제36회 한국정보처리학회 추계학술발표대회 논문집 제18권 제2호 (2011. 11)

프라이버시 보호를 위한 소셜 네트워크의 익명화 비용에 관한 연구*

박치성, 강주성, 이옥연
국민대학교 수학과, 정보보안연구소
{parkcs01,jskang,oyyi}@kookmin.ac.kr

A study on anonymization cost of social network for privacy preservation

Chi-Seong Park, Ok-Yeon Yi, Ju-Sung Kang
Dept. of Math. and CISI, Kookmin University

요 약

소셜 네트워크를 통해 수집된 수많은 데이터들은 여러 분야에 중요한 자료로 활용되고 있으며, 소셜 네트워크상의 데이터들이 이용되면서 개인정보가 노출되는 프라이버시 문제가 발생하고 있다. 프라이버시 문제를 해결하기 위한 실용적인 방안으로 k-익명성, l-다양성 등의 개념과 이를 토대로 한 데이터 익명화 방법이 제안되어 있다. 데이터의 익명화에서는 원본데이터의 왜곡을 최소화하면서 프라이버시 보호를 극대화하는 것이 목적이다. 이러한 목적을 달성하기 위해 익명화 비용을 측정하기 위한 합리적인 방법이 필요하다. 본 논문에서는 소셜 네트워크 그래프의 익명화 알고리즘 수행을 위해 필수적 요소인 익명화 비용을 합리적이고 실용적으로 측정하는 방법을 제안한다.

1. 서론

최근 빠르게 확산된 소셜 네트워크는 개인의 정보를 공유함으로써 기존의 오프라인에서 형성된 인맥을 강화시키고 온라인에서 새로운 인맥을 형성할 수 있다는 큰 장점을 갖고 있다. 이렇게 형성된 소셜 네트워크 데이터는 기업의 마케팅 활동 및 다양한 연구 목적으로 유용하게 이용되고 있다. 반면에 소셜 네트워크 데이터의 개방적인 특성으로 인해 개인의 중요 정보 노출을 방지하기 어렵다는 프라이버시 문제가 발생하게 된다.

소셜 네트워크는 자신을 중심 개체로 하여 다른 개체들과 하나의 네트워크 또는 인맥을 형성하는 구조적인 특징을 갖고 있다. 소셜 네트워크는 하나의 네트워크를 구성하는 개체들을 점(vertex)으로 나타내고 이들 사이의 관계를 선(edge)으로 표현하면 하나의 그래프(graph) 형태로 도식화할 수 있다. 각 개체의 통계적 데이터 기록(record)은 해당하는 점에 라벨(label) 형태로 추가하면 소셜 네트워크 데이터를 온전히 표현할 수 있게 된다. 이렇게 표현된 소셜 네트워크는 구조적인 특징에 의해 이웃 공격(neighborhood attack)[1]에 취약점을 드러내고 있다. 최근 Zhou와 Pei[1]는 통계적 데이터에서의 프라이버시 보호를 위해 사용된 데이터 익명화 방법인 k-익명성(k-anonymity)[2], l-다양성(l-diversity)[3] 개념을 소셜 네트워크 데이터에 적용할 수 있는 알고리즘을 제안하였다.

소셜 네트워크 데이터를 익명화 하게 되면 원본데이터의 왜곡은 피할 수 없다. 원본데이터의 왜곡이 많으면 정보의 정확성(accuracy)이 떨어지게 되므로 소셜 네트워크 데이터를 익명화하는 것에 있어서 원본데이터의 왜곡을 최소화하고 익명성을 극대화하는 것이 가장 큰 목적이 된다. 이를 위해서는 데이터가 합리적이고 효율적으로 익명화되었는지를 측정하기 위한 방법이 요구된다. Zhou와 Pei[1]는 소셜 네트워크에서 두 이웃 구조의 익명화 효율을 측정하는 방법으로 비용($Cost$)이란 측도(measure)를 제안하였다. Zhou와 Pei[1]가 제안한 $Cost$는 라벨을 일반화하고 점과 선을 추가하는 것에 각각 가중치를 부여함으로써 그 값이 계산된다. Zhou와 Pei[1]는 각 가중치의 값에 따라 라벨의 일반화, 점과 선의 추가 사이에 취사선택(tradeoff) 가능함을 확인하였으나 가중치의 값에 대한 결정 방법은 정확히 언급하지 않았다.

본 논문에서는 Zhou와 Pei[1]의 연구에서 제안된 익명화 효율 측도인 $Cost$에 주목하여 그 값을 구하기 위해 필수적인 가중치들의 합리적인 값을 결정하는 방법을 제안한다. 또한 제안한 가중치를 $Cost$에 적용하여 그래프로 표현된 소셜 네트워크의 익명화를 수행한 예를 제시한다.

2. 소셜 네트워크

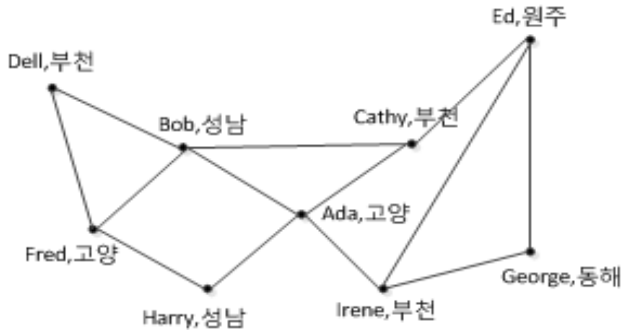
한 개체를 중심으로 하나의 네트워크를 형성하는 소셜 네트워크에서 Zhou와 Pei[1]는 소셜 네트워크 그래프를

* 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (20100024870)

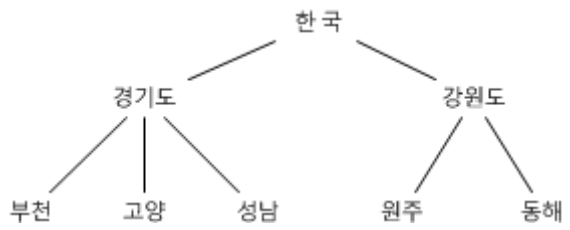
$G=(V,E,L,\mathcal{L})$로 정의한다. 여기에서 V는 점들의 집합, $E \subseteq V \times V$는 선들의 집합, L은 라벨들의 집합, $\mathcal{L} : V \rightarrow L$은 각 점에 라벨을 부여하는 라벨링함수를 의미한다. 그래프 G에서는 점들의 집합은 $V(G)$, 선들의 집합은 $E(G)$, 라벨들의 집합은 L_G, 라벨링함수은 $\mathcal{L}_G$의 표기를 사용하였다. 이와 같은 점, 선, 라벨등의 구성요소를 이용하여 소셜 네트워크 그래프를 (그림 1)과 같은 그래프의 형태로 도식화 할 수 있다.

그래프로 표현된 소셜 네트워크는 구조적인 특징으로 인해 이웃 공격에 취약점을 드러낸다. 이웃 공격은 공격자가 공격대상의 이웃 구조에 대한 정보를 갖고 있다면 그래프로부터 어떤 점이 공격대상에 해당하며 그 주위의 이웃들은 어떤 구조를 갖고 있는 지 등의 프라이버시를 노출시킬 수 있는 공격이다.

본 논문에서는 Zhou와 Pei[1]가 제안한 소셜 네트워크 그래프에 임의의 라벨을 적용하여 (그림 1)과 같은 그래프를 생성하였다. (그림 1)의 그래프에서 모든 점들의 이름을 제거하고 라벨만 유지한 소셜 네트워크 그래프를 익명화하고자 한다.



(그림 1) 소셜 네트워크 그래프



(그림 2) 라벨 계층도

3. B.Zhou와 J.Pei의 소셜 네트워크 익명화

3.1 소셜 네트워크 데이터의 k-익명성

통계적 데이터에서의 k-익명성은 각 속성의 속성 값들을 일반화시켜 공격자가 속성 값에 대한 정보를 이용해 개인을 구분할 수 없도록 하여 프라이버시를 보호한다. 통계적 데이터는 속성 값들의 일반화만으로 익명성을 만족시킬 수 있다. 반면 소셜 네트워크 데이터는 속성 값뿐만 아니라 네트워크 구조에 대한 익명성을 고려해야 한다. 네트워크 구조에 대한 익명성을 만족시키기 위해 소셜 네트워크 그래프의 점들을 구별할 수 있는 이름을 제거하고 점과 선의 추가/삭제, 라벨의 일반화를 이용해 구조적으로 k-익명성을 만족하도록 해야 한다. Zhou와 Pei[1]는 소셜 네트워크 그래프의 익명화를 위해 점과 선의 추가를 이용하지만 소셜 네트워크 그래프에 가짜 점을 추가하지는 않는다. 또한 (그림 2)와 같은 라벨 계층도에 따라 라벨을 일반화 시켜줌으로써 라벨을 익명화한다.

Zhou와 Pei[1]은 소셜 네트워크에서 k-익명성을 만족시키기 위한 알고리즘을 제안하였다. 이 알고리즘은 소셜 네트워크 그래프 내에서 가장 유사한 이웃 구조를 갖고 있는 최소 k개의 점들을 그룹화하고 그룹 내에서 각 점들의 이웃 구조를 동일하게 만들어줌으로써 익명화를 수행한다. 이웃 구조의 유사성을 판단하기 위한 방법이 반복적으로 이루어지며 그 방법으로 $Cost$가 사용된다. $Cost$는 두 점이 갖고 있는 이웃 구조를 동일하게 만들어 익명화하는데 드는 소요를 계산한다. $Cost$의 값이 작을수록 두 점의 이웃 구조가 유사하다고 판단한다. 소셜 네트워크 그래프의 익명화는 유사한 점들을 익명화하는 방법으로 이루어지므로 어떠한 점들이 유사한지를 판단할 수 있는 기준값으로써 $Cost$는 매우 중요하다. $Cost$의 값이 합리적으로 계산되어야 익명화시킬 점들을 올바르게 구분하여 합리적이고 효율적으로 소셜 네트워크 그래프를 익명화할 수 있다.

3.2 익명화 비용값 계산에서의 가중치

Zhou와 Pei[1]는 익명화 알고리즘을 수행하는 과정에서 두 점이 갖고 있는 이웃 구조의 유사성을 판단하기 위해 $Cost$를 제안하였다. $Cost$는 소셜 네트워크 그래프를 익명화하기 위해 변화를 주는 방법, 즉 라벨의 일반화, 선의 추가, 점의 추가에 따라 각각 α, β, γ의 가중치를 주어 계산된다. 이 가중치에 어떠한 값을 주느냐에 따라 어떠한 점들이 유사한 구조를 갖는지가 정해지므로 가중치의 역할이 중요하다. Zhou와 Pei[1]는 가중치에 어떠한 값을 주느냐에 따라 소셜 네트워크 그래프를 익명화하기 위해 변화를 주는 방법들의 양이 취사선택될 수 있다는 것을 실험적으로 확인하였다. 하지만 가중치의 값에 대한 결정 방법을 정확히 언급하지 않았다. 본 논문에서는 가중치의 값이 임의로 결정되면 소셜 네트워크 그래프에 대한 익명화가 수행되지 못하는 경우가 있음을 확인함으로써 가중치의 값에 대한 결정 방법이 정해져야 함을 확인하였다.

4. 익명화 비용값 설정 방법

Zhou와 Pei[1]가 제안한 익명화 알고리즘을 수행하는데 있어서 $Cost$는 두 점의 이웃 구조에 대한 유사성을 판단할 수 있는 기준값으로써 매우 중요한 요소이다. 본 장에서는 $Cost$의 계산 방법을 설명하고 가중치의 값에 대한 결정 방법을 제안한다.

$Cost$를 계산하는 방법은 다음과 같다.

$$Cost(u,v) = \alpha \cdot \sum_{v' \in H'} NCP(v') + \beta \cdot \left| \{(v_1,v_2) | (v_1,v_2) \notin E(H), (A(v_1),A(v_2)) \in E(H')\} \right| + \gamma \cdot (|V(H')| - |V(H)|)$$

α, β, γ는 사용자에 의해 지정된 가중치를 의미한다. 첫 번째 항은 $NCP(v')$들의 합으로 일반화한 라벨의 정보 손실을 측정하고, 두 번째와 세 번째 항은 각각 익명화 이후에 u, v에 추가된 선과 점들의 개수를 의미한다. 또한 점 $A(v_1)$과 $A(v_2)$는 각각 점 v_1과 v_2의 익명화 이후의 점을 의미한다. 여기에서 NCP(Normalized Certainty Penalty)는 라벨 계층도에서 최하위 노드 전체 개수에 대해 일반화한 이후 라벨의 최하위 노드 개수의 비를 의미한다. 예를 들어 일반화된 라벨이 (그림 2)에서 경기도라면 NCP(경기도)의 값은

$$NCP(\text{경기도}) = \frac{(\text{경기도의 최하위 노드 개수})}{(\text{최하위 노드 전체 개수})} = \frac{3}{5}$$

이 된다.

본 논문에서는 가중치 α, β, γ의 값을 결정하는데 있어서 라벨 계층도와 소셜 네트워크 그래프를 이용하여 α, β, γ 값의 비를 결정해 그 값을 결정하고자 한다. α는 라벨에 관련된 가중치로 각 최하위 노드에 대한 점들의 평균 개수를 그 값으로 한다. 라벨을 익명하는 것은 라벨 계층도에서 노드를 일반화 시키는 것으로 최하위 노드마다 갖게 되는 점들의 평균 개수를 가중치의 값으로 하면 기댓값으로써 의미가 있다. 예를 들어 (그림 1)과 (그림 2)를 함께 생각해서 보면 부천 3개, 고양 2개, 성남 2개, 원주 1개, 동해 1개의 점이 있다. 이것을 이용해 α를 계산하면

$$\alpha = \frac{(3+2+2+1+1)}{5} = \frac{9}{5}$$

의 값을 갖게 된다. β는 추가되는 선의 개수에 관련한 가중치로 소셜 네트워크 그래프에 포함되어 있는 선의 총 개수를 그 값으로 한다. 또한 γ는 추가되는 점의 개수에 관련한 가중치로 소셜 네트워크 그래프에 포함되어 있는 점의 총 개수를 그 값으로 한다.

결과적으로 α, β, γ 값의 비는

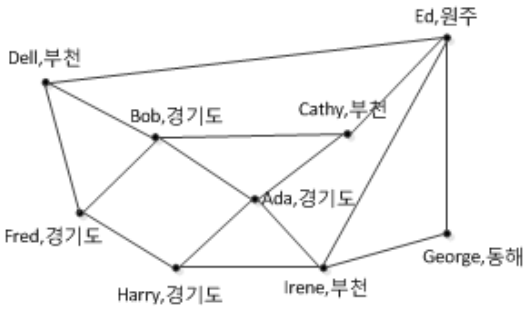
$$\alpha : \beta : \gamma = \left(\frac{\text{점의 총 개수}}{\text{최하위노드의 총 개수}} \right) : (\text{선의 총 개수}) : (\text{점의 총 개수})$$

로 계산된다. (그림 1)의 그래프에서 α, β, γ의 값의 비를 정하면 $\alpha : \beta : \gamma = (9/5) : 13 : 9$가 된다.

5. 합리적인 $Cost$를 적용한 그래프 익명화

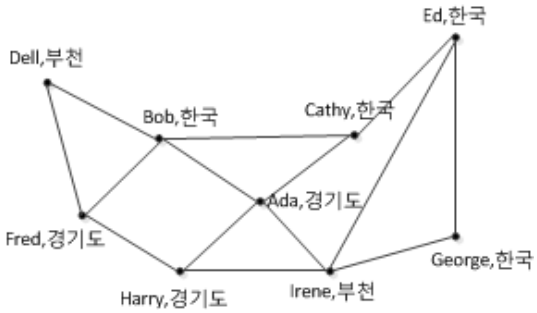
$Cost$에서 가중치의 값에 대한 결정 방법을 제안하였다. 본 장에서는 4장에서 제안한 가중치의 값에 대한 결정 방법을 이용한 $Cost$를 적용하여 (그림 1)에 제시된 소셜 네트워크 그래프에 대한 익명화를 수행한다.

비교를 위해 α, β, γ 값의 비를 1:1:1로 설정하고 2-익명성을 만족하도록 소셜 네트워크 그래프의 익명화를 수행하였다. 익명화를 수행하는 과정에서 (그림 3)과 같은 그래프를 얻을 수 있다. Zhou와 Pei[1]가 제시한 익명화 알고리즘에 대한 기본 가정에서는 더 이상의 익명화를 진행 할 수 없다.

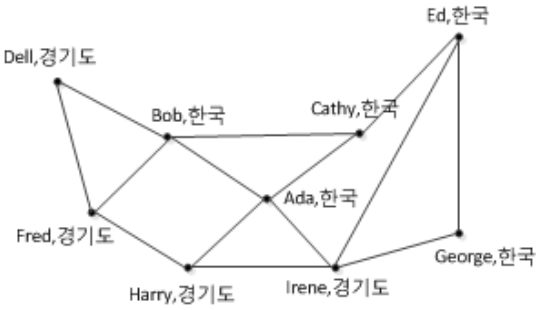


(그림 3) 가중치 $\alpha : \beta : \gamma = 1:1:1$, $k=2$일 때 익명화 알고리즘 3번째 수행의 익명화 그래프

본 논문에서 제안된 가중치 계산법을 이용해 α, β, γ 값의 비는 $(9/5):13:9 = 9:65:45$의 비로 설정할 수 있다. 앞의 예제와 마찬가지로 2-익명성를 만족하도록 익명화를 수행하였다. 앞의 예제에서와 같은 단계에서 얻은 그래프는 (그림 4)와 같다. 이 후, Zhou와 Pei[1]의 기본 가정에서 익명화 알고리즘을 지속적으로 수행한 결과로 (그림 5)와 같은 소셜 네트워크 익명화 그래프를 얻을 수 있다.



(그림 4) 가중치 $\alpha : \beta : \gamma = 9:65:45$, $k=2$일 때 익명화 알고리즘 3번째 수행의 익명화 그래프



(그림 5) 가중치 $\alpha : \beta : \gamma = 9:65:45$, $k=2$일 때 최종 익명화 그래프

원본 소셜 네트워크 그래프와 각각 다른 가중치의 값을 적용한 $Cost$를 이용해 익명화 알고리즘을 수행한 이후의 그래프인 (그림 3)과 (그림 5)의 그래프를 각각 비교해 보자. 그림에서 보듯이 (그림 3)에 비해 (그림 5)의 그래프

는 선의 추가가 덜 되어 그래프의 변형이 적고, 2-익명성을 만족하는 것을 확인할 수 있다. (그림 3)은 익명화 알고리즘을 끝까지 수행하지 못하여 2-익명성을 만족하지 못한다. 본 논문에서는 가중치의 값을 결정하는데 있어서 합리적인 방안을 제시함으로써 더 적은 그래프의 변형으로 소셜 네트워크 그래프에 대한 익명화를 수행할 수 있음을 확인하였다.

6. 결론

본 논문에서는 소셜 네트워크 그래프의 합리적이고 효율적인 익명화를 달성하기 위해 *Cost*에서 가중치의 값을 결정하는데 있어서 합리적인 방법을 제안하였다. 이 방법을 적용한 *Cost*를 이용해 더 적은 그래프의 변형으로 k-익명성을 달성하여 그 결과를 확인하였다.

통계적 데이터에는 익명화 비용을 측정하기 위해 *precision*, *score*등의 방법들이 연구되어 있다. 이러한 방법들을 *Cost*와 마찬가지로 소셜 네트워크의 익명화 비용을 측정하는데 사용될 수 있을 것으로 판단된다. *Cost*이외의 익명화 측정 방법에 대한 연구를 통해 좀 더 합리적으로 소셜 네트워크를 익명화하는 것에 대해 연구가 이루어져야 한다.

또한 본 논문에서는 소셜 네트워크 그래프에서 각 점들이 단일 라벨일 때를 가정하여 익명화를 실시하였다. 실제 소셜 네트워크에서는 여러 개의 라벨을 가질 수 있다. 이와 같이 여러 개의 라벨을 갖는 경우에 소셜 네트워크 그래프를 익명화하기 위하여 다중 라벨에 대한 익명화 비용 측정 방법에 대해 추가적인 연구가 이루어져야 한다.

참고문헌

[1] Bin Zhou, Jian Pei "The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks", Springer-Verlag London Limited 2010, June, 2010

[2] L. Sweeney, "k-anonymity : a model for protecting privacy.", International Journal on Uncertainty Fuzziness knowledge-Based System 10(5):557-570

[3] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M (2006), "L-diversity: privacy beyond k-anonymity." In: Proceedings of the 22nd IEEE international conference on data engineering (ICDE'06),
IEEE Computer Society, Washington, DC

[4] Samarati P, Sweeney L (1998), "Generalizing data to provide anonymity when disclosing information.", In: Proceedings of the 7th ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems(PODS'98), ACM Press, New York, p 188