JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 24, No. 5: 567~575, May. 2020

그래프 속성을 이용한 온라인 소셜 네트워크 스팸 탐지 동향 분석

정시현[1] · 오하영[2*]

Exploratory study on the Spam Detection of the Online Social Network based on Graph Properties

Sihyun Jeong[1] · Hayoung Oh[2*]

[1]Ph.D Candidate, Department of Computer and Engineering, Seoul University, Seoul, 08826 Korea
[2*]Assistant Professor, Global Convergence, Sungkyunkwan University, Seoul, 03063 Korea

요 약

온라인 소셜 네트워크가 현대인의 정보 공유 및 교류의 핵심적인 매체로 사용됨에 따라, 그 이용자는 매해 급격하게 증가하고 있다. 이는 단순히 사용량 증가뿐만 아니라 정보의 신뢰성에서도 기존 언론 매체를 능가하기도 하는데, 최근 등장하는 마케팅 전략들은 이 점을 노리고 교묘하게 소셜 네트워크를 공격하고 있다. 그에 따라 자연스럽게 형성되어야 할 여론이 온라인 공격으로 인해 인위적으로 구성되기도 하고, 이를 신뢰하는 사람들도 많아지게 되었다. 따라서 온라인 소셜 네트워크를 공격하는 주체들을 탐지하고자 하는 연구들이 최근 많이 진행되고 있다. 본 논문에서는 이러한 온라인 소셜 네트워크 공격자들을 탐지하고자 하는 연구들의 동향을 분석하는데, 그 중 소셜 네트워크 그래프 특성을 이용한 연구들에 집중하고 있다. 기존의 contents-based 기법이 사생활 침해 및 공격 전략 변화에 따른 분류 오류를 나타낼 수 있음에 반해, 그래프 기반 방법은 공격자 패턴을 이용하여 보다 강건한 탐지 방법을 제안하고 있다.

ABSTRACT

As online social networks are used as a critical medium for modern people's information sharing and relationship, their users are increasing rapidly every year. This not only increases usage but also surpasses the existing media in terms of information credibility. Therefore, emerging marketing strategies are deliberately attacking social networks. As a result, public opinion, which should be formed naturally, is artificially formed by online attacks, and many people trust it. Therefore, many studies have been conducted to detect agents attacking online social networks. In this paper, we analyze the trends of researches attempting to detect such online social network attackers, focusing on researches using social network graph characteristics. While the existing content-based techniques may represent classification errors due to privacy infringement and changes in attack strategies, the graph-based method proposes a more robust detection method using attacker patterns.

키워드 : 소셜 네트워크, 공격 탐지, 그래프 특성, 패턴 분석, 스팸

Keywords : Social Network, Attack detection, Graph property, Pattern analysis, Spam

Received 22 October 2019, **Revised** 11 December 2019, **Accepted** 31 December 2019

*** Corresponding Author** Hayoung Oh (E-mail: hyoh79@gmail.com Tel:+82-2-583-8585)
Assistant Professor, Global Convergence, Sungkyunkwan University, Seoul, 03063 Korea

Open Access http://doi.org/10.6109/jkiice.2020.24.5.567 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

SNS(Social Networking Services)는 유저들 간의 신뢰를 바탕으로 아이템 추천부터 정보 공유까지 다양한 역할을 수행하고 있다. 이는 온라인 소셜 네트워크가 실제 오프라인 소셜 네트워크와 매우 유사한 작용을 하고 있다는 것을 보여 준다. 예를 들면, 우리는 친한 친구 혹은 믿을 수 있을 만한 사람이 추천해 준 제품에 좀 더 관심을 갖기도 하고 그들이 주는 정보를 쉽게 받아들이기도 한다. 이러한 SNS의 파급력은 기업 등의 마케팅 전략을 위해 사용되기도 한다. 트위터의 follower 계정을 다수 구매하여 특정 계정의 follower수를 높여주는 방법이나, Facebook의 다수 계정 구매로 특정 Facebook page의 구독자 수를 늘리는 방법 등이 대표적이다. 하지만 이러한 SNS 활용이 오/남용되면 스팸 (Spam) 컨텐츠로서 SNS의 신뢰성을 크게 저하시키고 결국 서비스의 실패로 이어지게 된다.

현재까지 SNS 스팸 등의 공격 탐지 분야에서 가장 많이 사용된 방법은 크게 Contents-based approach[1,2,3,4]와 Behavior-based approach[5,6,7,8]였다. Contents-based approach는 스팸 탐지 초기에 사용된 방법으로 SNS에서 보내는 메시지나 SNS상의 개인 포스팅에 스팸성 단어가 포함되어 있는가를 분석하여 스패머를 탐지하는 기본적인 방법이다. 하지만 이 방법의 경우 이미지를 이용하거나 피싱 링크를 주로 이용하면서 스팸성이 없는 어휘로 컨텐츠를 구성했을 경우 탐지가 어렵다는 단점이 있다. Behavior-based approach는 Contents-based approach 이후에 등장하였으며, 스팸 탐지를 위해서는 컨텐츠 내용 외에도 개인의 SNS 사용 패턴과 관련된 정보들을 사용하는 방법이다. 예를 들면, 사용자가 주로 사용 언어나 사용 시간대, 포스팅 업로드 시간 간격, 리트윗 수 등의 정보가 있다[1]. 하지만 이 방법의 경우 그룹으로 일관성 있는 패턴의 공격을 할 경우의 탐지에는 적합하나 개별적으로 활동하는 스패머를 탐지하기는 어렵다는 문제가 있다.

소셜 네트워크 그래프 속성 기반 탐지 방법의 초기에는 [9]와 같이 단순히 다수에게 스팸 컨텐츠를 전송하는 유저를 찾는 것을 목적으로 하는 연구가 많았다. 이 경우에는 스패머가 상대적으로 많은 outgoing edge비율을 가진다는 특징이 있을 경우 탐지하기 용이했다. 즉, 스패머는 indegree에 비해 outdegree가 많다는 점을 스팸 탐지에 이용하는 경우가 많았다. 하지만 스패머도 일반 사용자처럼 가장하기 위해 의도적으로 indegree까지 높이는 방법을 사용하고 있기도 하다. 따라서, degree정보 및 상대적인 비율만 이용한 단순한 방법보다는 좀 더 복합적이고 다각도로 스패머 특징을 뽑을 수 있는 그래프 특성을 측정해야 한다.

특히 최근에는 기존보다 좀 더 복잡한 공격 유형들이 등장하면서, 최근 연구들은 크게 특정 attack에 집중하여 해결하는 방법[10,11,12]과 이상 행동을 통한 SNS 오/남용 사례를 탐지하는 방법들[13,14,15]로 나뉘어진다. 특정 attack이라 하면, 일반적으로 탐지하기 힘든 계정 탈취 공격을 대상으로 하는 경우가 많다. 반면, 이상 행동을 탐지하는 연구는 대부분 일반적인 사용자의 SNS 사용 패턴과 이상 행동자, 즉 아웃라이어(Outlier)의 사용 패턴을 비교하는 방법으로 이루어진다.

최근 소셜 네트워크 특성이 각광받고 있는 이유는, 스패머가 아무리 일반 사용자처럼 보이려 하더라도 애초에 SNS 사용 목적이 그들과 다르기 때문에 그 행동 양식은 다르게 나타날 수밖에 없다는 데서 기인한다. 즉, 빅데이터의 특징 및 앙상블 평균 등의 중장기 분석 개념을 반영하면 결국 일반 사용자와 다른 패턴을 보인다는 것이다. 특히, 소셜 네트워크 그래프의 경우 스패머 자신의 의지만으로 변화시키는 것이 아닌 다른 사용자와의 상호작용으로 변화하는 것이기 때문에 현재 스패머 탐지에 가장 유용한 연구 분야라고 할 수 있다. 결과, 다음 장에서 소셜 그래프 기반의 Fraud(혹은 Outlier) 탐지 연구를 어떻게 분류하고, 어떤 연구들이 있는지 살펴보도록 하겠다. 또한, 그래프 특성을 이용해 소셜 미디어에 만연한 계정 탈취 공격을 탐지하기 위한 방법 역시 마지막 장에서 제안한다.

Ⅱ. 연구 동향

소셜 네트워크 그래프 기반 Fraud 탐지는 크게 Local social network 기반[16,17,18]과 Global social network 기반 탐지 방법[13,14,19,20,21,22,23,24]으로 분류할 수 있다. 이 중 Global social network 기반 탐지 방법은 Propagation method[14,19,20]와 Latent factor model [13,21,22,23,24]로 분류할 수 있다. 소셜 네트워크 그래프 기반 접근 방법(그림1) 외에, 최근 연구된 몇 가지

Outlier detection 방법들[25,26,27,28]도 위 분류의 연구들 소개 이후에 기술되었다.

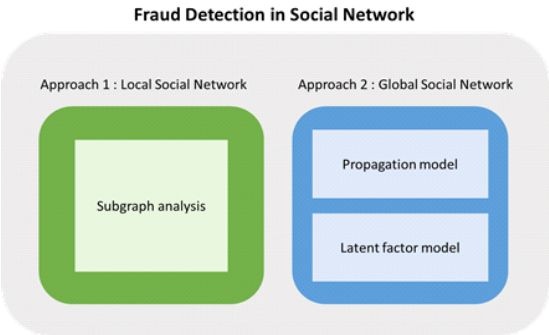

Fig. 1 Fraud detection based on Social network graph

2.1. Local Social Network 기반 Fraud 탐지 방법

Local Social Network 기반 탐지 방법은 Subgraph의 분석을 중요시하는 방법이다. Subgraph란 노드들 (i.e., vertex)의 subset과 이들 간에 연결된 연결선 (i.e., 엣지, edge)들을 의미한다. Subgraph 분석 방법은 Global social network를 대상으로 하지 않고, 주로 Local Social network를 대상으로 한다. Local Social Network의 예로는 Ego-network를 들 수 있다. 이는 앞 장에서 언급한 개인 소셜 네트워크와도 일맥상통하는 부분이다. Ego-network란 중심 사용자인 ego와 이들간의 1-hop 이웃들을 노드로 하고, 이들간에 발생한 관계를 엣지로 하는 서브그래프이다(그림 2).

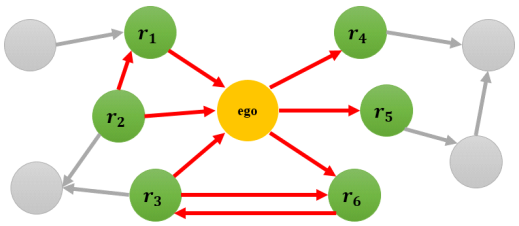

Fig. 2 Ego-Network (green, yellow nodes and red edges) [29]

Ego-network가 많이 사용되는 이유는 특정 사용자의 스패머 여부를 알기 위해서는 이웃 사용자들 간의 관계가 그 특징을 가장 많이 반영하기 때문이다. Homophily 이론에 따르면[30] 비슷한 특성을 공유하는 사용자들 간에는 관계가 발생할 확률이 더 높다.

앞서 언급했듯이, Ego-network는 한 특정 노드와 서로 공통점들이 존재하는 이웃 노드들 (i.e., 그룹별)로 구성된 Subgraph를 의미한다. 결과, 우리가 어떠한 특정 노드가 일반 사용자인지, 혹은 스패머인지 판별하기 위해 Ego-network를 기준으로 분석하게 된다면, 전체 소셜 네트워크를 보는 것보다 비슷한 특성을 공유하는 직접 이웃 사용자들을 참고하여 판단할 수 있다. 분류 특성이 더 강하게 나타날 것이다. 따라서 Local social network 기반 Fraud 탐지라 하면 Ego-network 기반 Fraud 탐지인 경우가 대부분이다. 이들은 공통적으로 Ego-network에서 특정한 패턴의 Subgraph가 있는지 찾는 Graph query를 사용한다. 특정한 패턴의 한 가지 예로 triangle 그래프 패턴을 들 수 있는데, Ego-network 내에 triangle이 얼마나 많이 등장하는지 분석하는 Graph query를 사용할 수 있다. 다음은 Ego-network 기반 Fraud 탐지 방법들이다.

스팸 탐지 기법에 사용된 서브그래프 기반 탐지 방법은 일반적으로 스패머의 특정 공격 패턴을 몇 가지로 분류하고, 그 정의에 맞는 서브그래프 패턴을 미리 정의하는 것부터 시작한다. 대표적으로 [17]는 Youtube스팸을 탐지하기 위해 Spam campaign을 두 종류로 분류하여 각 campaign에 관련된 서브그래프 패턴을 정의하였다 (그림 3). 정의한 서브그래프는 사이즈가 3에서 5 사이의 Network motif profile 측정 방법에 따라 분류하였으며, 그림 3에서 파란색은 비디오를 나타내고, 옅은 주황색은 사용자를 나타낸다.

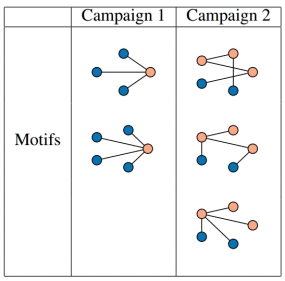

Fig. 3 Campaigns for YouTube spam detection[17]

이와 유사하게, [18]은 The Triad Significance Profile (TSP)을 이용하여 스패머의 ego-network 내에 등장하는 motif의 분포가 일반 사용자의 분포와 다르다는 점에 착안한 스팸 탐지 기법을 제안하였다. TSP란, 13가지

Triad(그림 4)의 각 등장 횟수 분포를 Z score를 이용해 normalized 한 것을 말한다.

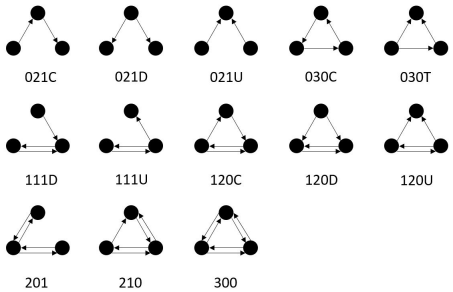

Fig. 4 13 isomorphic triads for Triad Significance Profile (TSP)[18]

[18]에서 사용한 TSP는 [31]에서 이종 네트워크 비교에서 먼저 사용하였다. 이는 비교 대상 그래프에서 등장한 각 Subgraph들이 종류별로 얼마나 많이 등장하였는지 측정하고, random graph로 생성된 null model에서의 등장 비율과 비교하였다. 실제로, 실험 결과 다른 종류의 네트워크 간에는 Subgraph의 등장 비율이 달랐고, 같은 종류 간에는 등장 비율이 유사하였다.

스패머 탐지를 위해서는 단순히 서브그래프 구조를 비교하는 것도 중요하지만, 노드간의 관계 의미성을 파악할 수 있는 role 분석 역시 유용하다. [16]는 Ego-network 패턴을 임의로 정의하였는데, 이는 스패머의 공격 패턴을 정의한 서브그래프가 아닌 neighbor 노드 간에 만들어지는 엣지와 이들의 weight가 발생한 패턴이다. 같은 서브그래프 내에서도 weighted edge가 발생한 패턴이 다르기 때문에 엣지 weight로 인한 노드의 role이 정해질 수 있다.

Oddball[16]은 Web spam을 찾는데 그래프 내에서 triangle 패턴의 등장 비율을 이용하였다. 이처럼 주어진 노드 수에서 발생 가능한 모든 서브그래프 패턴을 이용하는 방법은 대표적으로 Motif와 Graphlet이 있다. Motif는 노드 3개로부터 만들어지는 가능한 모든 서브그래프를 사용하는 방식이고, Graphlet은 [32]에 의해 노드의 role과 관련하여 연구되었는데, 1개부터 5개까지의 노드에서 만들어지는 undirected 서브그래프 패턴을 사용하는 방식이다.

Local social network를 Ego-network가 아닌 dense subgraph 대상으로 하는 연구들도[33,34] 있다. 이 방법은 Global social network를 Adjacency matrix상에 표현 후, 해당 matrix를 재배열하여 dense region을 찾아 이를 Subgraph로 보고 활용하는 것이다. 만약 몇 개의 스패머 계정들이 다수 일반 사용자들을 Link 하고 있다면 Submatrix가 Subgraph가 되기 때문에 이 방법을 통해 스패머 Subgraph를 찾아낼 수도 있다. 이와 유사하게, crowdsourced worker들 간의 similarity로 social network를 만들고, sybil cluster를 탐지하는 방법의 연구[35]도 있다.

2.2. Global Social Network 기반 Fraud 탐지 방법

소셜 네트워크 그래프 기반 Fraud 탐지 기법의 대부분은 Global Social Network 기반 탐지 방법이다. 이 방법은 크게 Propagation model[14,19,20]과 Latent factor model[13,21,22,23,24]로 분류할 수 있다. Propagation model을 사용한 스팸 탐지 방법은 초기 web spam 탐지 방법으로부터 최근 SNS 스팸 탐지 방법까지 발전을 거듭하며 이어지고 있다. Latent factor model은 Matrix factorization을 이용한 스팸 탐지 방법이다. 그 밖에 Outlier detection model 방법도 있다.

2.2.1. Propagation model

이 방법은 전통적인 Pagerank과 HITS 알고리즘으로부터 발전하였다. 이들은 그래프 내에서 중요한 노드를 찾기 위한 방법인데, 이를 역으로 중요하지 않은 노드를 찾는 방법으로 Fraud detection에 적용한 것이다. 중요하지 않은, 혹은 영향력이 없는 스패머 계정 노드의 경우 Pagerank[35]나 HITS[36] 알고리즘에서 상대적으로 낮은 점수를 받게 될 것이다. 하지만 스패머들이 탐지를 피하기 위해 자신의 영향력을 높이는 Link farming, 계정 탈취 등의 방법들을 사용하기 시작하며 약점이 발생하였다.

SybilBelief[19]는 Seed set을 이용한 스팸 탐지 Propagation 알고리즘으로, Label noise의 위험으로부터 보다 자유롭다. 확실한 label을 가진 seed set을 이용하면 신뢰성 있는 노드와 신뢰성 없는 노드에 가중치가 주어지기 때문에 결과적으로 스팸과 비 스팸의 차이가 크게 나타난다. [19]는 특히 소셜 네트워크를 Markov Random Field (MRF)로 모델링하는 방법인데, 구체적으로 각 노드가 sybil 혹은 benign의 binary random variable을 가질 수 있도록 하여 노드의 스팸 가능성을 계산하였고, 이

모델은 seed set에 label noise가 있더라도 높은 스팸 탐지율을 보여 주었다.

[20]은 Twitter에서 발생하는 "Follow spam" 공격을 탐지하기 위한 Collusionrank를 제안하였다. 이는 전통적 Pagerank를 기반으로 한 알고리즘인데, Follow spam의 스패머가 기존 스패머와 달리 in-degree가 상당히 높게 나타난다는 특징을 이용해서 이에 맞게 최적화하였다. 따라서 Collusionrank는 Twitter에 신고된 Follow spam 공격자의 94%를 탐지하였다. 하지만 Collusionrank를 비롯한 대부분의 Propagation method들이 가지고 있는 문제는, 일반 사용자라도 SNS상에서 많은 activity를 보이지 않으면 스패머로 간주되기 쉽다는 점이다.

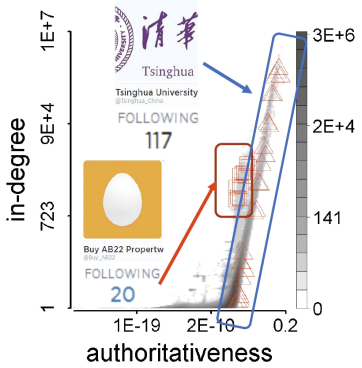

Fig. 5 Sychronized behavior detection of CatchSync[14]

Propagation algorithm만을 사용한 방법도 있지만, 다른 feature들을 함께 고려한 방법도 있다. Catchsync [14]는 HITS 알고리즘을 이용해 얻은 Authority와 Hub value를 각각 in-degree와 out-degree와의 관계로 mapping하였다. Twitter에서 높은 out-degree를 가지면서 낮은 hub value를 갖는 노드들이 집중적으로 분포해 있고, 높은 in-degree를 가지면서 다소 낮은 authority value를 갖는 노드들을 synchronized malicious user로 규정하였다(그림 5). 따라서 일반 사용자의 행동 패턴에 비해 Synchronity가 높게 나타난다면 anomaly로 규정된다. 이 방법은 synchronized behavior를 탐지하는 데 좋은 탐지율을 보여 주었지만, 스패머가 in-degree나 out-degree를 고려하여 공격할 가능성이 있다는 점에서 약점이 있다. 이 외에도, 질문 및 답변 공유 사이트에서의 crowdsourced worker를 탐지하는 [26] 에서는 Social network attribute로 HITS score, closeness/betweenness centrality[15]를 참고하여 사용하고 있다.

2.2.2. Latent factor model

이 방법은 Matrix factorization을 이용한 스팸 탐지 방법으로, Recommender system에서는 사용자가 특정 아이템에 대해 어느 정도의 선호도를 가지고 있을지 등을 예측하는 모델로 쓰였다. 이러한 Latent factor model을 이용한 스팸 탐지 방법들[13,22,23,24] 중 CoBaFi[23]와 CDOutliers[24]는 각각 엣지에 attribute, 노드에 attribute가 있는 그래프를 분석한 스팸 탐지 방법이라는 특징이 있다. 특히, CoBaFi는 탈취 계정의 클러스터를 탐지하는 데 특화된 방법이다. 이렇게 Latent factor model을 사용하면 Propagation method에서 탐지하기 힘든 계정 탈취를 탐지하기 용이하고, 더욱이 집단으로 행동하는 스패머들의 클러스터를 탐지할 수도 있다.

2.3. Outlier detection model

이 방법은 소셜 네트워크 그래프만 사용한 방법은 아니지만, 최근 스팸 탐지에 데이터 마이닝 분야에서 사용되던 Outlier detection 방법들을 적용한 사례이고 자주 등장하고 있기에 활용 가치가 큰 기법이다. Outlier detection 방법을 사용했을 때의 장점은 어떠한 유형의 공격이라도 일반 사용자들과의 평균 패턴과 다르면 Outlier로 간주하여 탐지 가능하다는 것이다. 특히, 스팸 탐지에서의 고질적인 문제인 계정 탈취 공격 문제를 일부 해결 가능하기 때문에 의미가 크다.

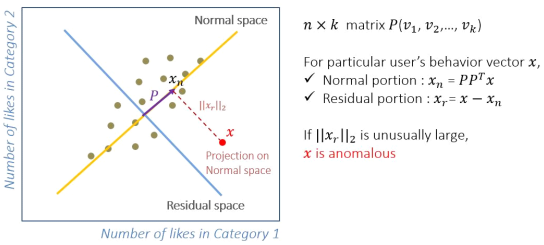

Fig. 6 Like spam detection of facebook with PCA

[25]는 Principal Component Analysis (PCA) 주성분 분석을 사용한 스팸 탐지 방법이다(그림 6). Facebook을 대상으로 한 Like spam 탐지 실험에서, 일반 사용자들이 주로 어떤 카테고리 내의 어떤 페이지들에 Like 표시를 했는가에 대한 정보로 주성분을 계산하였다. 즉, 사용자와 카테고리의 관계를 연결한 정보가 입력이 된다.

만약, 스패머가 일반 패턴과 다르게 Like 표시를 한 경우, 스패머는 주성분에서 멀리 떨어진 곳에 위치하게 되고 Outlier로 분류된다.

[26]은 Yahoo! Answer의 스패머들을 잡기 위해 선형 회귀를 사용하였다. 실제 Abuse report 수와 예상되는 Abuse report 수의 차이를 이용해 일반 사용자에 대해 Deviant score를 계산하고, 일반 사용자와 크게 다른 경우 스패머로 규정하였다. [27]은 anomalous subgraph를 찾기 위해 inner connectivity와 external separability를 동시에 관찰하였다. 이러한 Outlier detection model들의 경우 feature로 어떤 것을 사용하느냐에 따라 결과가 크게 달라지기도 하지만, normal pattern으로부터 얼마나 차이가 나야 Outlier로 규정할지에 대한 Threshold 결정 문제도 항상 따르는 편이다. 최근에는 이러한 사용자의 행동 정보 자체를 임베딩하여 분류하는 모델[28] 등도 제안되고 있다.

Ⅲ. 스팸 공격 유형 및 제안 기법

3.1. 온라인 소셜 네트워크 스팸 유형

스팸은 일반적으로 광고 컨텐츠의 전파가 목적이기 때문에, 전통적인 이메일 스팸과 같이 일대다 형식의 broadcasting type 공격이 대부분이다. 이는 하나의 계정이 다수의 불특정 사용자에게 메시징을 통해 직접 컨텐츠의 전파를 하는 것으로, 적은 비용으로 큰 직접적 효과를 얻을 수 있어 많이 사용되고 있다. 보통 Twitter, Facebook, Instagram 등 메시징 서비스를 제공하는 많은 온라인 소셜 네트워크 서비스에서 자주 등장한다.

반면, 간접 노출을 유도하는 유형이 있다. 가장 대표적인 것이 웹 스팸의 spamdexing 방식과 비슷한 특정 계정의 영향력을 높이는 distributed type이다. 이는 다수의 스패밍 협력 계정들이 link farm을 형성하여 서로의 영향력을 높여준 상태에서 스팸 컨텐츠를 업로드하는 특정 계정을 구독하여 명시적으로 구독자 수를 높여 주는 방식이다. 이렇게 가짜 영향력을 얻게 된 스패밍 계정은 다른 일반 사용자 계정을 구독하여 자신의 존재감을 알리고, 맞 구독을 유도하여 자신의 타임라인에 업로드 되는 스팸 컨텐츠를 볼 수 있도록 한다. 다른 유형으로 hashtag를 이용한 Tag spam이 있는데, 이는 광고하고자 하는 컨텐츠에 유행하고 있는 hashtag를 달아 해당 hashtag를 검색한 사용자들에게 노출시키는 방법이다.

3.2. 계정 탈취 공격(Account Hijacking)

계정 탈취 공격은 스팸 봇을 이용한 공격[37]과 다르게 피싱 링크 등을 이용하여 사용자의 계정을 해킹하고, 이를 스패밍 계정으로 사용하는 방법이다. 예를 들면, 해킹한 계정으로 친구들에게 광고를 보내게 되면 친구가 아닌 사람이 보낸 메시지보다 더 주의깊게 보게 된다는 장점이 있다. 또한, 메시징이 아닌 타임라인에 광고를 올리게 되면 그 계정을 이전부터 구독하고 있던 사용자들에게 손쉽게 광고를 노출시킬 수 있다는 장점도 있어 많이 사용된다. 이는 본래 일반 사용자의 계정이었기 때문에 이상 행동이 발생하는 시점에 유의하여 탐지해야 한다는 어려움이 있다.

3.3. 제안 기법

broadcasting type 공격의 경우, 다수의 불특정 사용자들을 대상으로 메시징을 하기 때문에 메시지 수신자 간에 기존 관계가 크게 없을 가능성이 높다. 이 경우에 graph 특성을 이용한다면 발신자를 기준으로 local clustering coeffieient를 관찰하는 것이 도움이 된다. 이는 수신자들 간에 발생할 수 있는 모든 관계의 수 중에 실제로 얼마나 많은 관계가 발생하였는지 측정하는 방법으로, 일반적인 broadcasting type 스패머의 경우 clustering coefficient가 낮게 나타난다.

간접 노출을 유도하는 유형 중 distributed type은 각 계정을 구독하는 사용자들(즉, 스패밍 협력 계정을 타깃으로 한다.)을 중심으로 행동 패턴을 살펴보아야 한다. 이들이 일반 사용자들처럼 충분한 활동량을 보이지 않고 자기들만의 구독수 증가 관계에만 집중하고 있다면 의심해 볼 수 있다. 계정 생성 시간이나 활동 시간을 검토해 보거나 프로필 정보를 참고해 보는 등의 contents 및 behavior 특성과 hybrid 기법으로 사용하는 것이 탐지율 향상에 도움이 된다. Tag spam과 같은 유형은 컨텐츠와 관계가 없는 hashtag를 첨부한다는 것이 문제가 되는데, 이 경우 hashtag들간의 관계에서 엔트로피를 분석해 보거나 타임라인의 특정 hashtag 등장 비율을 계정 사용자의 사회관계 특성과 함께 이용하는 방식이 도움이 될 수 있다.

계정 탈취 공격의 경우, 시간에 따른 행동 분석이 도움이 되는데 time window별로 관계 그래프의 증감 특성이

나 행동 관련 특성(접속 시간, 작성 글자 수, 이용한 기능 등)등을 계산하여, 기존의 패턴에서 크게 벗어나는 경우 anomaly로 간주하는 방법을 사용할 수 있다. 이전 연구의 경우 PCA와 같은 주성분분석 방법이 도움이 된다.

Ⅳ. 결 론

Facebook, Twitter와 같은 온라인 소셜 네트워크 서비스 (Social Network Service, 이하 SNS)는 현재 세계적으로 가장 주목 받고 있는 통신 매체이다. 그 사용량 급증에 따라 이를 악용하고자 하는 온라인 공격자들의 수 역시 날로 증가하고 있다. 본 논문은 이러한 온라인 공격자를 탐지하기 위해 이용되는 소셜 네트워크 그래프 기반 방법들의 연구 동향을 파악하고자 하였다. 기존에 많이 사용되는 Contents-based approach와 같은 방법들은 텍스트 정보에 의존하는 경향이 크기 때문에 스팸성이 높게 나타나지 않는 스팸은 탐지하기 어렵다는 문제가 있다. 소셜 네트워크 그래프 속성 기반 스팸 탐지 방법은 크게 소셜 네트워크 정보로 구성한 Local graph 및 Global graph에 기반한 방법들로 나누어진다. 또한, 소셜 네트워크 그래프 특성을 일부 이용하는 outlier detection model도 등장하였다. 본 논문에서는 계정 탈취 공격을 탐지하기 위한 소셜 네트워크 그래프 속성 기반 방법을 함께 제안하여, 그래프 속성이 향후 온라인 소셜 미디어 보안 문제에서 중요한 역할을 할 수 있음을 시사한다.

ACKNOWLEDGEMENT

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2017R1D1A1B03035557).

References

[1] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna. "COMPA: Detecting Compromised Accounts on Social Networks." *NDSS*. 2013.

[2] G. Magno, T. Rodrigues, V. Augusto, and F. Almeida. “Detecting spammers on twitter. In Collaboration, Electronic messaging”, *Anti-Abuse and Spam Conference (CEAS)*, 2010.

[3] J. M. Romo and L. Araujo. “Detecting malicious tweets in trending topics using a statistical analysis of language”. *Expert Systems with Applications* 40.8, 2013

[4] S. Y. Schoenebeck, D. M. Romero, G. Schoenebeck, and D. Boyd. “Detecting spam in a twitter network.” *First Monday*, 15(1), January, 2009.

[5] X. Li, M. Zhang, Y. Liu, S. Ma, Y. Jin, and L. Ru. "Search engine click spam detection based on bipartite graph propagation." *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 2014.

[6] T. Tian, J. Zhu, F. Xia, X. Zhuang, and T. Zhang. "Crowd fraud detection in internet advertising." *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015.

[7] Q. Cao, X. Yang, J. Yu, and C. Palow. "Uncovering large groups of active malicious accounts in online social networks." *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2014.

[8] Q. Cao, X. Yang, J. Yu, and C. Palow. "VolTime: Unsupervised Anomaly Detection on Users' Online Activity Volume." *Proceedings of the 2017 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics*, 2017.

[9] L. H. Yu, and D.Y. Yeung. *“A learning approach to spam detection based on social networks”. Diss. Hong Kong University of Science and Technology*, 2007

[10] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F.Bonchi. "The social world of content abusers in community question answering." *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015.

[11] E. Zhai, Z. Li, Z. Li, F. Wu, and G. Chen. "Resisting tag spam by leveraging implicit user behaviors." *Proceedings of the VLDB Endowment* 10.3 (2016): 241-252.

[12] H. Zheng, M. Xue, H. Lu, S. Hao, H. Zhu, X. Liang, and K. W. Ross "Smoke screener or straight shooter: Detecting elite sybil attacks in user-review social networks." *arXiv preprint arXiv*:1709.06916 (2017).

[13] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang "Inferring strange behavior from connectivity pattern in social networks." *Advances in Knowledge Discovery and*

Data Mining. Springer International Publishing, 2014. 126-138.

[14] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang. "CatchSync: catching synchronized behavior in large directed graphs." *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014.

[15] M. Newman,. *Networks. Oxford university press, 2018*.

[16] L. Akoglu, M. McGlohon, and C. Faloutsos. "Oddball: Spotting anomalies in weighted graphs." *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2010. 410-421.

[17] D. O'Callaghan, M. Harrigan, J. Carthy, and P. Cunningham." *Network Analysis of Recurring YouTube Spam Campaigns." ICWSM. 2012*.

[18] S. Jeong, G. Noh, H. Oh, and C. Kim. "Follow spam detection based on cascaded social information." *Information Sciences* 369 (2016): 481-499.

[19] N. Z. Gong, M. Frank, and P. Mittal. "SybilBelief: A Semi-Supervised Learning Approach for Structure-Based Sybil Detection." *Information Forensics and Security, IEEE Transactions on* 9.6 (2014): 976-987.

[20] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi. "Understanding and combating link farming in the twitter social network." *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012.

[21] D. Yuan, G. Li, Q. Li, and Y. Zheng. "Sybil defense in crowdsourcing platforms." *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017.

[22] N. Shah, A. Beutel, B. Gallagher, and C. Faloutsos. "Spotting suspicious link behavior with fBox: an adversarial perspective." *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE, 2014.

[23] A. Beutel, K. Murray, C. Faloutsos, and A. J. Smola. "Cobafi: collaborative bayesian filtering." *Proceedings of the 23rd international conference on World wide web*. ACM, 2014.

[24] M. Gupta, J. Gao, and J. Han. "Community distribution outlier detection in heterogeneous information networks." *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 2013. 557-573.

[25] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. "Towards detecting anomalous user behavior in online social networks." *Proceedings of the 23rd USENIX Security Symposium (USENIX Security)*. 2014.

[26] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. Bonchi. "The Social World of Content Abusers in Community Question Answering." *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015.

[27] B. Perozzi, and L. Akoglu. "Scalable anomaly ranking of attributed neighborhoods." *Proceedings of the 2016 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics*, 2016.

[28] S. Dhawan, S. C. R. Gangireddy, S. Kumar, and T. Chakraborty. 2019. “Spotting collective behaviour of online frauds in customer reviews.” *In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*, Sarit Kraus (Ed.). AAAI Press 245-251.

[29] S. Jeong, J. Lee, J. Park, and C. Kim. "The Social Relation Key: A new paradigm for security." *Information Systems* 71 (2017): 68-77.

[30] M. McPherson, L. S. Lovin, and J. M. Cook. "Birds of a feather: Homophily in social networks." *Annual review of sociology* 27.1 (2001): 415-444.

[31] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. S. Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. "Superfamilies of evolved and designed networks." *Science* 303.5663 (2004): 1538-1542.

[32] Ö. N. Yaveroğlu, N. M. Dognin, D. Davis, Z. Levnajic, V. Janjic, R. Karapandza, A. Stojmirovic, and N. Pržulj. "Revealing the hidden language of complex networks." *Scientific reports* 4 (2014).

[33] Y. Koren, "Collaborative filtering with temporal dynamics." *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009.

[34] R. Salakhutdinov, and A. Mnih. "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo." *Proceedings of the 25th international conference on Machine learning*. ACM, 2008.

[35] L.Page, S. Brin, R. Motwani, and T. Winograd. “The PageRank citation ranking: Bringing order to the web.” *Stanford InfoLab*, 1999.

[36] J. Kleinberg, "Hubs, authorities, and communities." *ACM computing surveys (CSUR)* 31.4es (1999): 5.

[37] Lee Chan-chan, Seo Go-eun, Shin-yong Shin, Dong-gun Kim, & Jae-hee Cho. (2015). “Improved tweet bot detection using geographic space and device information.” Journal of the Korea Information and Communication Society, 19(12), 2878-2884.

정시현(Sihyun Jeong)

서울대학교 컴퓨터공학부 박사과정

※ 관심분야 : 소셜정보망 및 데이터 분석

오하영(Hayoung Oh)

이화여자대학교 컴퓨터 공학 석사
서울대학교 컴퓨터 공학 박사
성균관대학교 글로벌융합학부 조교수

※ 관심분야 : 소셜정보망 및 데이터 분석