온라인 사회 연결망을 위한 개인정보 보호 방안들의 평가

이종민, 배덕호, 김상욱
한양대학교 전자컴퓨터통신공학과
e-mail: {ggoy123, dhbae, wook}@agape.hanyang.ac.kr

Evaluation of Privacy Preserving Methods in Online Social Networks

Jong-Min Lee, Duck-Ho Bae, Sang-Wook Kim
Department of Electronics and Computer Engineering, Hanyang University

요 약

본 논문은 온라인 사회 연결망을 위한 개인정보 보호 방안들에 대해 알아보고, 각 방안이 사회 연결망의 특성 변화에 미친 영향을 분석한다. 분석 결과, 개인정보 보호 방안들은 사회 연결망의 특성을 크게 훼손시키는 것으로 나타났다.

1. 서론

최근 인터넷의 발달과 함께 온라인상에 다양한 사회 연결망들이 출현하고 있다. 이러한 사회 연결망들을 노드(개인)와 에지 (관계)로 이루어진 그래프 형태로 모델링한 후, 해당 그래프를 분석하여 유용한 정보를 도출하고자 하는 시도들이 많아지고 있다. 그러나 그래프를 분석하는 과정에서 공격자들에 의해 연결망 구성원들의 개인정보가 노출되는 문제가 발생하였으며, 이를 해결하기 위한 다양한 그래프 개인정보 보호 방안들이 제안되었다.

기존의 그래프 개인정보 보호 방안들은 각각 자신만의 공격 모델을 설립하고 이로부터 개인 정보를 보호하기 위해 새로운 에지를 삽입, 기존 에지를 삭제, 에지 간의 교환 등을 수행하여 그래프의 구조를 변형한다. 이를 통해 공격자가 특정 노드 혹은 노드와 노드 간의 관계를 알아내지 못하게 함으로써 구성원들의 개인정보를 보호한다.

지금까지 개인정보 보호에 관한 연구는 공격 모델을 설립하고, 이를 보호하기 위한 방안에만 초점이 맞추어져 있었다. 그러나 연결망 분석의 근본적인 취지를 생각해볼 때, 개인정보 보호 방안이 원본 그래프 특성 변화에 미치는 영향을 분석하는 것 또한 매우 중요하다. 본 논문에서는 원본 그래프와 개인정보 보호 방안을 적용한 그래프와의 비교를 통해 기존의 개인정보 보호 방안들이 그래프의 특성에 미치는 영향을 분석한다.

2. 그래프 개인정보 보호 방안

기존의 테이블 데이터를 위한 개인정보 보호 방안은 단순히 노드의 ID를 임의의 ID로 익명화하여 개인정보를 보호하였다[1]. 그러나 그래프는 테이블 데이터와는 달리 노드의 degree, 노드와 노드 사이의 에지, 서브 그래프 등의 구조적인 특징 등을 갖는다. 이로 인해 기존의 테이블 데이터를 위한 개인정보 보호 방안은 그래프의 구조적 특징을 이용한 질의 공격으로부터 개인정보를 보호할 수 없다. 이를 위해 다음과 같은 그래프 개인정보 보호 방안들이 제안되었다.

2.1 k-Anonymity Privacy Preservation via Edge Modification

(1) k-degree anonymization: 노드의 degree를 이용한 질의 공격으로부터 개인정보를 보호하기 위해 그래프 내에 동일한 degree를 갖는 노드가 최소한 k개 이상 존재하도록 에지를 삽입하는 방안이다[2].

(2) k-automorphism anonymity: 서브 그래프를 이용한 질의 공격으로부터 개인정보를 보호하기 위해 그래프 내에 동일한 구조를 갖는 서브 그래프가 최소한 k개 이상 존재하도록 에지를 삽입하는 방안이다[3].

2.2 Privacy Preservation via Randomization

(1) Random deletion: 노드의 degree를 이용하여 노드들 간의 에지 존재 여부를 확인하는 질의 공격에 대해 노드의 degree를 기준으로 그룹핑 한 뒤, 그룹 간의 에지 존재 확률을 낮추도록 랜덤하게 에지를 삭제하는 방안이다[4].

(2) Random switching: 서브 그래프를 이용한 질의 공격에 대해 공격자가 특정 서브 그래프를 찾지 못하거나 공격 결과를 신뢰하지 못하도록 랜덤하게 에지를 교환하는 방안이다[5].

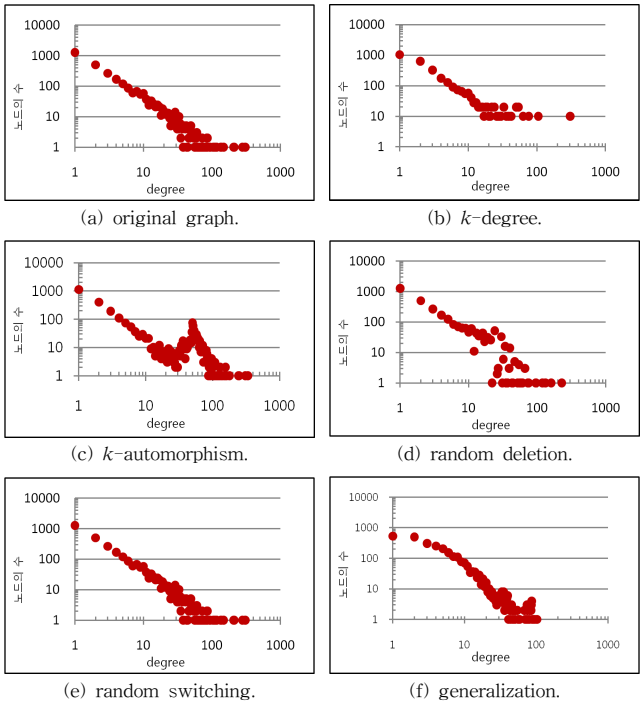
2.3 Cluster-based Graph Generalization

그래프를 클러스터링 하여 각 클러스터에 속한 노드와 에지의 수에 관한 정보만을 간략히 제공하는 방안이다. 공격자는 요약된 정보만을 볼 수 있기 때문에 구조적 질의를 통한 공격이 의미가 없게 된다. 그러나 해당 그래프를 분석하기 위해서는 요약된 그래프 정보를 통해 임의의 그래프를 재구성해야만 한다[6].

3. 실험 결과

본 장에서는 개인정보 보호 방안이 원본 그래프의 특성에 미치는 영향을 분석하기 위해, 원본 그래프와 개인정보 보호 방안을 적용한 그래프에서 각각 degree 분포와 hop-plot을 측정하고 이를 비교하였다. 분석을 위해 샘플링 된 Epinion 데이터[1]를 사용하였다. 데이터는 3,094개의 노드와 9,680개의 에지로 구성된다.

1) www.epinions.com



(a) original graph. (b) k-degree.

(c) k-automorphism. (d) random deletion.

(e) random switching. (f) generalization.

그림 1. Degree 분포.

3.1 Degree 분포

일반적인 온라인 사회연결망의 degree 분포는 power-law 분포를 따른다[7]. 그림 1은 원본 그래프와 각 개인정보 보호 방안을 적용한 그래프들의 degree 분포를 나타낸다. 그래프의 x축은 노드의 degree, y축은 해당 degree를 갖는 노드의 수를 나타낸다. 원본 그래프 (그림 1(a))의 degree 분포는 power-law 분포를 따름을 알 수 있다. k-degree anonymization 방안을 적용한 그래프 (그림 1(b))는 degree가 높은 노드들의 degree 분포가 power-law를 따르지 않음을 알 수 있다. 일반적으로 사회연결망에서 degree가 높은 노드들은 그 수가 매우 적으므로 k-anonymity를 만족하지 않는다. 이를 해결하기 위해 k-degree anonymization 방안은 degree가 높은 노드들에 많은 에지를 추가하며, 이로 인해 degree 분포가 power-law 분포를 벗어나게 된다. k-automorphism anonymity 방안을 적용한 그래프 (그림 1(c)) 또한 power-law 분포를 따르지 않는다. degree가 높은 노드들을 포함하는 서브 그래프는 k-anonymity를 만족하지 않는다. 이를 해결하기 위해 k-automorphism anonymity 방안은 degree가 비교적 낮은 노드들 (10 ~ 100)을 포함하는 서브 그래프에 에지를 추가하는 전략을 사용함으로써, degree 분포가 power-law 분포를 벗어나게 된다. Random deletion 방안을 적용한 그래프 (그림 1(d)) 역시도 degree 분포에 변화가 일어났다. 이는 10 ~ 100정도의 degree를 갖는 노드들 간의 에지 존재 확률이 상대적으로 높아 해당 그룹 간에 에지 삭제가 많이 발생하였기 때문이다. Random switching 방안 (그림 1(e))은 에지를 삽입, 삭제한 것이 아니라 단지 교환하기 때문에 degree의 분포는 바뀌지 않았다. Generalization 방안 (그림 1(f))은 그래프의 요약된 정보를 이용해 임의의 그래프를 재구성하기 때문에 사회 연결망에서 나타나는 degree가 높은 허브 노드가 생성되지 않는 것을 볼 수 있다.

3.2 Hop-plot

Hop은 그래프에서 도달 가능한 두 노드 사이에 존재하는 에지의 수를 의미한다. 그림 2는 각 hop에 대해 해당 hop에서 도달 가능한 모든 노드 쌍의 수를 나타낸다.

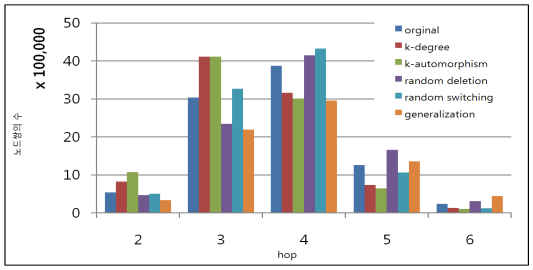


그림 2. Hop-plot.

원본 그래프는 2 hop에서 4 hop으로 갈수록 도달 가능한 노드 쌍의 수가 점점 증가 하고, 4 hop에서 가장 높은 값을 보인 뒤 5 hop부터는 급격히 감소함을 볼 수 있다. k-degree anonymizatin과 k-automorphism anonymity 방안은 3 hop에서 가장 높은 값을 가지고 4 hop부터는 값이 감소하는 경향을 보인다. 이는 두 방안 모두 에지를 추가함으로써 노드 간의 평균 거리가 가까워졌기 때문이다. Random deletion 방안은 에지의 삭제로 인해 도달 가능한 모든 노드의 쌍의 수가 전반적으로 감소함을 볼 수 있다. Random switching 방안의 경우 degree 분포는 변함이 없었지만, 노드들 간의 구조가 변경됨으로 인해 원본 그래프와 비교하여 hop-plot이 달라짐을 볼 수 있다. 앞에서 언급하였듯이, graph generalization 방안은 그래프 재구성 시 허브 노드가 생성되지 않는다. 이로 인해 노드 사이의 평균 거리가 멀어져 hop-plot의 분포가 완만해짐을 볼 수 있다.

4. 결론

본 논문에서는 그래프 개인정보 보호 방안이 그래프 특성 변화에 미친 영향을 분석하였다. 이를 위해 원본 그래프와 개인정보 보호 방안을 적용한 그래프의 degree 분포와 hop-plot을 차이를 비교하고, 차이가 나타나는 원인을 규명하였다. 비록 각각의 그래프 개인정보 보호 방안이 보호하고자 하는 질의 공격이 다르기 때문에 각 방안 간의 직접적인 비교가 불가능하다. 그러나 본 연구에서 제시하는 그래프 특성 보존의 측면도 고려하여야 할 것이라고 생각한다.

감사의 글

"본 연구는 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단 (No.2008-0061006) 및 지식경제부 및 정보통신산업진흥원의 'IT융합 고급인력과정 지원사업' (NIPA-2011-C6150-1101-0001)의 지원을 받았습니다."

참고문헌

[1] M. Hay et al. Anonymizing social networks. Technical Report 07-19, University of Massachusetts Amherst, 2007.

[2] K. Liu and E. Terzi, "Towards Identity Anonymization on Graphs," In *ACM SIGMOD*, 2008.

[2] L. Zou, L. Chen, and M. Ozsu, "k-Automorphism: A General Framework for Privacy Preserving Network Publication," In *VLDB*, 2009.

[3] L. Zhang and W. Zhang, "Edge Anonymity in Social Graphs," In *SocialCom*, 2009.

[4] X. Ying and X. Wu, "Randomizing Social Networks: A Spectrum Preserving Approach," In *SDM*, 2008.

[5] M. Hay et al., "Resisting Structural Re-identification in Anonymized Social Networks," In *VLDB*, 2008.

[6] M. Floutsos et al. "On Power-Law Relationships of the Internet Topology," Computer Communicatins Review 29, pp. 251 - 262, 1999.