KIPS Trans. Softw. and Data Eng.
Vol.8, No.7 pp.275~288 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2019.8.7.275

Estimating the Rumor Source by Rumor Centrality Based Query in Networks

Jaeyoung Choi†

ABSTRACT

In this paper, we consider a rumor source inference problem when sufficiently many nodes heard the rumor in the network. This is an important problem because information spread in networks is fast in many real-world phenomena such as diffusion of a new technology, computer virus/spam infection in the internet, and tweeting and retweeting of popular topics and some of this information is harmful to other nodes. This problem has been much studied, where it has been shown that the detection probability cannot be beyond 31% even for regular trees if the number of infected nodes is sufficiently large. Motivated by this, we study the impact of query that is asking some additional question to the candidate nodes of the source and propose budget assignment algorithms of a query when the network administrator has a finite budget. We perform various simulations for the proposed method and obtain the detection probability that outperforms to the existing prior works.

Keywords : Rumor Source Detection, Epidemic Models, Maximum Likelihood Estimator, Query

네트워크에서 루머 중심성 기반 질의를 통한 루머의 근원 추정

최 재 영†

요 약

본 논문에서는 네트워크에서 충분히 많은 노드가 루머를 들었을 때 그 근원이 어디서부터 시작 되었는지를 추론하는 문제를 고려한다. 이것은 신기술의 확산, 인터넷에서의 컴퓨터 바이러스/스팸 감염, 인기 있는 주제의 tweeting 및 retweeting과 같은 많은 실제 환경에서 네트워크의 정보 확산이 빠르게 진행되고, 이 정보 중 일부는 다른 노드에게 악영향을 미칠 수 있기 때문에 매우 중요한 문제이다. 이 문제는 선행연구에 의해 감염된 노드의 수가 충분히 많으면 정규 트리의 경우에도 탐지 확률이 31%를 초과 할 수 없다는 것이 입증되었다. 이를 바탕으로 네트워크에 감염된 후보 노드에게 몇 가지 추가 질의를 하는 방법에 대해 조사하고 네트워크 관리자가 한정된 자산을 가지고 있을 때 각 노드에 대한 질의의 수를 어떻게 분배하는지에 대한 자산 할당 알고리즘을 제안한다. 마지막으로 제안한 방법에 대하여 다양한 시뮬레이션을 수행하였고 기존 선행 연구보다 우수한 성능을 확인하였다.

키워드 : 루머 근원탐지, 루머확산 모델, 최우추정량, 질의

1. 서 론

최근 많은 사람들이 모바일이나 테블릿과 같은 휴대용 PC 혹은 개인 PC를 통해서 언제 어디서나 인터넷을 통해 세계 각지의 소식을 빨리 접하고 있다. 이런 인터넷의 급속한 발전으로 인해 복잡한 네트워크 위에서 다양한 정보들이 전파되고 있는데 이런 정보의 확산 현상은 세계적인 대기업인 삼성이나 애플과 같이 자신의 회사의 이득을 위해 새로운 제품을 되도록 많은 사람들에게 빨리 알리고자 하는 목적으로 전파되는 경우도 있고 혹은, 정치인이나 유명인들이 자신의 신조를 사람들에게 알리고자 하는 목적으로 전파하기도 한다. 하지만, 이와 동시에 Fig. 1과 같이 악성루머나 바이러스와처럼 사람들에게 혹은 컴퓨터에게 좋지 않는 정보 또한 이런 인터넷망을 타고 급속도로 퍼져나가는 경우도 있다. 이는 인터넷망에서 익명성이 보장이 되는 경우에 대해서 일부 사람들이 이를 악용하여 발생되는 현상 중 하나이다. 루머나 악성코드처럼 좋지 않은 정보가 퍼지는 상황을 제어하는 방법은 크게 두 가지로 구분이 된다. 하나는 네트워크에 다양하게 연결된

† 정 회 원 : 호남대학교 미래자동차공학부 조교수
Manuscript Received : December 11, 2018
First Revision : January 24, 2019
Accepted : March 1, 2019
* Corresponding Author : Jaeyoung Choi(jychoi@honam.ac.kr)

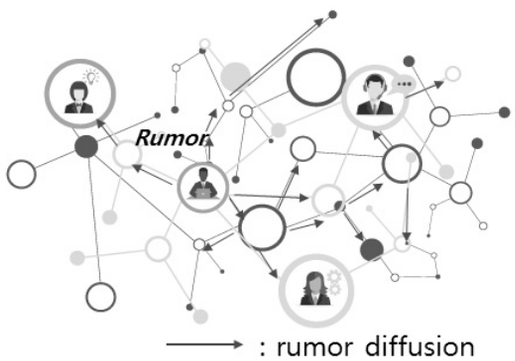

Fig. 1. Rumor Spreading on Social Networks
[Source: https://pngtree.eu/neural-interfaces.html]

여러 경로들을 찾아서 긴급하게 이를 차단하는 방법이 있을 수 있고 또 다른 하나는 정보를 처음 퍼뜨린 근원(Source)을 찾아서 더 이상 퍼지지 않도록 막는 방법이 있다. 전자의 경우는 현재 복잡하게 형성된 거대한 네트워크에서 이미 흘러가고 있는 정보를 쉽게 차단하는 것은 여러 현실적인 어려운 부분들이 있을 수 있기 때문에, 이 문제와 관련하여 주로 처음 퍼뜨린 근원을 먼저 찾아내는 방법에 대한 연구가 전 세계적으로 많이 진행이 되었다. 그러나 아직 본 문제는 네트워크에서 연결성이 변하지 않는 정적(Static)인 경우에 그 적용 가능성이 고려될 수 있다.

본 문제를 처음 다뤘던 연구[1]에서는 현존하는 가장 좋은 추정방법 중 하나인 최우추정량(Maximum Likelihood Estimator: MLE)을 사용해서 네트워크에 아주 많은 노드들이 본 정보를 듣게 되면[1], 즉 감염이 되면 처음 정보를 퍼뜨린 근원을 찾을 수 있는 확률이 얼마나 되는지에 대해 이론적으로 분석하였는데 모든 노드가 같은 수의 이웃 노드를 가진 비교적 간단한 정규트리(Regular tree)에서도 그 확률이 31%를, 일반적인 그래프에서는 10%를 넘지 못하는 것이 밝혀졌다. 이는 네트워크에 굉장히 많은 감염된 노드가 모두 루머를 처음 전파한 근원이 될 수 있는 상황에서는 나쁘지 않은 성능일 수 있으나, 아무리 잘 찾아도 이것 이상의 확률로 찾지 못한다는 근본적인 한계에 대해서 보인 것이기도 하다. 이후, 이를 더 높이고자 다른 여러 방법들이 제안이 되었다. 그 중에서도, 본 논문에서는 네트워크에서 감염된 노드들 중 일부를 선택하여 추가적인 질의(Query)를 하게 되는 경우 실제 근원을 얼마나 잘 찾을 수 있는지에 대해 연구하였다. 이런 질의 방법은 선행연구[2, 3]에서 처음 제안이 되었는데, 여기서는 네트워크에 감염된 노드들 중 최우추정량을 기반으로 일정한 거리에 떨어진 노드들을 후보자로 선택하여 물어보는 방법을 제안하였고, 이것은 근원이 될 수 있는 노드들을 초기 선택하는 최적의 방법이 아니다. 따라서 본 연구에서는 감염된 각 노드가 가진 근원이 될 가능성(Likelihood)의 순서대로 후보 노드들을 선택하여 물어보는 최적의 방법에 대해 분석하였다, 또한, [2, 3]에서는 선택된 노드가 잡음(Noise)이 있는 불확실한 대답을 확률적으로 한다는 가정에서 질문에 대한 정답을 필터링(Filtering)을 위해 모든 노드들에게 동일한 수의 질의를 물어보는 경우에 대하여 분석한 것에 대해 본 연구에서는 이 부분에 대해서도 네트워크에서 퍼진 정보를 보고 근원이 될 확률이 큰 노드들에게 더 많이 물어보는 좀 더 효율적인 방법에 대한 연구로 확장하였다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 본 연구와 관련된 여러 선행 연구들에 대한 소개를 비롯해서 그 이후로 새롭게 제기된 다른 문제들에 대해서 간략히 소개한다. 3장에서는 본 논문에서 고려하고 있는 정보 확산에 대한 수학적인 모델을 자세히 기술하고 이런 모델에서 루머의 근원을 잘 찾을 수 있는 두 가지 질의 방법에 대해 설명한다. 그리고 4장에서는 제안된 두 가지 질의 방법에 대하여 각각의 노드에게 질의 횟수를 어떻게 할당하는지에 대한 자산할당 알고리즘을 제안하고 5장에서는 이런 알고리즘 기법으로 새롭게 제안된 방법이 루머의 근원을 얼마나 잘 찾을 수 있는지에 대한 다양한 네트워크 환경에서의 실험을 보여준 후, 마지막장에서 논문에 대한 전체적인 요약을 하며 결론을 맺는다.

2. 관련 연구

본 장에서는 관련 연구로서 네트워크에서 루머가 퍼진 후 근원을 찾는 문제를 Table 1과 같이 세 가지 방법으로 정리한다.

2.1 단일 소스(Single Source) 추론

네트워크에서 정보가 퍼진 후 근원을 찾는 문제는 앞서 언급한 바와 같이 [1]에서 처음 제안이 되었다. 여기서는 소셜 네트워크뿐 아니라 일반적인 컴퓨터 네트워크에서 발생할 수 있는 바이러스나 루머와 같은 정보가 충분히 많은 시간동안 퍼져나간 후에 네트워크 관리자가 이를 찾는 방법으로서 최우추정량을 제안하였다. 이는 네트워크 관리자는 정보가 언제 퍼지기 시작했는지에 대한 정보를 모르고 다만 임의의 시점에서 네트워크를 관측했을 때, 루머가 확산된 형태(Snapshot)만 보고 그 근원에 대해 추론하는 방법인데, 여기서는 루머가 연결된 각 이웃들에게 확률적으로 퍼져나가는 모델을 고려하였기 때문에 확률에 기반 한 최우추정량을 사용하였다. 하지만, 해당연구에서는 그래프가 가장 다루기 쉬운 정규트리 구조인 경우에만 최우추정량을 유한시간에 계산하는 것이 가능하다는 사실을 수학적으로 밝혔고 이를 토대로 루머 중심(Rumor

1) 이 경우를 감염(infection) 되었다고 표현한다.

Table 1. Taxonomy of Rumor Source Detection Problems

<table>
<tr><th colspan="2">Single-source</th><th rowspan="2">Multi-sources</th><th rowspan="2">General Setting</th></tr>
<tr><th>Seeking</th><th>Hiding</th></tr>
<tr><td rowspan="2">Rumor Center [1]
Suspect Set [4]
Multiple Observations [5]</td><td rowspan="2">Adaptive Diffusion [7]</td><td>K-Center [11]</td><td>Partial Observation [14]</td></tr>
<tr><td colspan="2">ER-Random Graph [15]</td></tr>
<tr><td>Anti-Rumor [6]</td><td rowspan="2">Preference Attachment [8]

Spy-Adversary [9]</td><td rowspan="2">Different Starting [12]</td><td rowspan="2">MAPE-General Graph [16]</td></tr>
<tr><td>Query Approach [2,3]</td></tr>
<tr><td colspan="2">Game Theoretic Approach [10]</td><td colspan="2">Model Free [13]</td></tr>
</table>

Center: RC)이라고 하는 새로운 그래프 중심에 대한 개념을 만들어 냈다. 이것은 다름이 아니라, 각 노드를 근원으로 가정한 경우, 현재의 루머가 퍼진 모양이 나타나는 경우의 수가 가장 많은 노드를 말한다. 이를 바탕으로 최우추정량을 사용한 경우, 정규트리에서 얼마나 원래 근원을 잘 찾아낼 수 있는지에 대해서 이론적으로 결과를 얻어냈는데, 차수(Degree)가 2인 직선 네트워크인 경우에는 시간이 충분히 지난 후에는 최우추정량으로 실제 근원을 찾을 확률이 0이 되고, 차수가 3이상인 경우에는 시간이 아무리 많이 지나고 네트워크에 감염된 노드가 많아져도 0보다는 큰 확률로 찾아낼 수 있다는 것을 증명하였다. 하지만, 최우추정량만을 사용한 방법은 정보의 근원을 발견 할 확률이 최대 0.31을 넘지 못한다는 근본적인 한계도 밝혀냈다. 이 연구를 기반으로 근원을 찾을 확률을 높일 수 있는 여러 가지 다른 방법들이 제안이 되었는데 본 논문의 선행연구인 [2, 3]에서는 네트워크에 정보가 퍼진 후에 관리자가 각 노드에게 추가적인 질의(누가 너에게 이 정보를 알려 주었는지 등)를 함으로서 근원을 파헤쳐 나가는 새로운 방법을 제안하였고 실제 관리자가 이렇게 물어보는 경우 발생 할 수 있는 비용을 고려하여 주어진 자산(Budget)에서 얼마나 이런 추가적인 질문을 사용하면 근원을 잘 찾아낼 수 있는지에 대해 이론적으로 밝혔다. 하지만, 이론적인 분석을 위해 최적의 방법이 아닌 비교적 간단한 가정들을 하였고 본 논문에서는 이런 한계들을 극복하고자 더 효율적인 방법들에 대한 연구를 통해 결과를 분석하였다. 이 부분에 대한 자세한 설명은 다음 장인 모델 및 알고리즘 부분에서 다시 자세히 하도록 하겠다. [4]에서는 네트워크에서 전체 감염된 노드를 근원에 대한 후보로 보는 것 대신에 이전에 미리 정보에 근원에 대한 사전 지식(Prior information)이 있는 경우 즉, 모든 감염된 노드를 고려하지 않고 일정한 후보 그룹에서 반드시 근원이 있다고 가정을 할 때, 사후추정량(Maximum a Posterior Estimation: MAPE)을 사용하는 경우 근원을 찾을 확률이 1/2을 넘기도 하고 어떤 경우에는 확률 1로서 근원을 찾을 수 있다는 사실을 이론적으로 증명하였다. 하지만, 이것은 실제 네트워크에서는 근원에 대한 기존의 정보를 알기 어려울 수 있다는 한계가 있었다. [5]에서는 동일한 근원이 정보를 여러 번 퍼뜨리는 경우에 주어진 정보 확산 데이터를 가지고 근원을 얼마나 더 잘 찾아낼 수 있는지를 분석하였다. 재미있는 결과 중 하나로 동일한 근원이 단 두 번만 루머를 퍼뜨리는 경우에도 1/2가 넘는 확률로 찾을 수 있다는 사실을 이론적으로 밝혔다. [6]에서는 원래 정보에 반대되는 정보를 퍼뜨림으로서 얼마나 원래의 정보를 더 잘 찾게 되는가를 살펴보았는데 즉, 근원을 찾기 위해 두 가지 다른 정보가 같이 퍼지는 현상을 처음으로 제안하였다. 하지만, 이 논문에서 제안하고 있는 네트워크에 방어자(Protector)라고 하는 반대의 정보를 퍼뜨리는 노드가 원래 정보를 그 이웃으로부터 전파가 된 후 반대의 정보를 퍼뜨리는 한계적인 상황만을 고려하고 있고, 이 경우에서는 최우추정량이라는 가장 우수한 추정방법을 사용해도 반대정보의 확산이 원래 정보의 근원을 찾는데 시간이 충분히 지나면 전혀 도움이 되지 않는다는 사실을 밝혀냈다. 정보의 근원을 네트워크 관리자의 입장에서 얼마나 잘 찾아낼 수 있는가와 동시에 실제 근원의 입장에서 얼마나 자신을 잘 숨길 수 있는지에 대한 연구도 진행이 되었다. 특히, [7-9]에서는 자신의 정보를 보다 더 잘 숨기기 위한 방법으로 적응 형 확산(Adaptive diffusion)이라고 하는 새로운 확산 방법을 제안하였고, 이 경우 네트워크 관리자가 사용할 수 있는 가장 좋은 최우추정량을 가지고 찾는다 해도 잘 찾지 못한다는 것을 이론적으로 밝혔다. [10]에서는 이 둘을 각각 하나의 게임의 주체(Player)로 두고 근원은 정보가 퍼져나가는 비율을 조절함으로서, 그리고 관리자는 퍼진 노드들 중 일정한 크기의 집단을 선택함으로서 각각의 주체들이 자신의 유익(Payoff)을 최대화 하는 게임이론(Game theory)을 적용하여 분석하였고 이런 게임모델에서 시간이 충분히 지난 후 궁극적으로 평형을 이루게 되는 내쉬 균형(Nash Equilibrium: NE)은 어떻게 표현이 되는지를 분석하였다.

2.2 다중 소스(Multi-Sources) 추론

앞 장에서는 주로 네트워크에 정보의 근원이 하나인 경우 그것을 추론하는 방법에 대한 연구였다면, [11-13]에서는 이

보다 더 일반적인 경우로 여러 노드가 동시에 루머와 같은 정보를 퍼뜨리게 될 때, 이런 근원들의 집합을 잘 찾을 수 있는 방법들에 대해서도 살펴보았다. [11]에서는 네트워크에 k 개의 근원이 있다고 가정하고 시간이 충분히 지난 후에 모든 가능한 근원들의 집합을 고려하여 가장 그럴 듯한 후보를 찾는 알고리즘을 제안하였고 이것의 정확성에 대해 분석하였다. [12]에서는 같은 종류의 정보이지만 각각 다른 근원들이 다른 시간에 퍼뜨리는 일반적인 경우에 대해서도 모든 근원을 찾아 낼 수 있는 추론 방법을 제안하였고 이것의 성능을 실험적으로 얻어냈다. [13]에서는 정보가 퍼져나가는 모델이 알려지지 않은 경우 이를 적절하게 학습하여 다중 근원을 찾는 방법을 제안하였다.

2.3 일반적인 상황에서의 근원 추론(General Setting)

앞에서 분석을 위해 고려된 모델이 가진 한계를 극복하기 위한 보다 더 일반적인 상황에서의 방법들도 이후에 많이 제기가 되었다. [14]에서는 지금까지는 네트워크에 임의의 노드가 감염이 되었는지 아닌지를 알 수 있는 부분에 대해서는 완전한 정보가 있다고 가정하였으나, 실제로 네트워크 관리자는 이런 모든 노드를 관측하지 못하고, 그 중 일부분만 관측하게 되는 경우도 있는데, 이 경우에는 근원을 얼마나 잘 찾을 수 있는지에 대한 연구를 진행하였다. [15]에서는 앞에서 이론적인 분석을 위해서 주로 정규트리에서 결과를 얻은 부분에 대한 확장으로 확률적 그래프의 가장 간단한 그래프 중 하나인 Erdos-Renny(ER) 랜덤 그래프 위에서 루머가 퍼진 경우 근원을 찾는 문제를 연구하였고, [16]에서는 보다 더 일반적인 임의의 네트워크에서 근원을 찾는 방법으로 사후추정량을 사용하여 실험적으로 결과를 얻었다.

3. 시스템 모델 및 질의 접근 방법

본 장에서는 네트워크에서 루머를 듣고 감염된 노드에게 어떤 추가적인 질문들을 어떤 방법으로 물어보는지에 대한 query 종류와 방법 및 정보의 확산 모델에 대해 설명한다.

3.1 정보의 확산 모델(Rumor Diffusion models)

본 논문에서 고려하고 있는 네트워크는 $G=(V,E)$라는 그래프로 표기될 수 있고 여기서 V는 네트워크에 있는 모든 노드(Node)들의 집합이고 E는 각 노드를 연결하는 변(Edge)들의 집합이다. 여기서 노드란, 소셜 네트워크에서 한 사람의 개체를 말하기도 하고 혹은 인터넷망에서 한 컴퓨터를 말하기도 한다. 그리고 이런 노드들을 서로 연결하는 것이 그래프에서 변으로 표현이 되는데 이는 각 노드들의 소셜 연관성(Relationship) 혹은 각 컴퓨터들의 연결성을 의미한다. 본 연구에서는 기존의 다른 논문들에서 가정한 것과 같이 정보가 퍼져나가는 상황에서 네트워크의 경계에서의 영향을 피하기 위해 노드가 충분히 많이 있고 네트워크에 있는 모든 노드가 하나로 연결된(Connected) 경우를 고려한다. 본 논문에서는 앞서 언급한 바와 같이 네트워크에 한 명의 루머와 같은 정보를 퍼뜨리는 노드가 있고 이를 정보의 근원이라고 표현하고 s^*로 표기한다. 네트워크에서 이 루머가 근원으로부터 그 이웃에게 전달되어서 퍼져나가게 되는데, 여기서 정보가 퍼지는 현상은 이미 잘 알려진 확산 방법 중 하나인 Susceptible- Infected (SI) 모델을 따른다고 가정한다. 즉, 이것은 한 노드가 이미 감염된 상태라고 하면 이와 연결된 이웃노드는 비율이 $\lambda>0$인 지수분포(Exponential distribution)를 따라서 그 정보를 확률적으로 얻는 경우를 말한다. 이런 SI 모델에서는 만약 한 노드가 이웃노드로부터 감염된 후에는 시간이 지나도 그 상태가 변하지 않는다고 가정을 한다. 또한, 정보가 지수분포를 따라 퍼지게 될 때는 동일한 노드에 대해서 감염된 여러 이웃으로 부터 정보가 동시에 퍼질 수 없다는 것을 쉽게 확인할 수 있다.

1) 최우추정량

논문[1]에서 저자들은 위와 같이 SI 확산모델에서 동일한 비율 λ로 루머가 퍼지는 경우, 네트워크 관리자가 임의의 시점에서 관측했을 때 전체 N개의 노드가 감염이 된 그래프를 G_N이라고 하면, 정규트리 에서는 다음과 같이 최우추정량을 계산하였다.

$$\begin{aligned}\hat{s}_{MLE} &= \arg\max_{v\in G_N} P(G_N|v=s^*) \\ &= \arg\max_{v\in G_N} R(v,G_N)\end{aligned} \tag{1}$$

여기서 $P(G_N|v=v^*)$는 감염된 노드 v를 근원으로 가정했을 때, 감염그래프 G_N을 생성할 확률을 의미하고 따라서 최우추정량은 감염된 노드 중 이런 확률을 최대화 시키는 노드를 의미한다. 그리고 $R(v,G_N)$는 감염그래프 G_N에서 노드 v가 가진 루머 중심성인데, 이것은 v가 근원이라고 가정했을 때, G_N을 만들어 내는 다양한 종류의 확률 경로(Sample path)가 존재 할 수 있고 이를 더한 값을 의미한다. 자세한 계산 방법은 [1]을 참조하면 된다. 이것은 일반적으로 메시지 교환(Message passing)방법에 의해서 $O(N)$시간 안에 계산이 되는 것으로 알려졌다. 본 논문에서는 정규 트리뿐 아니라 보다 일반적인 그래프를 고려하기 위해서 다음과 같이 루머 중심 노드를 새롭게 정의한다.

정의 3.1 (루머 중심 노드와 루머 중심성) 일반적인 그래프에서 감염 그래프 G_N이 주어진 경우 루머 중심 노드 $\hat{s}_{RC}$는 다음과 같이 정의된다.

$$\hat{s}_{RC} = \arg\max_{v\in G_N} R(v, T_{bfs}(v)) \tag{2}$$

여기서 $T_{bfs}(v)$는 G_N에서 노드 v에 Breath-First-Search (BFS)로 만들어진 트리이고 $R(v, T_{bfs}(v))$는 이 경우의 루머 중심성이다.

위의 정의에서 BFS는 일반적인 루프(Loop)가 있는 그래프를 트리모양으로 근사시켜주는 잘 알려진 방법 [1] 중 하나이다. 따라서 만약, 그래프가 정규 트리인 경우에 (2)에서 정의된 루머 중심은 [1]에서 소개된 것과 정확하게 일치한다는 것을 확인할 수 있다.

3.2 Query 종류 및 부정확한 대답 모델

본 연구에서 고려하는 질의는 선행 연구[2, 3]에와 비슷하게 아래와 같이 두 가지의 질문으로 구성되어 있다.

Fig. 2. Identity/Direction Query

1) 정체성/방향 질의(Identity/Direction query)

네트워크 관리자(Querier)는 Fig. 2에서와 같이 먼저 감염된 노드에게 자신이 실제 근원인지 아닌지를 묻는데 이를 정체성(Identity) 질의라고 명명한다. 만약, 질문 받은 노드가 자신이 근원이라고 말하면 관리자는 랜덤으로 이웃의 감 염 노드 중 한 노드를 골라서 다시 물어본다. 하지만, 만약 자신이 근원이 아니라고 대답한 경우에는 관리자는 그럼 그 정보를 누구한테 들었는지에 대해 물어보는데 이를 방향 질문(Direction query)이라고 한다. 그러면, 그 노드는 자신의 이웃노드 중 한 노드를 지적하고 그 노드로부터 정보를 들었다고 대답한다. 이 둘의 질문이 결합된 질의를 본 논문에서는 정체성/방향 질의(Identity/Direction query)라고 정의한다. 실제 네트워크에서는 질의를 받은 노드가 아무런 응답을 해 주지 않을 수도 있지만, 본 논문에서는 분석의 용이함을 위해서 질의를 받은 노드는 일단 반드시 응답을 주는 상황을 고려한다.

2) 부정확한 대답

본 연구에서 고려하고 있는 질문에 대한 대답은 부정확성을 가지고 있다고 가정하고 있다. 다시 말하면, 질문을 받은 노드 v는 정체성질문에 대해서는 확률 p_v로 진실을 얘기하고 방향성 질문에 대해서는 확률 q_v로 진실을 얘기한다고 가정한다. 이렇게 가정을 하는 이유는 일반적으로 질의에 대한 대답은 사람이든지 컴퓨터든지 잡음이 있을 수 있는 환경을 고려할 수 있기 때문이다. 본 논문에서는 분석의 용이함을 위해 $p_v = p$, $q_v = q$로 즉, 모든 노드가 진실을 말할 확률이 같은 상황을 가정하고 또한 대답에 대한 완전한 랜덤성을 피하기 위해 $p > 1/2$ 와 $q > 1/nb(v)$ (여기서, $nb(v)$는 노드 v의 이웃의 수이다)의 상황을 고려한다.

3.3 질의 모델

일반적으로, 네트워크 관리자가 질의를 사용하는 경우 그에 대한 비용이 따를 수 있다. 본 논문에서는 위와 같이 정체성/방향 질문을 한 번 할 때마다 일정한 비용이 지불되는 경우를 고려하고 있고, 실제 네트워크 관리자는 총 $K > 0$의 자산(Budget)이 있다고 가정한다. 이런 가정을 바탕으로 본 논문에서는 다음과 같은 두 가지 질의 방법에 대해 고려한다.

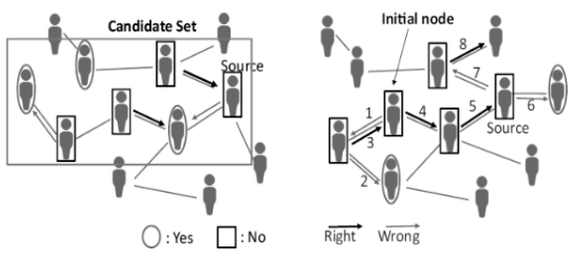

Fig. 3. Batch Query and Interactive Query with Untruthful Answers[3]

1) 집단형 질의(Batch Query)

집단형 질의 방법은 Fig. 3(a)에서와 같이 감염된 노드의 집단을 한 번에 선택하여 정체성/방향 질문을 하는 것을 의미한다. 즉, 네트워크 관리자에게 주어진 자산 K를 기준으로 몇 명의 노드에게 물어볼 수 있는지를 계산하고 이들을 선택해서 그림에서와 같이 집단으로 질문을 하게 되면 어떤 노드들은 자신이 근원이라고 하고(그림에서 Yes라고 대답한 노드들) 만약 자신이 근원이 아니라고 대답하면 누구로부터 들었는지에 대한 추가적인 방향 질문에 대한 대답을 한다.

2) 상호작용형 질의(Interactive Query)

상호작용형 질의 방법은 Fig. 3(b)에서와 같이 감염된 노드 중 한 노드(Initial node)를 선택해서 정체성/방향 질문을 한다. 만약, 이 노드가 자신이 근원이라고 답하면 주위 다른 노드를 랜덤하게 선택해서 같은 질문을 하고, 그렇지 않고 자신이 근원이 아니라고 하면 주위에서 누구로부터 들었는지에 대한 답을 주고 이 노드에 의해 선택된 노드에게 같은 질문을 수행한다. 이런 과정을 네트워크의 관리자가 소유한 자산을 다 소진할 때까지 반복한다.

4. 자산할당 알고리즘(Budget Allocation Algorithm)

앞 장에서와 고려하고 있는 모델에서 노드가 네트워크 관리자의 질문에 대해 불확실한 대답을 하는 상황을 고려하고

있기 때문에, 한 노드가 한 번의 대답에 대한 결과를 사용하는 방법은 정답을 추론하는데 어려울 수 있다. 따라서 동일한 질문을 여러 번 물어보는 경우를 고려하고 대답에 대한 모든 데이터를 가지고 진실을 추정한 후 이런 데이터를 적절히 사용하여 최종적으로 근원이 누구였는지를 알아내는 것이 더 네트워크 관리자의 한계적인 자산을 사용하는 데 있어서 효율적일 수 있다. 따라서 본 장에서는 첫째, 어느 노드에게 정체성/방향 질문을 얼마나 많이 물어볼 것이지 둘째, 각 질의 방법에 따라 질문하여 얻어진 데이터를 어떻게 필터링을 할 것인지에 대한 알고리즘을 제안한다. 기존의 연구[2, 3]에서도 모든 노드에게 여러 번 반복적으로 질문하는 것을 고려하고 있으나 모든 노드에게 같은 횟수만큼 물어보는 것을 가정하고 있다. 하지만, 네트워크에 루머가 퍼진 정보를 통해 어떤 노드가 더 근원이 될 확률이 큰지 어느 정도 알고 있는 상황이므로 이 정보를 가지고 각 노드마다 다른 횟수로 물어보는 방법을 고려하였다.

4.1 루머 중심성 기반 집단 질의 알고리즘 (Rumor Centrality based Batch Query Algorithm)

루머 중심성 기반 집단 질의 알고리즘은 다음과 같이 3단계로 이루어져 있다.

1) 초기 후보 노드 선정

네트워크에 정보가 퍼지게 되면 [1]의 결과를 바탕으로 감염된 모든 노드들에 대한 루머 중심성을 계산할 수 있고 이에 비례하여 먼저, 네트워크 관리자가 소유한 예산에 맞춰 노드의 집합을 선택하게 되는데 그 방법은 다음과 같다. 먼저, Fig. 4에서와 같이 전체 선택하는 노드의 수를 K_r이라고 하자. 즉, 감염된 노드들 중에 루머 중심성을 기준으로 K_r 만큼의 상위노드들을 선택한다. 그리고 이 노드의 집합에 정체성/방향에 대한 질문을 한다. 그 다음은, 선택된 노드 집합 중 가장 루머 중심성이 낮은 노드(그림에서 노란색으로 박스)를 제외시키고 남은 K_r-1명의 노드에게 정체성/방향 질문을 한다. 이런 과정을 마지막 노드 즉, 루머중심성이 가장 큰 노드에 질문을 던지게 될 때까지 진행한다. 그러면 다음과 같이 간단한 식으로 초기에 얼마의 K_r을 선택해야 하는지를 알 수 있다.

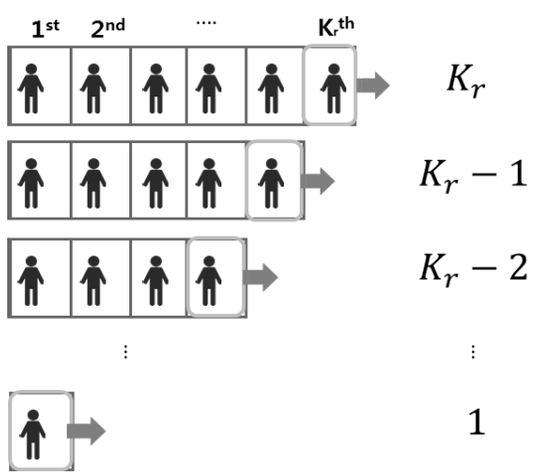

Fig. 4. Proposed Budget Assignment Method

$$1+2+\ldots+K_r=\frac{K_r(K_r+1)}{2}=K \tag{3}$$

따라서 위의 식을 적절하게 반올림해서 계산하면 $K_r=\lfloor\sqrt{2K}\rfloor$로 구해진다.

2) 대답 정보

위의 방법으로 후보 노드가 선정이 되면 루머 중심노드는 K_r번의 질문을 받게 되고 2번째로 루머 중심성이 큰 노드는 K_r-1의 질문을 그리고 가장 마지막으로 K_r번째로 루머 중심성이 큰 노드는 총 1번의 질문을 받게 된다. 각각의 동일한 질문에 대해 정체성 질문에 대해서는 진실을 말할 확률 p로 매번 독립적으로 대답을 하고 방향성 질문에 대해서는 확률 q로 진실을 말한다고 할 때, 이런 불확실한 여러 대답에 대해 정답을 잘 추려낼 수 있는 방법 중 하나는 다수결(Majority Voting: MV)인데 즉, 정체성 질문에 대해서 만약 전체 받은 질문 중 자신이 근원이라고 대답한 수가 과반수가 넘으면 그 노드는 정체성 질문에 대한 최종 후보의 집합 I에 속하게 되고 반대의 경우는 여기서 제외된다. 그리고 방향 질문에 대한 대답도 마찬가지로 다수결의 원칙에 따라 가장 많이 지적당한 이웃 노드가 그 노드에게로 정보를 준 노드로 선정하게 되고 각 노드를 중심으로 전체 네트워크에서 방향성 질문에 대한 확정된 답을 통해 가장 지적을 많이 받는 노드가 방향 질문에 대한 최종 후보 집합 D에 속하게 된다.[2)]

3) 최종 근원추정

이렇게 추려진 정체성 질문에 대한 최종적인 근원추정 방법은 위에서 얻은 두 질문에 대한 최종 후보 집합 I와 D에 속한 노드 중에 루머 중심성이 가장 큰 노드를 근원으로 추정하는 방법을 고려하고 수학적으로 표현하면 아래와 같다.

$$\hat{s}_B=\arg max_{v\in I\cap D}R(v,T_{bfs}(v)) \tag{4}$$

만약, 위의 경우 $I\cap D=\varnothing$인 경우는 $v\in I\cup D$인 상황을 고려해 그 근원을 추정한다. 이를 바탕으로 제안한 집단 질의 자산할당 방법이 만약 자산이 충분히 커지는 경우에서는 동일하게 물어보는 방법보다 항상 좋다는 것을 보일 수 있다. 이를 위해 먼저 $\vec{r}_{he}:=[r_1,r_2,,,r_{K_r}]$과 $\vec{r}_{ho}:=[r,r,...,r]$를 각각 선택된 후보노드에게 물어보는 질의의 수에 대한 벡터로 놓으면 다음이 성립한다.

2) 이렇게 접근하는 이유는 앞서 모델에서 각 노드가 진실을 말할 확률이 완전한 랜덤이 아닌 거짓을 말할 확률보다 조금 큰 상황을 고려하고 있기 때문이다.

정리 4.1 (집단 질의) 정규트리 G에서 $P(\hat{s}_B = s^*|\vec{r}_{he})$와 $P(\hat{s}_B = s^*|\vec{r}_{ho})$를 각각 제안한 자산할당 방법과 동일한 횟수로 질의하는 방법에 대한 근원 탐지 확률(Detection Probability)이라고 하자. 만약 초기 후보노드 집합의 크기가 같으면 다음이 성립한다.

$$\lim_{K\to\infty} P(\hat{s}_B = s^*|\vec{r}_{he}) \geq \lim_{K\to\infty} P(\hat{s}_B = s^*|\vec{r}_{ho}) \tag{5}$$

위의 결과는 집단 질의방법에서 동일한 자산을 가지고 동일한 수의 초기 후보노드들이 선정 할 수 있다면 제안한 루머 중심성 기반의 질의 방법이 일정한 횟수로 불어보는 방법보다 탐지확률이 크다는 것을 의미한다.

증명. 본 정리에 대한 증명을 위해 먼저, 근원 탐지확률을 다음과 같이 조건부 확률로 표현하면 다음과 같다.

$$P(\hat{s}_B = s^*) = P(\hat{s}_B = s^*|s^* \in C_{\vec{r}})P(s^* \in C_{\vec{r}}) \tag{6}$$

Equation (6)은 먼저, 탐지확률을 루머의 근원이 초기 후보노드 집합인 $C_{\vec{r}}$에 (여기서 $\vec{r}$은 대답 횟수에 대한 벡터이다) 포함되는 사건에 대한 조건부 확률로 표현한 것이다. 여기서, Equation (6)의 두 번째 확률은 선행연구[2]에 의해서 정규트리에서 동일한 크기의 초기 후보노드가 고려되는 상황에서는 루머 중심성을 기준으로 선택되는 집합에 실제 근원이 속할 확률이 최대화 된다는 것이 밝혀졌다. 따라서, 정리 1을 보이기 위해서는 Equation (6)의 첫 번째 확률에 대해서 제안한 자산할당 방법이 동일한 횟수의 방법보다 항상 크다는 것을 보이면 된다. 이를 위해서 고려되어야 하는 부분은 초기후보로 선택된 노드들에서 여러 번 물어보는 질의에 대한 대답을 다수결 법칙에 따라 추론될 때 최종적으로 루머의 근원이 집합 $I \cap D$에 들어갈 확률에 대해 알아보는 것이다. 즉, $P(s^* \in I \cap D)$을 살펴보면

$$\lim_{K\to\infty} P(s^* \in I \cap D|\vec{r}_{he}) \geq \lim_{K\to\infty} P(s^* \in I \cap D|\vec{r}_{ho}) \tag{7}$$

가 성립이 되는 것을 알 수 있는데 그 이유는 다음 두 가지로 설명이 된다.

(i) $\lim_{K\to\infty} P(s^* \in I|\vec{r}_{he}) \geq \lim_{K\to\infty} P(s^* \in I|\vec{r}_{ho})$: 이 결과가 나오는 이유는 제안한 자원할당 방법은 앞의 Fig. 4에서 초기 후보노드 집합에서 루머 중심성이 가장 작은 노드부터 하나씩 제거하는 방법이고 이것은 선형적으로 루머 중심성이 큰 노드에게 더 많이 물어보는 방법이다. 따라서 전체 자원 K가 커지면 초기 후보 노드 집합의 크기도 커지게 되고 실제 루머 근원이 이 집합에서 평균적으로 할당받는 질의 횟수가 되는 동일 질의 방법에서의 r 번 보다 확률 1로서 더 많은 양의 횟수로 질의를 받는다. 그 이유를 살펴보기 위해, $K(s^*)$를 제안된 자산할당 방법에 의해서 루머 근원인 s^*에 할당되는 질의의 수라고 하자. 그러면 다음 식을 만족하는 어떤 상수 $r \leq c \leq K_r$ 이 존재한다.

$$\begin{aligned}\lim_{K\to\infty} P(K(s^*) \geq r) &= \lim_{K\to\infty} \sum_{n=r}^{K_r} P(K(s^*) = n) \\ &\geq \lim_{K\to\infty} \sum_{n=c}^{K_r} P(K(s^*) = c) = 1,\end{aligned} \tag{8}$$

여기서 둘째 줄에 있는 부등호는 [1]에서 루머 근원과 루머 중심이 정규 트리인 경우 확률 1로서 상수거리에 있다는 사실로부터 유도가 된다. 다수결 방법으로 근원인지 아닌지를 결정을 할 때, 본 논문에서 정체성 질문에 대해 항상 진실을 말할 확률이 1/2 보다 크기 때문에 더 많이 질의를 받으면 최종 정체성 질의에 대한 후보 집합인 I에 들어갈 확률이 더 커진다. 따라서 (i)이 성립한다.

(ii) $\lim_{K\to\infty} P(s^* \in D|\vec{r}_{he}) \geq \lim_{K\to\infty} P(s^* \in D|\vec{r}_{ho})$: 앞서 (i)에서 언급한 바와 같이 실제로 루머의 근원이 루머 중심 근처에 있다는 것이 알려져 있는데 이는 다시 말하면 루머 근원 주위의 많은 노드들의 루머 중심성이 근원에서 많이 떨어져 있는 노드에 비해서 크다는 것을 의미한다. 이 성질을 고려하면 제안된 알고리즘에서 루머 근원 주위에 노드에게 정체성/방향 질문을 많이 할당하게 되는 경우 자신이 근원이 아니라고 말하는 경우가 많이 발생하게 되고 이를 통해 어디서 루머를 듣게 되었는지에 대해 더 많은 횟수로 대답을 줄 수 있다. 즉, 다수결 원리를 적용해서 실제 루머의 근원을 찾게 되는 확률이 제안된 알고리즘에서 동일하게 물어보는 방법보다 확률적으로 더 잘 추려 낼 수 있게 된다는 것이다. 따라서 (ii)가 성립하게 되고 최종적으로 정리 1에 대한 증명이 된다.

4.2 루머 중심성 기반 상호작용 질의 알고리즘 (Rumor Centrality based Interactive Query Algorithm)

루머 중심성 기반 상호작용 질의 알고리즘도 다음과 같이 3단계로 이루어져 있다.

1) 초기 후보 노드 선정

상호작용 질의방법에서는 최초로 누구에게 물어보는 것을 정하는 것이 중요하다. 본 연구에서는 루머 중심성이 가장 높은 루머 중심을 최초의 노드로 선정해서 정체성/방향 질문을 한다. 이런 방법에서는 앞의 집단 질의와는 다르게 알고리즘에서 총 몇 명의 노드가 질문에 대한 대답자로 선정되는지 알 수 없다. 이것은 네트워크 관리자가 소유한 총 자산을 어떤 비율로 나눠주는지에 대한 기준을 잡기가 어려워지게 되므로 여기서는 i번째로 지목되는 노드에게 K_i번의 질문을 한다고 하면 다음과 같은 식이 만들어 진다.

$$\sum_{i=1}^{M} K_i \leq K. \tag{9}$$

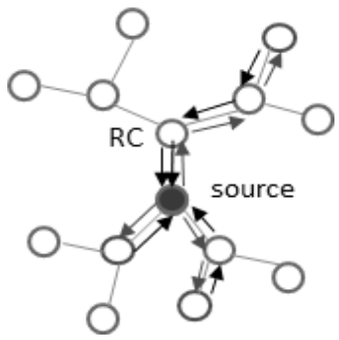

Fig. 5. Interactive Query from the Rumor Center

여기서 M은 네트워크 관리자가 주어진 자산을 가지고 물어보게 되는 노드의 수이고 랜덤변수(Random variable)이다. M이 크다는 것은 많은 노드에게 작은 양의 질문을 각 노드에게 할 수 있고 반대로 M이 작으면 적은 노드들에게 많은 질문을 할 수 있다. 초기 루머 중심으로부터 시작하여 선택되는 노드의 루머 중심성에 따라 K_i를 할당하는 방법은 전체 남은 자산에서 선택된 노드가 가진 루머 중심성이 전체에서 얼마의 비율인지에 따라 질문의 수가 결정이 된다. 이를 수학적으로 나타내기 위해 먼저 표현의 간결성을 위해서 $x=\eta_j(K-\sum_{i=1}^{i-1}K_i)$라고 하면

$$K_j=\begin{cases}\lfloor x\rfloor & \text{if } x\geq 1,\\ 1 & \text{if } 0<x<1.\end{cases} \tag{10}$$

여기서, η_j는 루머 중심성 기반 할당 비중으로 주어진다.

$$\eta_j=\frac{R(v_j,T_{bfs}(v_j))}{\sum_{i=1}^{N}R(v_i,T_{bfs}(v_i))} \tag{11}$$

2) 대답 정보

알고리즘은 처음에 루머 중심을 선택해서 $\lfloor\eta_1 K\rfloor$ 만큼의 정체성/방향 질문을 한다. 각각의 동일한 정체성 질문에 대해서는 진실을 말할 확률 p로 매번 독립적으로 대답을 하고 방향성 질문에 대해서는 확률 q로 진실을 말한다고 할 때, 집단 질의방법에서와 같이 불확실성이 있는 대답에 대해서 다수결 법칙에 의해서 필터링 한다. 즉, 정체성 질문에 대해서 만약 전체 받은 질문 중 자신이 근원이라고 대답한 수가 과반수가 넘으면 그 노드는 정체성 질문에 대한 최종 후보의 집합 I에 속하게 되고 반대의 경우는 여기서 제외된다. 그리고 방향 질문에 대한 대답도 마찬가지로 다수결의 원칙에 따라 가장 많이 지적당한 이웃 노드가 그 노드에게로 정보를 준 노드로 선정하게 되고 선택된 이웃 노드부터 네트워크 관리자가 남은 자산으로 동일한 작업을 반복한다. 각 노드를 중심으로 최종적으로 지적을 가장 많이 받는 노드가 방향 질문에 대한 최종 후보 집합 D에 속하게 된다.

3) 최종 근원추정

마지막으로 최종적인 근원추정 방법은 위에서 얻은 두 질문에 대한 최종 후보 집합 I와 D에 속한 노드 중에 루어 중심성이 가장 큰 노드를 근원으로 추정하는 방법을 택하고 수학적으로 표현하면 아래와 같다.

$$\hat{s}_I=\arg max_{v\in I\cap D}R(v,T_{bfs}(v)) \tag{12}$$

이를 바탕으로 제안한 상호작용 질의 자산할당 방법이 만약 자산이 충분히 커지는 경우에서는 동일하게 물어보는 방법보다 항상 좋다는 것을 다음과 같이 정리할 수 있다.

정리 4.2 (상호작용 질의) 정규트리 G에서 $P(\hat{s}_I=s^*|\vec{r}_{he})$와 $P(\hat{s}_I=s^*|\vec{r}_{ho})$를 각각 제안한 자산할당 방법과 동일한 횟수로 질의하는 방법에 대한 근원 탐지 확률이라고 하자. 만약 질의에 선택되는 노드의 수가 같으면 다음이 성립한다.

$$\lim_{K\to\infty}P(\hat{s}_I=s^*|\vec{r}_{he})\geq\lim_{K\to\infty}P(\hat{s}_I=s^*|\vec{r}_{ho}) \tag{13}$$

위의 결과는 상호작용 질의방법에서 만약 질의를 위해 선택이 되는 노드들의 수가 같은 경우에 대하여 자산의 수가 충분히 많은 경우에는 제안한 방법이 동일한 횟수로 질의 하는 방법에 비해서 훨씬 효율적이라는 것을 의미한다.

증명. 본 정리에도 증명은 크게 두 가지 확률을 얻음으로서 접근이 가능하다. 다시 말하면 탐지확률은 다음과 같이 조건부 확률로 표현할 수 있다.

$$P(\hat{s}_I=s^*)=P(\hat{s}_I=s^*|s^*\in C_r^*)P(s^*\in C_r^*) \tag{14}$$

여기서, 먼저 Equation (14)의 두 번째 확률은 정리 1의 증명과 동일한 이유로 루머 중심성 기반의 방법이 확률이 최대화 된다. 따라서, 정리 1을 보이기 위해서는 첫 번째 확률에 대해서 제안한 자산할당 방법이 동일한 횟수의 방법보다 항상 좋다는 것을 보이면 된다. 이를 위해서 질의를 위해 선택된 노드들에서 여러 번 물어보는 질의에 대한 대답을 다수결 법칙에 따라 추론 되는 경우 최종적으로 루머의 근원이 본문에서 정의된 집합 $I\cap D$에 들어갈 확률을 따지는 것이다. 따라서, $P(s^*\in I\cap D)$를 보면

$$\lim_{K\to\infty}P(s^*\in I\cap D|\vec{r}_{he})\geq\lim_{K\to\infty}P(s^*\in I\cap D|\vec{r}_{ho}) \tag{15}$$

가 성립이 되는 것을 알 수 있는데 그 이유는 다음 두 가지로 설명이 된다.

(i) $\lim_{K\to\infty}P(s^*\in I|\vec{r}_{he})\geq\lim_{K\to\infty}P(s^*\in I|\vec{r}_{ho})$: 제안한 자원할당 방법은 초기 질의 노드로 선택되는 루머 중심노드에게 일단 자산의 많은 양을 사용하여 정체성/방향 질의를 한다. 만약, 루머 중심노드가 루머 근원인 경우에는 위의 정리 1에서의

증명과 같이 다수결 방법에 의해 많이 물어 볼수록 근원이 I에 속할 확률이 커진다. 반대로 만약, 루머 중심이 근원이 아닌 경우라면 루머 중심으로부터 방향성 질문에 대한 대답을 많이 확보할 수 있고 이는 실제 근원이 있는 방향으로 더 정확하게 추려낼 수 있다. 루머 중심의 이웃노드 중 하나가 다음 질의 노드로 선택이 되어도 루머 중심성이 클 확률이 크기 때문에 다수결 방법을 적용하면 실제 근원을 밝혀내거나 아니면 그 방향으로 찾아가는데 더 효율적이다. [1]에 의해 일반적으로 루머 근원은 루머 중심에 높은 확률로 가까이 있다. 따라서 루머 근원이 질의 노드로 선택 되었을 경우에 동일한 횟수로 질의 하는 것보다 더 많은 양의 질의가 할당될 확률이 크게 된다. 그러므로 (i)이 성립한다.

(ii) $\lim_{K\to\infty} P(s^*\in D|\vec{r}_{he}) \geq \lim_{K\to\infty} P(s^*\in D|\vec{r}_{ho})$: 앞의 (i)에서 언급한 바와 같이 루머 중심노드로 부터 만약 질의 받는 노드가 실제 근원이 아니면 방향성 질문에 대한 많은 대답을 듣게 된다. 이를 통해서 상호작용 질의 방법에서는 더 효율적으로 루머 근원이 있는 쪽으로 찾아갈 수 있다. 그리고 $s^*\in D$ 부분에 대해서는 루머 근원이 이웃노드로 부터 방향성질문에 대해 그 부모노드로 많이 지적을 받아야 하는데 일단 루머 중심 근처에 높은 확률로 루머 근원이 있기 때문에 루머 근원의 이웃도 높은 확률로 루머 중심성이 크게 된다. 즉, 이웃노드가 동일한 횟수로 질의하는 방법에 비해 상대적으로 많은 정체성/방향 질의를 받게 되면 루머 근원으로부터 루머를 듣게 되었다는 답을 얻을 확률이 커지고 이는 $s^*\in D$에 대한 확률이 동일한 횟수로 질의하는 방법에 비해 더 크게 된다. 따라서 (ii)가 성립하게 되고 최종적으로 정리 2에 대한 증명이 된다.

5. 분석 및 탐지 실험

본 장에서는 앞서 제안한 두 가지 알고리즘에 기반 하여 시뮬레이션을 통해 실제 그래프에서 정보의 근원을 얼마나 잘 찾을 수 있는가에 대한 실험을 진행 하였다. 실험을 위해서 C언어로 만들어진 응용 프로그래밍인 Matlab을 사용하였고 각 그래프에서 총 1,000 노드가 감염이 되도록 근원으로부터 정보를 확산시켰고 이런 확산 현상을 총 600회를 진행한 후에 성능분석을 하였다. 고려된 네트워크와 기본적인 실험 세팅에 대한 설명은 다음과 같다.

5.1 그래프(네트워크) 세팅 및 탐지확률

1) 그래프 종류 및 세팅

본 실험에서는 다음과 같은 세 가지 종류의 네트워크에 대한 그래프를 고려한다. (1)먼저, 가장 간단한 그래프인 정규 트리를 고려하였다. (2)둘째로, 트리보다는 좀 일반적일 수 있는 합성그래프(Synthetic graph)에 대해서 살펴보았는데, 가장 잘 알려진 합성 그래프 중에서 ER 랜덤 그래프에 대해서 살펴보았다. ER 랜덤 그래프 같은 경우 네트워크에 노드가 주어지고 각 노드가 연결될 확률이 p>0 인 베르누이 분포를 따라 동전던지기를 하여 성공을 하면 연결하고 실패를 하면 연결을 하지 않는 방법으로 구성된 랜덤 그래프이다. 본 실험에서는 전체 그래프의 노드가 3,000 개인 경우 한 노드가 가진 평균 차수가 4가 되도록 형성시켰다. (3)마지막으로, 실험에서 고려한 그래프는 실제 데이터를 가지고 만들어진 그래프(Real-world graph)로서 페이스북 그래프[17]를 생성시켜서 실험을 하였다. 페이스북 네트워크인 경우는 총 4,039노드가 88,234개의 변으로 구성되어 있는 그래프를 Fig. 6과 같이 몇 개의 집단(cluster)이 그 안에서 긴밀하게 연결되어 있고 (많은 변들이 서로 연결됨) 이런 집단끼리는 각 집단의 내부보다는 덜 긴밀하게 연결이 되도록 형성하였다(실제 노드들이 연결되어 있는 모양이다). 위와 같이 형성된 그래프 위에서 랜덤하게 정보를 퍼뜨리는 근원을 선택하여 SI 모델에 따라 네트워크에서 전체 확산된 노드의 수가 1,000이 되도록 한 후 제안한 추정법이 실제 근원을 찾는지 아닌지를 확인하는 실험을 진행하였다.

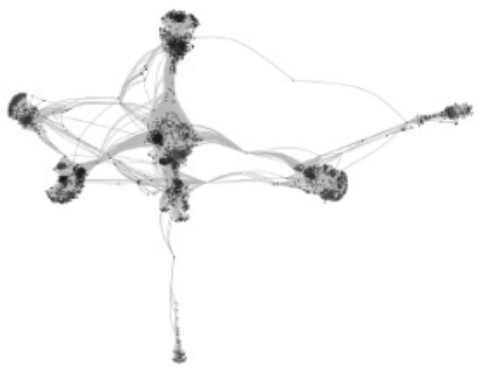

Fig. 6. Facebook Network[2]

2) 탐지확률

집단 질의의 탐지확률의 척도는 아래와 같은 식을 사용하였는데 이는 전체 시도된 정보의 확산 중에서 실제로 제안한 추정법이 원래 정보의 근원을 찾는 수로 나눠진 것이다.

$$\text{Detection Probability} = \frac{\#\text{of Detections}}{\#\text{of Trials}} \tag{16}$$

본 실험에서는 총 600번의 확산현상 중에서 정확히 정보의 근원을 찾는 횟수를 나누어 각 경우에 대한 탐지확률을 계산하였다.

5.2 비교대상 알고리즘

제안한 알고리즘의 탐지확률에 대한 성능비교의 대상으로

두 가지 질의방법에 대해서 아래와 같은 방법들을 실험에서 고려하였다.

1) 집단 질의

집단 질의의 경우에는 먼저 (1)초기 후보 노드 선택방법에 따라 그리고 (2)몇 회의 동일한 질문을 할 것인가에 따라 다음과 같이 4가지의 방법이 고려된다.

a) Random

이 방법은 네트워크에 N개의 노드에 루머가 퍼진 경우 초기 후보노드를 무작위(Random)로 선택하여 집단 질의를 동일한 수 만큼 물어보고 얻은 데이터를 통해 근원을 찾는 방법을 말한다. 비교 대상이 되는 알고리즘 중에 가장 기본적인 방법이다.

b) Hop (Homogeneous)

이 방법은 선행연구[2, 3]에서 사용했던 방법으로서 초기 후보노드를 루머 중심성에 기반 하여 선택하는 것이 아니라 루머 중심을 중심으로 일정한 거리만큼 떨어진 감염된 노드들을 후보자로 선택하고 동일한 수의 질문을 해서 얻은 데이터를 가지고 근원을 찾는 방법이다.

c) RC (Homogeneous)

이 방법은 b)에서 거리에 따라 후보자를 선택하는 것 대신 루머 중심성이 큰 노드를 중심으로 후보자를 선택하여 동일한 횟수의 질문을 통해 데이터를 얻어서 근원을 찾는 방법이다.

d) RC (Proposed)

이 방법은 c)에서 루머 중심성이 큰 노드를 중심으로 후보자를 선택하여 얻은 후 루머 중심성에 기반 한 질문의 수를 앞에서 제안한 방법대로 적절히 조절해서 한 후 데이터를 얻고 근원을 찾는 방법으로서 본 연구에서 가장 우위에 두고 있는 알고리즘이다.

2) 상호작용 질의

상호작용 질의의 경우에는 (1) 초기 노드를 어떻게 선택하는지에 따라 (2) 선택된 노드에게 얼마나 많이 물어볼지에 따라 다음과 같이 3가지의 방법을 고려한다.

a) Random

이 방법은 초기노드를 임의적으로 선택하여 질문을 통해 얻은 데이터를 가지고 다음 노드를 선택하고 각 노드에게 같은 양의 질문을 하는 방법을 의미한다.

b) RC (Homogeneous)

이 방법은 초기 노드를 루머 중심을 선택하여 질문을 하고 다음노드들을 얻은 데이터를 통해 선택해가는 방법으로서 각 선택되는 노드마다 동일한 횟수의 질문을 하는 선행연구[2, 3]에서 제안된 방법이다.

c) RC (Proposed)

이 방법은 b)에서와 같이 초기 노드를 루머센터를 선택하여 물어보면서 얻은 데이터를 가지고 다음 노드를 선택하는 것에 각각 물어보는 횟수가 Equation (10)과 같이 노드가 가진 루머 중심성에 비례하게끔 물어보는 본 연구에서 제안된 방법의 알고리즘이다.

5.3 실험 결과

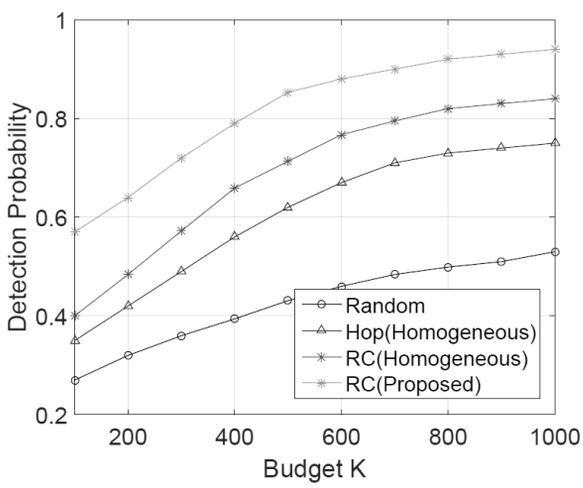

Fig. 7. Detection Probability for Batch Query on Regular Tree (degree = 3)

1) 정규 트리

먼저 Fig. 7에서 $p=q=2/3$인 경우에 대하여 집단 질의에 대한 4가지 알고리즘들에 대한 결과를 얻었다. 본 실험은 차수가 3(d=3)인 경우에 대해서 진행이 되었는데, 그림에서 볼 수 있듯이 4가지 알고리즘 모두 네트워크 관리자가 질문을 많이 하게 되는 경우 즉, 자산 K가 증가하는 경우 근원에 대한 탐지 확률이 증가한다는 것을 알 수 있다. 그 이유는 질의와 같은 추가적인 질문은 항상 네트워크에서 퍼진 정보에 대한 최우추정량으로만 근원을 추정하는 것보다 더 많은 정보를 제공해 주시 때문이다. 하지만, 그림에서와 같이 초기 후보노드를 선택하는 방법이 무작위인 경우에는 자산에 비해 탐지 확률이 크게 증가가 되지 않는 반면, 퍼진 정보에 대한 루머센터를 중심으로 일정한 거리 혹은 루머 중심성을 따라 선택하는 경우에는 많은 증가양상을 확인 할 수 있다. 이는 기존의 연구[1]에서 최우추정량과 실제 근원은 높은 확률로 일정한 거리 안에 있게 된다는 것이 알려졌고 거리에 따라 선정하든 루머 중심성을 기반으로 선정하든 실제 근원이 선정된 노드 안에 있을 확률이 커지기 때문이다. 그리고 또한,

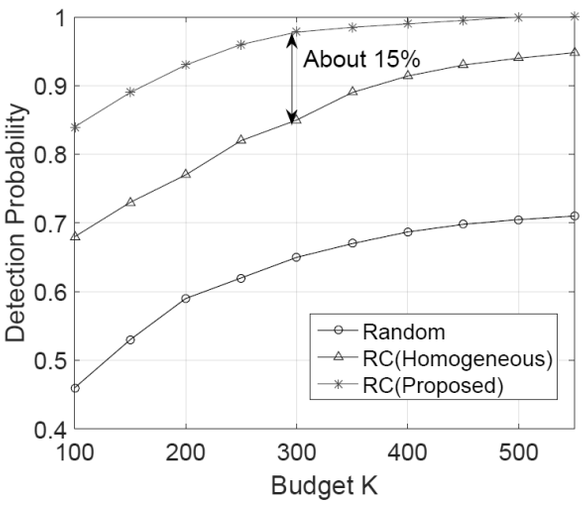

Fig. 8. Detection Probability for Interactive Query on Regular Tree (degree=3)

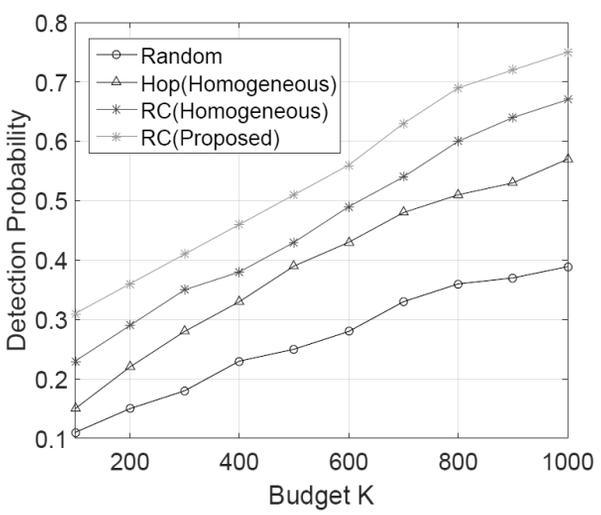

Fig. 9. Detection Probability for Batch Query on ER Random Graph

거리로만 초기 후보노드의 기준을 잡고 선정하는 것보다 실제 루어 중심성을 가지고 선정하는 것이 더 좋고 마지막으로, 동일하게 질문을 해서 진실을 가려내는 것보다 루머 중심성이 큰 노드 즉, 실제 근원이 될 가능성이 큰 노드에게 더 많이 물어보는 방법이 대략 10% 이상으로 탐지확률을 높혀 준다는 것을 확인할 수 있다.

다음으로 Fig. 8에서는 정규트리에서 상호작용 질의의 경우에 대해 3가지 알고리즘에 대한 탐지확률을 얻었다. 먼저, 알 수 있는 사실은 상호작용으로 질문을 하는 방법이 집단으로 하는 방법보다 훨씬 효율적이라는 사실인데, 이는 훨씬 더 작은 자산으로 더 높은 탐지확률을 나타내는 것을 통해 확인할 수 있다. 그 이유는 집단 질의는 실제 근원이 초기후보 노드집합 안에 있기 위해서 상대적으로 상호작용 방법보다 더 많은 양의 초기 후보자를 선정하게 되기 때문이다. 즉, 네트

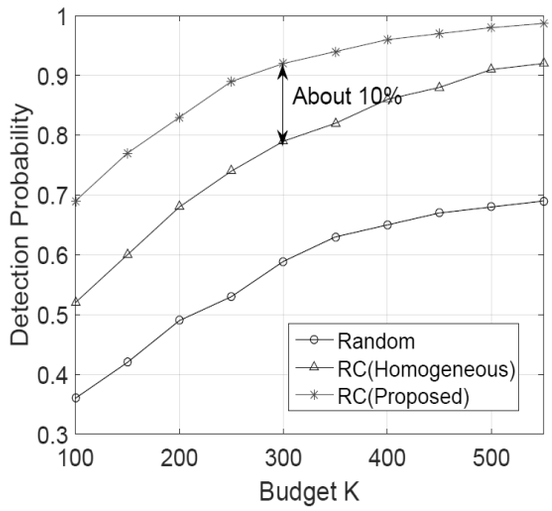

Fig. 10. Detection Probability for Interactive Query on ER Random Graph

워크 및 그래프에서 상호작용 질의방법은 루머 중심을 중심으로 몇 개의 경로를 집중적으로 확인해서 찾기 때문에 더 효율적이다. 또한 결과에서 알 수 있듯이, 초기 노드를 선정할 때, 무작위로 선정하는 것 보다는 확산된 루머의 형태를 보고 최우추정량을 기준으로 질의를 진행하는 것이 더 좋고 또한 동일하게 물어보는 것보다 루머 중심성에 대한 정보를 가지고 이에 비례하여 물어보는 것이 더 좋다는 것을 확인할 수 있었다.

2) ER 랜덤 그래프

본 실험에서는 $p=0.6$, $q=0.3$의 진실을 말할 확률에 대해서 실험을 진행하였다. Fig. 9는 정규트리에서와 같이 집단질의에 대해 4가지 다른 알고리즘에 대해서 결과를 얻었는데 이 경우에서도 자산의 양이 커질수록 탐지확률이 커진다는 사실을 먼저 확인할 수 있었다. 그리고 루머 중심에서의 거리를 바탕으로 초기 후보노드를 선택하는 것보다 본 연구에서 제안한 루머 중심성을 기반으로 후보 노드를 선택하고 이에 비례하여 질의를 해 주는 방법이 다른 것에 비해서 가장 성능이 좋다는 것을 확인 할 수 있었다.

다음으로 Fig. 10에서는 ER 랜덤그래프에서 상호작용 질의가 있는 경우에 대한 3가지 알고리즘에 대한 결과를 보여주고 있다. 정규트리에서와 같이 집단으로 선택하여 질의를 하는 것에 비해서 더 적은 자산으로 높은 탐지확률에 대한 결과를 확인 할 수 있었고, 제안한 알고리즘이 선행연구[2, 3]에서 보여줬던 방법에 비해서 대략 10%정도 성능이 향상이 되는 것을 확인할 수 있었다.

3) 페이스북 네트워크

마지막으로, 실제 인터넷에 있는 데페이스북 그래프에 대한 데이터를 바탕으로 네트워크를 형성한 후 두 질의방법에

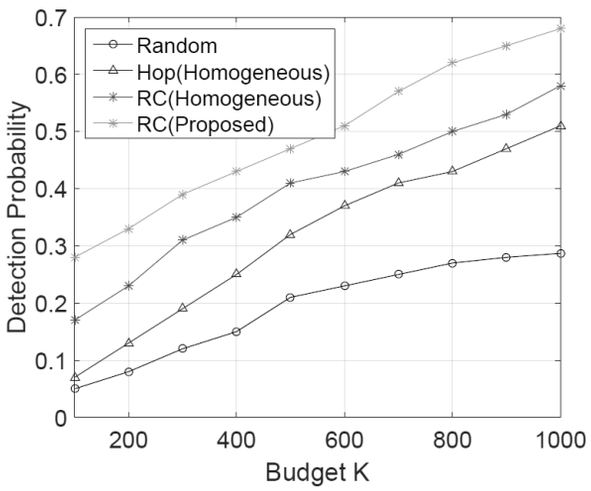

Fig. 11. Detection Probability for Batch Query on Facebook Graph

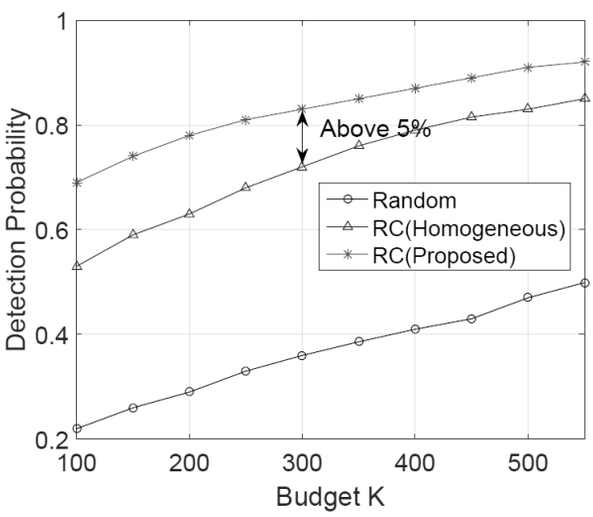

Fig. 12. Detection Probability for Interactive Query on Facebook Graph

대한 결과를 얻었다. 본 실험에서는 ER 에서와 같이 $p=0.6$, $q=0.3$의 정답을 말할 확률을 선택하여 진행하였다. Fig. 11 에서는 집단 질의방법에 대해 탐지확률에 대한 결과를 얻었는데, 앞에서의 정규 트리와 ER 랜덤그래프와 비교해서 가장 작은 값을 보여주고 있다. 그 이유는 페이스북 그래프는 비교적 간단한 앞의 두 그래프와 다르게 노드들이 굉장히 많이 연결이 되어 있는 복잡한 그래프여서 자산이 많아져도 네트워크 구조가 가진 때문에 앞에서 언급한 BFS 방법으로 실제 근원의 위치를 추정하는데 한계가 있기 때문이다. 하지만, 본 연구에서 제안한 방법이 기존연구들에 대한 결과에 비해 월등히 높은 탐지확률을 가지는 것을 볼 수 있고 이런 실제 네트워크에서도 비록 근사적인 방법일 수 있으나 BFS 기반 루머 중심성에 기반 하여 각 노드에게 적절한 양의 질문을 하면 근원을 찾는데 더 도움이 된다는 사실을 보여준다.

Fig. 11에서는 집단 질의방법에 대해 탐지확률에 대한 결과를 얻었는데, 앞에서의 정규 트리와 ER 랜덤그래프와 비교해서 가장 작은 값을 보여주고 있다. 그 이유는 페이스북 그래프는 비교적 간단한 앞의 두 그래프와 다르게 노드들이 굉장히 많이 연결이 되어 있는 복잡한 그래프여서 자산이 많아져도 네트워크 구조가 가진 때문에 앞에서 언급한 BFS 방법으로 실제 근원의 위치를 추정하는데 한계가 있기 때문이다. 하지만, 본 연구에서 제안한 방법이 기존연구들에 대한 결과에 비해 월등히 높은 탐지확률을 가지는 것을 볼 수 있고 이런 실제 네트워크에서도 비록 근사적인 방법일 수 있으나 BFS 기반 루머 중심성에 기반 하여 각 노드에게 적절한 양의 질문을 하면 근원을 찾는데 더 도움이 된다는 사실을 보여준다.

Fig. 12에서는 페이스북 네트워크에서 상호작용 질의가 적용되는 경우 제안한 알고리즘의 탐지확률을 보여주고 있는데, 비록 페이스북 같이 복잡한 네트워크이지만, 집단으로 물어보는 것보다 한 노드씩 추적하면서 물어보는 상호작용 방법이 훨씬 효율적이라는 사실을 확인 할 수 있었고, 제안한 방법이 다른 선행연구에 대한 결과보다 대략 5% 정도의 성능 향상을 보였다.

Table 2. Detection Probabilities for Batch Query using Various (p,q)(K=500)(H: Homogeneous, P: Proposed)

	(p,q)	Random	Hop(H)	RC(H)	RC(P)
Regular Tree (d=3)	(2/3,2/3)	0.41	0.60	0.71	0.82
	(3/4,3/4)	0.48	0.69	0.78	0.86
	(4/5,4/5)	0.53	0.76	0.84	0.89
ER	(0.6,0.3)	0.25	0.39	0.42	0.51
	(0.8,0.6)	0.31	0.46	0.51	0.59
	(0.9,0.9)	0.33	0.51	0.58	0.64
Facebook	(0.6,0.3)	0.20	0.31	0.40	0.46
	(0.8,0.6)	0.22	0.37	0.45	0.52
	(0.9,0.9)	0.24	0.41	0.49	0.55

Table 3. Detection Probabilities for Interactive Query using Various (p,q)(K=500)(H: Homogeneous, P: Proposed)

	(p,q)	Random	RC(H)	RC(P)
Regular Tree (d=3)	(2/3,2/3)	0.70	0.91	0.99
	(3/4,3/4)	0.73	0.94	0.99
	(4/5,4/5)	0.75	0.96	0.99
ER	(0.6,0.3)	0.67	0.90	0.97
	(0.8,0.6)	0.68	0.91	0.98
	(0.9,0.9)	0.69	0.92	0.98
Facebook	(0.6,0.3)	0.42	0.81	0.84
	(0.8,0.6)	0.45	83	0.84
	(0.9,0.9)	0.47	0.84	0.86

마지막으로, Table 2와 Table 3에서 각 네트워크 토폴로지들에 대해서 각 노드가 정체성/방향 질문에 대한 진실을 말할 확률을 다양하게 변화시켜서 근원 탐지 확률을 구하였다. 기본적으로, 진실을 말할 확률이 높으면 실제 근원을 찾을 확률도 더 커지는데 그 이유는 각 알고리즘에서 얻은 대답의 데이터를 다수결 법칙으로 필터링을 할 때, 더 적은 질문으로도 더 정확한 정보를 추려 낼 수 있기 때문이다. 그리고 집단 질의방법에 비해서 상호작용 질의방법이 다양한 정답을 말할 확률이 주어진 경우에도 훨씬 효율적이라는 사실을 확인할 수 있다.

6. 결 론

본 논문에서는 네트워크에서 루머와 같은 정보가 퍼진 경우 그 근원이 어디인지를 찾아내는 방법에 대한 연구로 추가적인 질문을 통해서 보다 더 잘 찾아낼 수 있는 방법에 대해 연구하였다. 각 노드가 질문에 대해서 확률적으로 불확실한 정답을 알려주는 상황에서 네트워크에 퍼진 정보를 통해 보다 더 근원일 것 같은 노드에게 질문을 더 많이 함으로서 실제 근원을 찾는 데 있어서 그 정확성을 기존에 제안된 선행 연구들에 비해 많이 높이는 결과를 얻었다. 특히, 본 논문에서 제안된 자원 할당 알고리즘은 집단 질의 방법에서 정규 트리인 경우 대략 10% 이상의 증가율을, ER 랜덤 그래프 및 페이스북 그래프에서는 대략 5% 이상의 증가율을 보였다. 또한, 상호작용 질의 방법에서는 정규 트리인 경우 대략 15% 이상의 증가율을, ER 랜덤 그래프 및 페이스북 그래프에서는 대략 5-10% 의 증가율을 확인할 수 있었다. 본 논문에서 제안된 자원 할당 방법이 동일한 횟수나 그냥 랜덤하게 질의하는 것보다 성능이 더 좋은 이유는 루머가 퍼져있는 정보를 효과적으로 사용하였기 때문이었다. 이는 루머의 근원에 대해 근사적으로 사후확률을 알려주고 있고 이와 비례하여 자원할당 하는 것이 루머의 퍼진 모습과 상관없이 자원을 사용하는 것보다 더 좋다는 것을 이론적으로 밝혔다.

후속 연구로는, 각 노드가 가진 진실을 말할 확률분포가 사전에 여러 데이터를 통해 주어진 경우 동일한 확률로 진실을 말해 주는 것이 아닌 서로 다른 확률로 질의에 대한 대답을 할 때 자원할당을 어떻게 하는 것이 더 효율적일 수 있는지를 살펴볼 계획이다.

References

[1] D. Shah and T. Zaman. Detecting Sources of Computer Viruses in Networks: Theory and Experiment. *In Proceedings of ACM SIGMETRICS,* 2010.

[2] J. Choi, S. Moon, J. Woo, K. Son, J. Shin, and Y. Yi. Rumor, "Source Detection under Querying with Untruthful Answers. *In Proceedings of IEEE INFOCOM,* 2017.

[3] J. Choi and Y. Yi, "Necessary and Sufficient Budgets in Information Source Finding with Querying: Adaptivity Gap," in *Proceedings of IEEE ISIT,* 2018.

[4] W. Dong, W. Zhang, and C. W. Tan, "Rooting Out the Rumor Culprit from Suspects." in *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, 2013.

[5] Z. Wang, W. Dong, W. Zhang, and C. W. Tan, "Rumor Source Detection with Multiple Observations: Fundamental Limits and Algorithms," in *Proceedings, ACM SIGMETRICS,* 2014.

[6] J. Choi, S. Moon, J. Shin, and Y. Yi, "Estimating the Rumor Source with Anti-Rumor in Social Networks," in *Proceedings of IEEE ICNP Workshop on Machine Learning,* 2016.

[7] G. Fanti, P. Kairouz, S. Oh, and P. Viswanath, "Spy vs. Spy: Rumor Source bfuscation," in *Proceedings of ACM SIGMETRICS,* 2015.

[8] G. Fanti, P. Kairouz, S. Oh, K. Ramchandran, and P. Viswanath, "Rumor ource Obfuscation on Irregular Trees," in *Proceedings of ACM SIGMETRICS,* 2016.

[9] G. Fanti, P. Kairouz, S. Oh, K. Ramchandran, and P. Viswanath, "Metadata-conscious Anonymous Messaging," in *Proceedings of ICML,* 2016.

[10] W. Luo, W. P. Tay and M. Leng, Infection Sprading and Source Identification: A Hide and Seek Game. *IEEE Transaction on Signal Processing,* Vol. 64, No. 16, AUGUST 15, 2016.

[11] J. Jaing, S. Wen, S. Yu, Y. Xiang, and W. Zhou, "K-Center: An Approach on the Multi-Source Identification of Information Diffusion," *IEEE Transactions on Information Forensics and Security,* Vol.10, pp.2616-2626, 2015.

[12] F. Ji and W. P. Tay, "An Algorithmic Framework for Estimating Rumor Sources With Different Start Times," *IEEE Transactions on Signal Processing,* Vol.65, pp. 2517-2530, 2017.

[13] Z. Wang, C. Wang, J. Pei, and X. Ye, "Multiple Source Detection without Knowing the Underlying Propagation Model," in *Proceedings of AAAI,* 2017.

[14] W. Luo, W. P. Tay, and M. Leng, "How to Identify an Infection Source With Limited Observations," *IEEE Journal of Selected Topics in Signal Processing,* Voi.8, No.4, pp. 586-597, 2014.

[15] K. Zhu and L. Ying, "Information Source Detection in Networks: Possibility and Impossibility Results," *IEEE INFOCOM,* 2017.

[16] B. Chang, F. Zhu, E. Chen, and Q. Liu, "Information source detection via Maximum A Postreia Estimation,?" in *Proceedings of IEEE ICDM,* 2015.

[17] J. Leskovec and J. McAuley, "Learning to Discover Social Circles in Ego Networks," in *Proceedings of NIPS,* 2012.

최 재 영

https://orcid.org/0000-0001-9118-8050
e-mail : jychoi@honam.ac.kr
2008년 고려대학교 수학과(학사)
2013년 고려대학교 수학과(석사)
2018년 한국과학기술원 전기 및 전자공학부(Ph.D.)
2018년~현 재 호남대학교 미래자동차공학부 조교수
관심분야 : 소셜네트워크, 데이터 마이닝, 통계적 추론, 자율주행자동차, 네트워크 보안