 할 수 있다. 결국, 집합 이론에 의해 $\rho(w, V_B) > 1-q$ 이며 정보 전파 과정이 끝났을 때 $V_B = S$ 이므로 $\rho(w, S) > 1-q$ 라고 할 수 있다. 즉, 네트워크 전반으로 정보 전파가 일어나지 않았을 때, 클러스터밀도가 $1-q$ 보다 큰 클러스터 S 가 존재한다고 할 수 있다. 따라서 Claim 2는 참이다. \square

정리 2와 정리 3에 의해 네트워크의 중첩 커뮤니티구조가 존재할 때, 중첩 커뮤니티 바깥에 존재하는 노드에서 중첩 커뮤니티에 존재하는 노드로의 정보 전파는 이뤄질 수 없다.

3.2.2 중첩 영역에 속한 노드에서의 정보 전파

중첩 영역에 속한 노드의 이웃 노드이면서 중첩 영역에 속하지 않은 노드의 집합을 $S(S \subset X)$ 라고 가정하고 S 에 속한 노드로의 정보 전파 조건을 확인한다. 이 때, 중첩 영역에 속한 모든 노드가 행동 A 를 취하고 있다고 가정하고 중첩 영역에 속한 노드의 집합을 D 라 하면 Claim 1(클러스터는 정보 전파를 방해한다)이 거짓임을 알 수 있다. 즉, 중첩 영역에서 정보 전파가 일어남을 확인할 수 있다. 다음은 중첩 영역에 속한 노드에서 정보 전파가 일어날 때 Claim 1이 거짓이라는 것에 대한 증명이다.

정리 4. 중첩 영역에 속한 노드로부터 정보 전파 상황에서 클러스터는 정보 전파를 방해하지 못한다. 즉 Claim 1은 거짓이다.

증명. S 에 속한 노드를 s 로, 행동 A 를 취하는 노드의 집합을 V_A 라고 할 때, s 가 행동 A 를 취하기 위한 조건은 다음과 같다.

$$P_A(s) \geq P_B(s) \Leftrightarrow \rho(s, V_A) = \rho(s, D) \geq q$$

노드 s 는 클러스터 X 에 속하므로 가정 3과 집합 이론을 이용하면 다음과 같은 조건을 얻을 수 있다.

$$\begin{aligned} & \rho(s, X) > 1-q \\ & \Leftrightarrow \rho(s, D) + \rho(s, XD) > 1-q \end{aligned} \quad (4)$$

$$\rho(s, XD) + \rho(s, D) \leq 1 \quad (5)$$

s 가 행동 A 를 취하기 위한 조건, 즉, 식 (4),(5)와 q 의 값에 대한 가정을 모두 만족시키는 구간이 존재한다. 즉, 적절한 q 값이 존재할 때 노드 s 는 행동 A 를 취하게 되어 정보 전파가 일어난다고 할 수 있다. 따라서 Claim 1은 거짓이다. \square

Claim 1이 거짓이므로 네트워크에 중첩 클러스터(overlapping cluster) 구조가 존재할 때, 중첩 클러스터에 속한 노드로부터 다른 노드로의 정보 전파가 일어난다고 할 수 있다.

3.3 정보의 공존성을 고려한 중첩 영역 구조에서의 정보 전파

실제 네트워크에 부합하는 정보 전파 메커니즘을 개발하기 위해 노드가 여러 정보를 습득할 수 있다는 정보의 공존성(compatibility), 즉, 노드가 여러 행동을 취할 수 있는 것을 고려해야 한다. 다시 말해, 한 노드의 이웃 노드들이 여러 행동을 취할 수 있으므로 해당 노드가 얻을 수 있는 이득에 관한 수식을 새롭게 수립할 필요가 있다.

네트워크에서 취할 수 있는 행동이 A , B 또는 AB

(행동 A 와 B 를 모두 취한 경우)만 있다고 할 때, 노드는 행동 AB 를 취하는 이웃 노드(1-hop neighbor)로부터 자신이 행동 A 또는 행동 B 를 선택하느냐에 따라 이득 a 또는 b 를 얻을 수 있다고 가정한다. 또한, 노드가 행동 AB 를 취하게 되면 행동 AB 를 취하는 이웃 노드로부터 행동 A 와 B 중에서 이득이 더 높은 쪽의 이득을 얻을 수 있다고 가정한다. 하지만, 노드가 여러 행동을 취하기 위해서는 두 행동을 꾸준히 유지해야 한다. 즉, 여러 행동을 유지하기 위한 유지비용이 발생[3]하며 해당 유지비용을 $c(\in (0,1))$ 라고 정의한다. 이 유지비용은 노드가 여러 행동을 취할 때의 이득에서 차감된다. 이와 같은 가정 하에 노드가 취하는 행동에 대한 이득을 다음과 같이 정의할 수 있다.

정의 4. (정보의 공존성을 고려했을 때 노드의 이득)

- 노드 v 가 행동 A 를 선택했을 때 노드의 이득

$$\begin{aligned} P_A(v) &= \{\rho(v, V_A)*a + \rho(v, V_{AB})*a\}/(a+b) \\ &= (1-q)*\{\rho(v, V_A) + \rho(v, V_{AB})\} \end{aligned}$$

- 노드 v 가 행동 B 를 선택했을 때 노드의 이득

$$\begin{aligned} P_B(v) &= \{\rho(v, V_B)*b + \rho(v, V_{AB})*b\}/(a+b) \\ &= q*\{\rho(v, V_B) + \rho(v, V_{AB})\} \end{aligned}$$

- 노드 v 가 행동 AB 를 선택했을 때 노드의 이득

$$\begin{aligned} P_{AB}(v) &= \{\rho(v, V_A)*a + \rho(v, V_B)*b + \rho(v, V_{AB})*a\}/(a+b) - c \\ &= P_A(v) + q*\rho(v, V_B) - c \end{aligned}$$

여기서 V_A , V_B , V_{AB} 는 각각 행동 A , B , AB 를 취하는 노드의 집합을 의미한다.

본 연구에서는 정보의 공존성(compatibility)을 허용한 뒤, 특정 조건에서 노드가 행동 AB 를 선택했을 때, 다른 행동을 선택했을 때보다 높은 이득을 얻을 수 있다는 것을 보인다. 또한, 행동 AB 가 중첩된 클러스터로 전파(diffusion into overlapped area)되며 중첩된 클러스터에서도 전파(diffusion from overlapped area)되는 것을 증명한다.

3.3.1 중첩 영역에 속한 노드로의 정보 전파

3.2에서와 같은 가정 하에 네트워크에 두 개의 클러스터 X 와 Y 가 존재하고 X 와 Y 는 중첩된다고 가정한다. 중첩된 클러스터에 속한 노드의 집합을 H 라고 하고 H 의 이웃노드이며 클러스터 X 에만 속하는 노드들은 행동 AB 또는 A 만 취한다고 가정한다. 그 외의 노드들은 행동 B 를 취한다고 가정한다. 이 때, 행동 AB 는 중첩 영역으로 전파될 수 있음을 다음과 같은 정리 5를 통해 확인할 수 있다.

정리 5. 정보의 공존성을 허용하였을 때, 클러스터는 행동 AB 에 대하여 중첩 영역으로의 정보 전파를 방해하지 못한다. 즉, Claim 1은 거짓이다.

증명. 행동 AB 가 H 에 속한 임의의 노드 h 로 전파되기 위해서는 $P_{AB}(h) \geq P_B(h)$ 와 $P_{AB}(h) \geq P_A(h)$ 두

조건 모두 만족해야한다. 따라서 행동 A, B, AB 를 취하는 노드의 집합을 각각 V_A, V_B, V_{AB} 라고 할 때, 다음과 같이 정리할 수 있다.

$$\begin{aligned} P_{AB}(h) &\geq P_B(h) \\ \Leftrightarrow P_A(h) + q^* \rho(h, V_B) - c &\geq q^* \{\rho(h, V_B) + \rho(h, V_{AB})\} \\ \Leftrightarrow P_A(h) - q^* \rho(h, V_{AB}) &\geq c \\ \Leftrightarrow (1-q)^* \{1 - \rho(h, V_B)\} - q^* \rho(h, V_{AB}) &\geq c \\ \Leftrightarrow 1 - q - q^* \rho(h, V_{AB}) - c &\geq (1-q)^* \rho(h, V_B) \\ \Leftrightarrow 1 - \frac{q}{1-q} * \rho(h, V_{AB}) - \frac{c}{1-q} &\geq \rho(h, V_B) \end{aligned} \quad (5)$$

$$\begin{aligned} P_{AB}(h) &\geq P_A(h) \\ \Leftrightarrow \rho(h, V_B) &\geq \frac{c}{q} \end{aligned} \quad (6)$$

노드 집합 H 는 클러스터 Y 에 속하고 Y 에 속한 노드들은 위 가정에 의해 모두 행동 B 를 취하고 있으므로 다음과 같이 정리할 수 있다.

$$\rho(h, V_B) > 1 - q \quad (7)$$

즉, 식 (5),(6),(7)을 만족할 때, 행동 AB 는 중첩 영역에 속한 노드로 전파된다. 식 (5),(6),(7)을 모두 만족하는 부등식은 다음과 같이 정리할 수 있다.

$$1 - \frac{q}{1-q} * \rho(h, V_{AB}) - \frac{c}{1-q} > \max(\frac{c}{q}, 1 - q) \quad (8)$$

식 (8)을 만족하는 노드의 집합 H 가 존재할 수 있는지 확인하기 위해 다음과 같은 두 가지 경우를 고려해야 한다.

경우 1. $1 - q \geq \frac{c}{q}$

가정에 의해 q 가 $1/2$ 보다 작거나 같으므로 경우 1은 다음과 같이 정리할 수 있다.

$$1 - q \geq \frac{c}{q} \Leftrightarrow q^*(1 - q) \geq c, \text{ for } \frac{1 - \sqrt{1 - 4^*c}}{2} \leq q \leq \frac{1}{2}$$

따라서 경우 1에서 식 (8)은 다음과 같다.

$$\begin{aligned} (8) &\Leftrightarrow 1 - \frac{q}{1-q} * \rho(h, V_{AB}) - \frac{c}{1-q} > 1 - q \\ &\Leftrightarrow q^*(1 - q) - c > q^* \rho(h, V_{AB}) \\ &\Leftrightarrow 1 - q - \frac{c}{q} > \rho(h, V_{AB}) \quad (9) \\ &, \text{ for } \frac{1 - \sqrt{1 - 4^*c}}{2} \leq q \leq \frac{1}{2} \text{ and } 0 \leq c \leq \frac{1}{4} \end{aligned}$$

식 (9)와 주어진 조건을 만족하는 노드 h 가 있을 때, 행동 AB 는 중첩된 클러스터로 전파될 수 있다. 따라서 적절한 q 와 c 값에서 식 (9)를 만족하는 노드 집합 H 가 존재하고, 이 때 행동 AB 는 H 로 전파된다. 즉, Claim 1은 거짓이다.

경우 2. $1 - q < \frac{c}{q}$

경우 2일 때는 식 (8)을 만족시키는 노드 집합 H 가 존재하지 않음을 쉽게 보일 수 있다. 본 연구에서는 이에 대한 증명은 생략한다. \square

경우 1을 만족하는 노드가 존재할 때 정리 1의 필요 충분조건이 만족하지 않으므로 행동 AB 는 중첩 영역으로의 정보 전파가 일어난다고 할 수 있다.

3.3.2 중첩 영역에 속한 노드에서의 정보 전파

네트워크에 클러스터 X 와 Y 가 존재하고 X 와 Y 모두에 속하는 노드가 존재한다고 가정한다. 중첩 영역에 속하는 노드는 행동 A 또는 AB 만 취한다고 가정하고 중첩 영역에 속하는 노드의 이웃 노드(1-hop neighbor)이며 클러스터 X 에만 속하는 노드의 집합을 S , S 에 속한 임의의 노드를 s 라고 한다. 행동 A, B, AB 를 취하는 노드의 집합을 각각 V_A, V_B, V_{AB} 라 했을 때, 행동 AB 는 중첩 영역에서 전파될 수 있음을 다음과 같은 정리 6을 통해 확인할 수 있다.

정리 6. 정보의 공존성을 허용하였을 때, 클러스터는 행동 AB 에 대하여 중첩 영역으로부터의 정보 전파를 방해하지 못한다. 즉, Claim 1은 거짓이다.

증명. 행동 AB 가 s 로 전파되기 위해서는 3.3.1과 마찬가지로 $P_{AB}(s) \geq P_B(s)$ 와 $P_{AB}(s) \geq P_A(s)$ 두 조건 모두 만족해야 한다. 따라서 다음과 같이 정리할 수 있다.

$$\begin{aligned} P_{AB}(s) &\geq P_B(s) \\ \Leftrightarrow 1 - \frac{c}{1-q} - \frac{q}{1-q} * \rho(s, V_{AB}) &\geq \rho(s, V_B) \end{aligned} \quad (10)$$

$$P_{AB}(s) \geq P_A(s) \Leftrightarrow \rho(s, V_B) \geq \frac{c}{q} \quad (11)$$

식 (10),(11)을 동시에 만족하는 부등식은 다음과 같다.

$$1 - \frac{c}{1-q} - \frac{q}{1-q} * \rho(s, V_{AB}) \geq \rho(s, V_B) \geq \frac{c}{q} \quad (12)$$

식 (12)를 만족하는 노드 $s (\in S)$ 가 있을 경우 Claim 1은 거짓이라고 할 수 있다. 식 (12)는 다음과 같이 정리할 수 있다.

$$\begin{aligned} 1 - \frac{c}{1-q} - \frac{q}{1-q} * \rho(s, V_{AB}) &\geq \frac{c}{q} \\ \Leftrightarrow q^*(1 - q) - q^*c - q^* \rho(s, V_{AB}) - c^*(1 - q) &\geq 0 \quad (13) \\ \Leftrightarrow -q^2 + q - c &\geq q^* \rho(s, V_{AB}) \\ \Leftrightarrow -1 + \frac{1}{q} - \frac{c}{q^2} &\geq \rho(s, V_{AB}) \end{aligned}$$

이 때, $c < \frac{1}{4}$ 이고 $q > \frac{1 - \sqrt{1 - 4^*c}}{2}$ 이면 식 (13)의

좌변은 0보다 커지기 때문에 식 (13)을 만족하는 노드 s 가 존재할 수 있다. 따라서 식 (12)를 만족하는 노드 s 가 존재할 수 있다. 즉, 집합 S 가 존재할 수 있다. 따라서 Claim 1은 거짓이며 클러스터가 정보 전파에 방해하지 못하고, 행동 AB 는 중첩 영역에 속한 노드로 전파된다.

\square

3절에서 보인 증명은 표 1에 정리되어 있다. 네트워크에 중첩 영역이 존재하고 정보의 공존성을 고려하지 않

표 1 정리 요약

Table 1 Summary of our theorems

	Clusters can overlap	Clusters can overlap and nodes can have multiple information (compatibility)
Cascade into overlapped area	Clusters are obstacles to cascades : True Clusters are the only obstacles to cascades : True	Clusters are obstacles to cascades : False
Cascade from overlapped area	Clusters are obstacles to cascades : False	Clusters are obstacles to cascades : False

았을 때, 중첩 영역으로의 정보 전파(cascade into overlapped area)는 일어나지 않는다. 하지만, 중첩 영역으로부터의 정보 전파(cascade from overlapped area)는 일어난다. 정보의 공존성까지 고려했을 경우, 행동 AB는 중첩 영역으로의 정보 전파가 일어나며 중첩 영역으로부터의 정보 전파도 일어난다.

4. 실험

4.1 실험 데이터

중첩 클러스터(overlapping cluster)가 존재하는 실제 네트워크에서 정보의 전파 정도를 확인하기 위해 다양한 네트워크 데이터 세트인 KarateClub[12], LiveJournal, YouTube, 그리고 DBLP[13]를 사용하였다. 각 데이터의 노드와 간선은 각각 사람과 친구관계, 블로그와 친구관계, 유튜브 유저와 친구 관계, 그리고 논문 저자와 공동저자(co-authorship)관계를 나타낸다. 각 데이터 세트에 대한 정보는 표 2와 같다.

표 2 데이터 세트

Table 2 Data sets

Dataset	No. of nodes	No. of edges	No. of clusters
KarateClub	34	78	2
LiveJournal	4,000,000	34,900,000	311,782
YouTube	1,130,000	2,990,000	8,385
DBLP	420,000	1,340,000	2,547

4.2 Karate Club

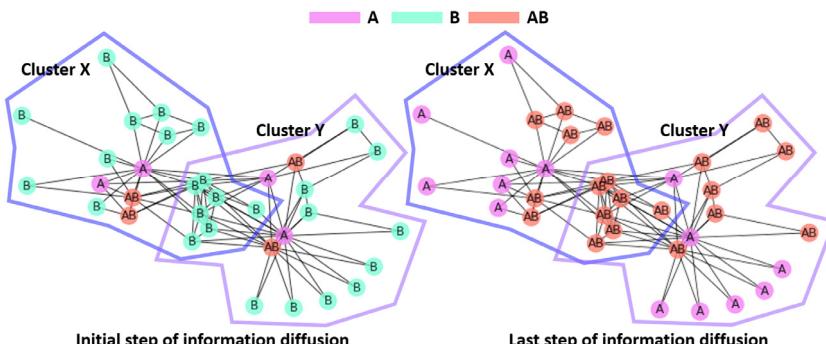
본 연구에서는 Karate Club 데이터 세트를 이용하여 정보 전파의 과정을 시각적으로 보인다. 본격적인 실험에 앞서 Karate Club 데이터 세트의 클러스터 밀도(cluster of density)가 본 연구의 가정에 부합하게끔 클러스터의 밀도를 $1-q$ 보다 크도록 triadic closure를 형성할 수 있는 노드와 노드 사이에 간선을 생성해준다. 이 후, 정보 전파 과정에 대해 실험을 진행한다.

4.2.1 중첩 영역으로부터의 정보 전파

Karate Club 데이터 세트에서 중첩 영역에 속하는 모든 노드가 행동 A를 취하도록 설정한 후 기준(threshold)이 되는 q 의 값을 정하고 정보 전파 과정을 확인했다. 이 실험에서는 q 의 값을 0.25로 설정하였다. 3.2.2에서 증명한 바와 같이 중첩 영역에서 행동 A가 전파되는 것을 확인할 수 있다(그림 2). Karate Club 데이터 세트에서 자줏빛을 띠는 노드는 중첩 영역에 속한 노드들이며 다른 색의 선으로 구분된 노드들은 다른 클러스터에 속한 노드들을 의미한다.

4.2.2 정보의 공존성을 고려한 중첩 영역으로의 정보 전파

Karate Club 데이터 세트에서 중첩 영역에 속하는 노드의 이웃 노드(1-hop neighbor)가 행동 A 또는 AB를 취하도록 설정한 후 실험을 진행했다. 이 실험에서는 q 의 값을 0.37, c 의 값을 0.1로 설정하였다. 3.3.1에서 증명한 바와 같이 중첩 영역으로 행동 AB가 전파되는 것을 확인할 수 있었다(그림 3).

그림 2 KarateClub에서 중첩 영역으로부터의 정보 전파($q=0.25$)Fig. 2 Diffusion from the overlapped area in Karate the Club ($q=0.25$)

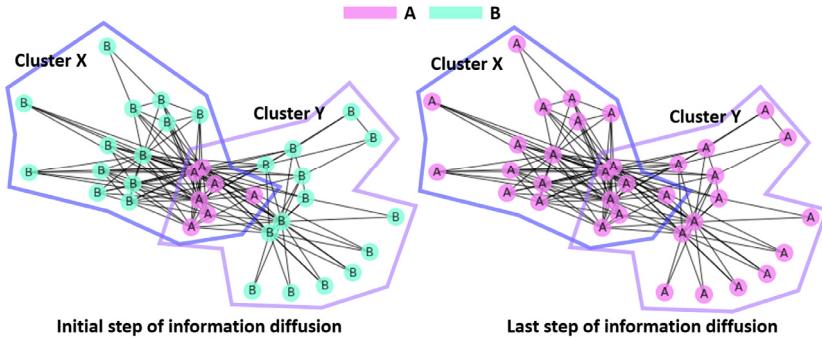


그림 3 KarateClub에서 정보의 공존성을 고려한 중첩 영역으로의 정보 전파($q=0.37$, $c=0.1$)
Fig. 3 Diffusion into the overlapped area with compatibility in the Karate Club ($q=0.37$, $c=0.1$)

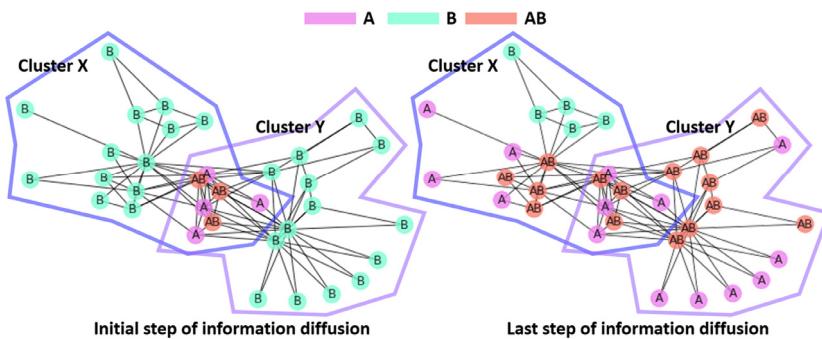


그림 4 KarateClub에서 정보의 공존성을 고려한 중첩 영역으로부터의 정보 전파($q=0.37$, $c=0.1$)
Fig. 4 Diffusion from the overlapped area with compatibility in the Karate Club($q=0.37$, $c=0.1$)

4.2.3 정보의 공존성을 고려한 중첩 영역으로부터의 정보 전파

Karate Club 데이터 셋에서 중첩 클러스터에 속하는 모든 노드가 행동 A 또는 AB 를 취하도록 설정한 후 실험을 진행했다. 이 실험에서는 q 의 값을 0.37, c 의 값을 0.1로 설정하였다. 3.3.2에서 증명한 바와 같이 중첩 영역에서 다른 노드로 행동 AB 가 전파되는 것을 확인할 수 있었다(그림 4).

4.3 LiveJournal, YouTube, DBLP

SNAP(<http://snap.stanford.edu/data/index.html>)에서 제공하는 real-world 데이터 셋을 이용하여 실험하였다.

4.3.1 중첩 영역으로부터의 정보 전파

각 데이터 셋(LiveJournal, YouTube, DBLP)에 중첩 영역에 속한 노드가 행동 A 혹은 AB 를 취하도록 설정하고 q 의 값을 다르게 적용한 후 정보 전파 결과를 관측했다. 이 경우 q ($\in (0, \frac{1}{2})$)의 값이 증가하는 것은 행동 A 를 취했을 때의 이득이 B 를 취했을 때의 이득과 비슷해짐을 의미한다. 따라서 q 의 값이 커질수록 행동 A 가 전파

되는 범위가 줄어들 것이다. 이 실험에서는 q 의 값이 0.05 일 때와 0.48일 때로 나누어 실험을 진행하였다(그림 5).

4.3.2 정보의 공존성을 고려한 중첩 영역으로의 정보전파 각 데이터 셋에 중첩 영역에 속한 노드의 이웃노드(1-hop neighbor)이며 중첩 영역에 속하지 않은 노드가 행동 A 혹은 AB 를 취하도록 설정하였다. c ($\in (0, \frac{1}{4})$)의 값이 고정되어있을 때 적절한 범위내의 q ($\in (\frac{1 - \sqrt{1 - 4*c}}{2}, \frac{1}{2})$)의 값을 증가시켜 정보 전파 결과를 관측했다. 이 실험에서는 c 의 값이 0.005로 매우 낮을 때 q 의 값을 0.01, 0.34로 두어 정보 전파 결과를 관측하였다(그림 6). 정리 5에서 증명한 바와 같이 행동 AB 가 중첩 영역으로 전파되어 네트워크 전반으로 퍼져나가는 것을 확인할 수 있으며 행동 AB 를 유지하는 비용 c 의 값이 낮기 때문에, q 의 값이 높아질수록, 즉, 행동 A 를 취했을 때 B 를 취했을 때의 이득이 비슷해질 수록 행동 AB 를 선택하는 노드의 수가 더 많아지는 것을 확인할 수 있다.

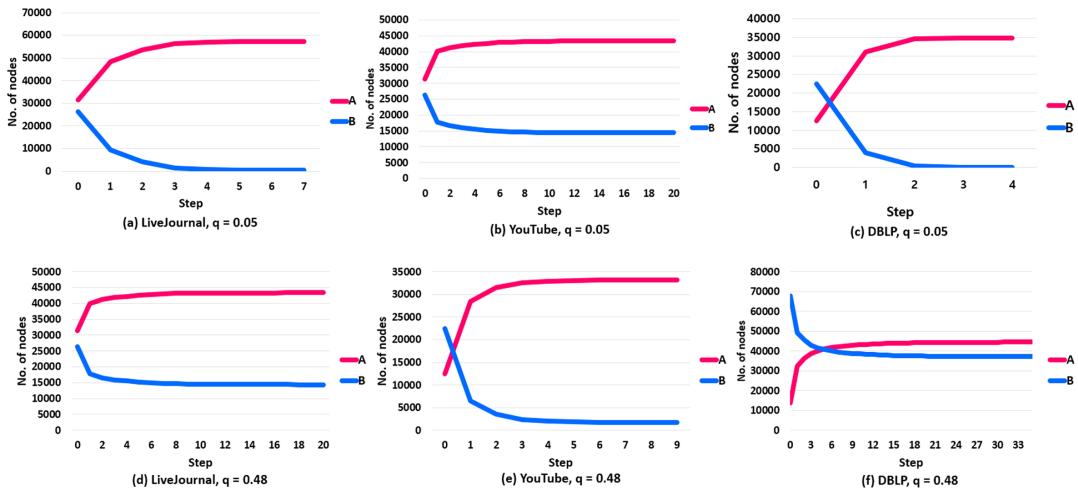


그림 5 q 의 값에 따른 중첩 영역으로부터의 정보 전파
Fig. 5 Diffusion from the overlapped area according to the q value

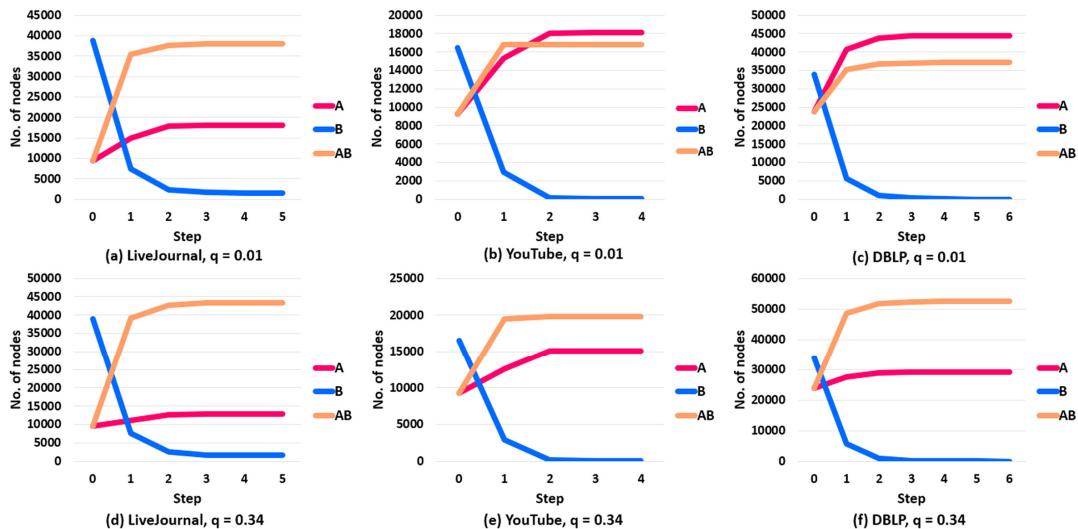


그림 6 c 를 0.005로 고정했을 때 q 의 값에 따른 중첩 영역으로의 정보 전파
Fig. 6 Diffusion into the overlapped area according to the q value with the fixed c value 0.005

4.3.3 정보의 공존성을 고려한 중첩 영역으로부터의 정보전파

각 데이터 셋에 중첩 영역에 속한 노드가 행동 A 혹은 AB를 취하도록 설정한 후, 4.3.2와 같은 q 와 c 값에 대해 실험을 진행하였다(그림 7). 이 경우, 정리 6에서 증명한 바와 같이 행동 AB가 중첩 영역으로부터 전파되는 것을 확인할 수 있다. 또한, 행동 AB를 유지하는 비용인 c 의 값이 작으면 q 의 값이 높아질 때, 즉, 행동 A와 B의 이득이 유사해질 때, 행동 AB를 취하는 노드의 개수가 증가하는 것을 확인할 수 있다.

4.4 실험 결과

본 연구에서는 중첩 영역으로부터의 정보 전파와 정보의 공존성(compatibility)을 고려하였을 때 중첩 영역으로부터의 정보 전파 및 중첩 영역으로의 정보 전파에 대해 실험을 진행했다. 행동 A를 취하기 위한 기준(threshold) q 와 행동 AB를 취할 때 생기는 유지비용 c 를 하이퍼 파라미터(hyper parameter)로 두어 q 와 c 의 값에 따른 정보 전파 경향을 확인하였다. 실험 결과, 행동 A의 이득이 높을수록, 즉, q 의 값이 작을수록 행동 A의 전파 범위가 넓어졌으며 c 의 값이 작고 q 의 값이

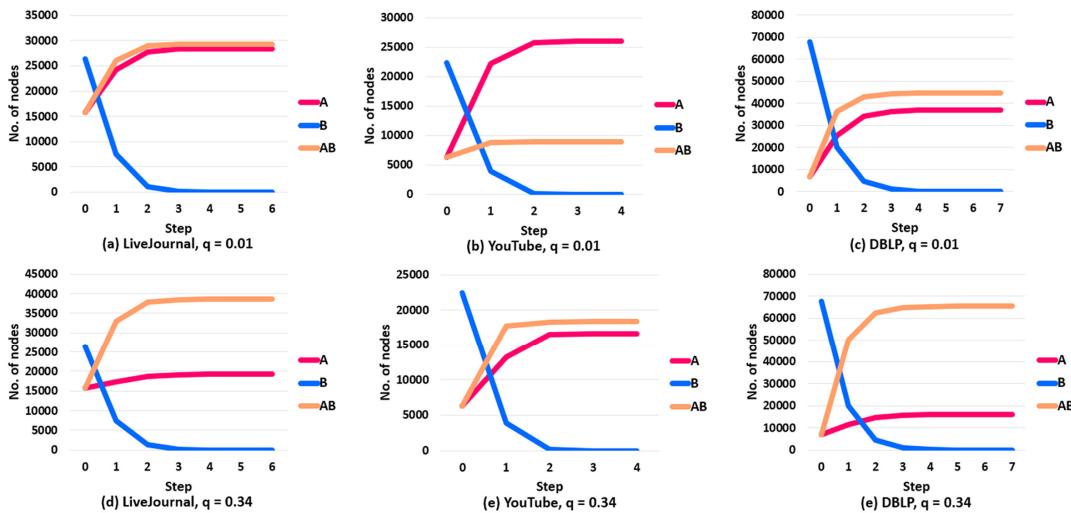


그림 7 c를 0.005로 고정했을 때 q의 값에 따른 중첩 영역으로부터의 정보 전파

Fig. 7 Diffusion from the overlapped area according to the q value with fixed the c value 0.005

큰 경우 행동 AB의 전파 범위가 넓어지는 것을 확인했다. 또한, 이 실험을 통해 본 연구에서 주장한 바와 같아, 중첩 클러스터 구조와 정보의 공존성을 고려하였을 때, 클러스터가 정보 전파의 방해요소가 아닐 수 있음을 확인하였다.

5. 결 론

본 연구에서는 실제 네트워크 구조에 부합하는 정보 전파 알고리즘을 제안하였다. 실제 그래프의 중첩 커뮤니티 구조에 맞는 새로운 정보 전파 메커니즘을 수학적으로 모델링하고 정보의 공존성도 고려하여 중첩 커뮤니티가 정보 전파의 방해요소가 되지 않을 수 있음을 증명했다. 또한, 중첩 커뮤니티에 속한 노드가 정보 전파에 핵심이 되는 것을 밝혔다.

References

- [1] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," *10th IEEE International Conference on Data Mining*, pp. 599–608, Dec. 2010.
- [2] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," *Proc. of the 13th international on World Wide Web*, pp. 491–501, May 2004.
- [3] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, Cambridge University Press, 2010.
- [4] J. J. Whang, I. S. Dhillon, and D. F. Gleich, "Non-exhaustive, Overlapping k-means," *Proc. of the 15th SIAM International Conference on Data Mining*, pp. 936–944, Apr. 2015.
- [5] J. J. Whang, Y. Hou, D. F. Gleich, and I. S. Dhillon, "Non-exhaustive, Overlapping Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 11, pp. 2644–2659, Nov. 2019.
- [6] J. J. Whang and I. S. Dhillon, "Non-exhaustive, Overlapping Co-Clustering," *Proc. of the 26th ACM Conference on Information and Knowledge Management*, pp. 2367–2370, Nov. 2017.
- [7] J. J. Whang, D. F. Gleich, and I. S. Dhillon, "Overlapping Community Detection Using Neighborhood-Inflated Seed Expansion," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, No. 5, pp. 1272–1284, May 2016.
- [8] J. Yang and J. Leskovec, "Structure and overlaps of ground-truth communities in networks," *ACM Transactions on Intelligent Systems and Technology*, Vol. 5, No. 2, pp. 1–35, Apr. 2014.
- [9] J. J. Whang and Y. Shin, "Localized Ranking in Social and Information Networks," *IEICE Transactions on Information and Systems*, Vol. E101-D, No. 2, pp. 547–551, Feb. 2018.
- [10] S. Lin, Q. Hu, G. Wang, and P. S. Yu, "Understanding community effects on information diffusion," *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 82–95, Apr. 2015.
- [11] J. J. Whang, "An Empirical Study of Community Overlap: Ground-truth, Algorithmic Solutions, and Implications," *Proc. of the 26th ACM Conference on Information and Knowledge Management*, pp. 2363–2366, Nov. 2017.
- [12] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, Vol. 33, No. 4, pp. 452–473, 1977.

473, Nov. 1977.

- [13] J. Yang and J. Leskovec, "Defining and Evaluating Network Communities based on Ground-truth," *12th IEEE International Conference on Data Mining*, pp. 745–754, 2012.



이 우 중

2018년 성균관대학교 수학과, 컴퓨터공학과 졸업(학사). 2018년~현재 성균관대학교 소프트웨어학과 석사 과정 관심분야는 그래프 마이닝, 데이터 마이닝, 기계학습, 빅데이터



황 지 영

2010년 이화여자대학교 컴퓨터공학과 졸업(학사). 2015년 텍사스 오스틴 대학교 (University of Texas at Austin) 컴퓨터 과학과 졸업(박사). 2016년~현재 성균관대학교 소프트웨어학과 조교수. 관심분야는 빅데이터, 데이터 마이닝, 그래프 마이닝, 기계학습