■ 2010년도 학생논문 경진대회 수상작

소셜 네트워크 데이터의 프라이버시 보호 배포를 위한 모델

(A Model for Privacy Preserving Publication of Social Network Data)

성 민 경 † 정 연 돈 ††
(Min Kyung Sung) (Yon Dohn Chung)

요 약 최근 빠르게 확산되고 있는 온라인 소셜 네트워크 서비스는 수많은 데이터를 저장하고 이를 분석하여 여러 연구 분야에 활용하고 있다. 정보의 효율성을 높이기 위해 기업이나 공공기관은 자신들이 가진 데이터를 배포하고, 배포된 데이터를 이용하여 여러 목적에 사용한다. 그러나 배포되는 소셜 네트워크에는 개인과 관련된 정보가 포함되어 있으므로 개인 프라이버시가 노출될 수 있는 문제가 있다. 배포되는 소셜 네트워크에서 단순히 이름 등의 식별자를 지우는 것으로는 개인 프라이버시 보호에 충분하지 않으며, 소셜 네트워크가 가진 구조적 정보에 의해서도 개인 프라이버시가 노출될 수 있다. 본 논문에서는 내용 정보를 포함하고 있는 소셜 네트워크 배포 시 개인 프라이버시 노출에 이용되는 복합된 공격법을 제시하고 이를 방지할 수 있는 새로운 모델인 ℓ-차수 다양성(ℓ-degree diversity)을 제안한다. ℓ-차수 다양성은 소셜 네트워크 데이터 배포에서 ℓ-다양성을 최초로 적용한 모델이며 높은 정보 보존율을 가짐을 실험을 통해 볼 수 있다.

키워드 : 소셜 네트워크, 프라이버시, 데이터 배포, k-익명성(k-anonymity), ℓ-다양성(ℓ-diversity), ℓ-차수 다양성(ℓ-degree diversity)

Abstract Online social network services that are rapidly growing recently store tremendous data and analyze them for many research areas. To enhance the effectiveness of information, companies or public institutions publish their data and utilize the published data for many purposes. However, a social network containing information of individuals may cause a privacy disclosure problem. Eliminating identifiers such as names is not effective for the privacy protection, since private information can be inferred through the structural information of a social network. In this paper, we consider a new complex attack type that uses both the content and structure information, and propose a model, ℓ-degree diversity, for the privacy preserving publication of the social network data against such attacks. ℓ-degree diversity is the first model for applying ℓ-diversity to social network data publication and through the experiments it shows high data preservation rate.

Key words : Social network, Privacy, Data publication, k-anonymity, ℓ-diversity, ℓ-degree diversity

· 이 연구에 참여한 연구자(의 일부)는 2단계 BK21 사업의 지원을 받았음

† 학생회원 : 고려대학교 컴퓨터 전파통신공학과
mj999@korea.ac.kr
†† 정 회 원 : 고려대학교 컴퓨터학과 교수
ydchung@korea.ac.kr

논문접수 : 2010년 5월 25일
심사완료 : 2010년 7월 7일

정보과학회논문지: 데이타베이스 제37권 제4호(2010.8)

1. 서 론

인터넷의 지속적인 발전에 따라 수많은 정보가 공유, 수집, 분석 되고 있으며 기관이나 조직에서는 저장된 개인정보를 분석하여 의학연구, 인구통계 등 다양한 연구 분야에 활용하고 있다. 또한 최근 들어 페이스북, 버즈, 트위터, 싸이월드와 같은 온라인 소셜 네트워크 서비스(online social network service)사용자가 빠르게 증가하여, 소셜 네트워크와 관련된 개인정보의 양도 늘어나고 있다. 기관이나 조직은 이를 이용하여 마케팅, 유전

표 1 배포된 주소록 데이터

이름	생년월일	주소
서현	1991.6.8	서울시 종로구 효자동
지연	1987.11.6	수원시 장안구 조원동
승엽	1982.6.28	서울시 중랑구 면목동
지성	1981.3.17	수원시 팔달구 파장동
재석	1972.12.1	서울시 강서구 등촌동

표 2 이웃의 수

이름	이웃
서현	2
지연	2
승엽	1
지성	5
재석	1

병 연구, 인적네트워크 분석 등의 연구를 한다. 정보의 활용도를 높이기 위해 기관이나 조직은 서로의 정보를 공유하거나 공공의 목적으로 배포하기도 한다. 예를 들어, 표 1은 배포된 동호회 주소록이며, 표 2는 온라인 소셜 네트워크 서비스에서 배포한 이웃의 수 정보이다. 그러나 배포된 정보로 인해 개인의 민감한 정보가 노출되지 않도록 해야 한다. [1]의 연구에 따르면 미국 인구의 약 87%는 성별, 생년월일, 5자리 ZIP코드의 단 세 가지 정보로 개인이 유일하게 판별된다고 하였다. 개인이 유일하게 판별될 경우 다른 정보와 결합을 통해 개인의 민감한 정보가 드러날 수 있다. 더욱이 소셜 네트워크는 사람들 간의 관계까지 표현하므로 개인정보 보호를 위해 소셜 네트워크와 관련된 데이터 배포 시 각별한 주의가 필요하다.

의학 연구 기관에서, 소셜 네트워크에 관한 데이터를 배포한다고 하자. 소셜 네트워크는 그래프로 표현되며 그래프의 정점은 개인을 나타내고 정점간의 간선은 개인 간의 관계를 나타낸다. 각 정점은 각 개인에 대한 정보를 가지고 있다. 그림 1(a)는 소셜 네트워크 원시데이터이며 각 정점은 개인의 이름, 주민등록번호, 주소, 나이, 질병 정보를 포함하고 있다(편의상 주민등록번호는 4자리로 나타낸다). 정점에 포함된 이러한 정보를 내용 정보라 한다. 만약 그림 1(a)가 다른 기관이나 공공으로 배포된다면 개인의 민감한 정보인 질병이 노출된다. 이와 같은 개인의 민감한 정보 노출을 방지하기 위해 소셜 네트워크 데이터를 배포할 때 개인을 유일하게 판별할 수 있는 이름과 주민등록번호를 삭제하여야 한다. 그러나 단순히 이름과 주민등록번호를 삭제하는 것만으로는 개인 정보가 보호되지 않는다. 이름과 주민등록번호가 삭제된 그림 1(b)의 소셜 네트워크가 표 1의 주소록 데이터와 결합될 경우 각 개인이 유일하게 판별되어 개인의 민감한 정보가 노출된다. 이러한 형태의 프라이버시 노출 공격을 결합 공격(linkage attack)이라 한다.

다른 데이터와의 결합을 통한 결합 공격으로 인한 정보 노출을 방지하기 위해 k-익명성(k-anonymity) 모델[2]이나 ℓ-다양성(ℓ-diversity) 모델[3]과 같은 데이터 익명화 모델들이 제안되었다. 이들 모델은 내용 정보를 수정하여 그림 1(c)와 같이 익명화된 소셜 네트워크 데이터를 만든다. k-익명성 모델은 데이터 집합에서 서로 구분이 가지 않는 개체가 적어도 k개 이상 존재하도록 함으로써 결합 공격을 방지하는 모델이다. ℓ-다양성 모델은 서로 구분이 가지 않는 객체들에서 민감한 정보가 최소 ℓ개 이상 존재하도록 함으로써 개체의 민감한 정보 노출을 방지하는 모델이다(두 모델에 대한 자세한 설명은 2장 관련연구에서 한다).

그러나 그림 1(c) 또한 소셜 네트워크가 가진 구조적 특성으로 인해 개인정보를 보호하기에 충분하지 않다. 구조적 특성 중 하나인 정점의 차수는 소셜 네트워크에서 개인이 얼마나 많은 사람들과 관계를 맺고 있는지 나타낸다. 그림 1(c)가 표 2의 이웃의 수(차수) 데이터와 합쳐질 경우 개인이 유일하게 판별되어 민감한 정보가 노출된다. 예를 들어, 표 2에서 이웃의 수(차수)가 5인 사람은 지성이 유일한데, 그림 1(c)에서 차수가 5인 정점은 <수원시, 24-28, 피부염>의 내용 정보를 가진 정점으로 유일해서 지성이 피부염을 앓고 있다는 개인의 민감한 정보가 드러난다.

이러한 개인정보 노출 문제를 해결하기 위해 본 논문은 정점에 각 개인에 대한 내용 정보를 가지고 있는 소셜 네트워크 데이터에 대하여 결합 공격에 의한 개인정보 노출을 방지하는 모델을 제안한다. 내용 정보 보호를 위해 소셜 네트워크에 포함된 내용 정보 수정을 통해 ℓ-다양성 모델을 만족하도록 하며, 구조 정보 보호를 위

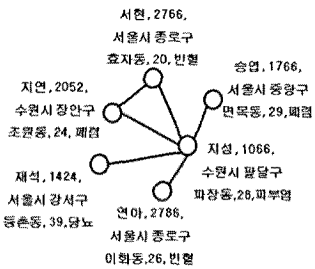

그림 1(a) 소셜 네트워크 데이터,

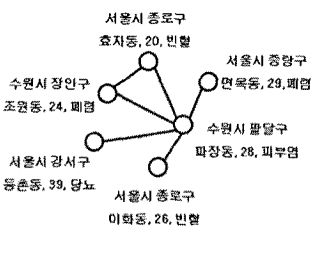

그림 1(b) 이름과 주민번호 삭제,

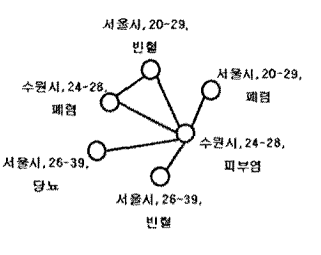

그림 1(c) k-anonymity 모델에 따라 수정

해 소셜 네트워크에 간선을 추가하여 이웃의 수에 의해 개인이 판별되지 않는 수정된 소셜 네트워크를 만들어 배포하는 것이 본 논문에서 제안하는 모델의 목적이다.

개인 내용 정보를 포함한 소셜 네트워크에서 프라이버시 침해 문제를 해결하기 위한 몇몇 기법들이 제시되었다[4-6]. Zheleva 등은 소셜 네트워크에서 간선의 민감한 정보, 즉 개인 간의 관계에 민감한 정보가 있는 상황에서 개인정보 보호 기법을 제시하였다[4]. 이 연구는 정점의 개인정보를 다루는 본 연구와는 해결하고자 하는 문제가 다르다. Campan 등은 소셜 네트워크의 정점 내용 정보를 k-익명성 모델을 이용하여 보호하는 기법을 제시하였다[5]. 그러나 본 논문은 k-익명성 보다 강력한 프라이버시 조건을 가진 ℓ-다양성 모델을 만족하는 기법을 다룬다. 또한 [5]는 구조적 공격에 의한 개인정보 노출방지를 위해 정점의 클러스터링을 통한 전체 소셜 네트워크를 추상적 형태로 표현하기 때문에 전체 소셜 네트워크 구조를 왜곡하는 문제점이 있다. Wei 등은 k-익명성을 만족하면서, 간선의 추가/삭제를 통한 구조적 공격 방지 기법을 제안하였다[6]. 그러나 이 기법은 정점에 포함된 개인 내용 정보가 나이, 연봉 등의 숫자정보인 경우만 고려하였다. 즉, 국적, 주소 등의 계층적 정보는 고려하지 않았다. 게다가 구조적 공격을 방지하기 위한 간선 추가/삭제 과정에서 단지 근접한 정점들끼리 묶어 부분그래프를 형성하여 간선을 추가/삭제하는 기법을 사용하기 때문에 전체 소셜 네트워크 구조가 크게 손상되는 문제점이 있다. 또한, ℓ-다양성 모델로 확장할 때 근접한 정점에 ℓ-다양성 모델을 만족시킬 수 있는 정점이 없으면 전체 구조가 더욱 손상되므로 쉽게 ℓ-다양성 모델을 적용할 수 없다.

소셜 네트워크의 구조 정보를 크게 훼손하는 기존 기법과 대비하여 본 논문에서 제안하는 기법의 차별성 및 독창성은 다음과 같다.

- 전체 소셜 네트워크 연결성을 고려한 구조 수정으로 소셜 네트워크 구조적 정보 보존
- k-익명성 모델만을 만족하는 기존 기법과 달리 ℓ-다양성 모델을 만족하는 소셜 네트워크를 생성하는 새로운 모델인 ℓ-차수 다양성(ℓ-degree diversity) 모델 제안
- 새로운 타입의 공격법인 DA공격에 대한 방어 기법 제안

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 살펴보고 3장에서는 문제를 정의한다. 4장에서는 ℓ-차수 다양성 모델에 대해 알아보며, 5장에서는 ℓ-차수 다양성 모델에 대한 알고리즘을 제안한다. 6장에서는 평가를 통해 제안기법의 성능을 살펴본다. 7장에서 결론과 함께 본 논문을 마무리 한다.

2. 관련 연구

데이터 배포 상황에서 개인정보 보호를 위한 연구는 지난 수년간 활발히 진행되어왔다. 테이블로 구성된 관계형 데이터(relational data)와 관련된 연구는 지속적으로 진행되고 있으며, 최근 그래프로 구성된 소셜 네트워크 데이터와 관련된 연구가 주목받고 있다.

2.1 관계형 데이터 관련연구

개인 정보가 포함된 관계형 데이터에서 테이블의 각 레코드(record)는 개인을 나타내고 어트리뷰트(attribute)는 개인의 각 속성을 나타낸다. 관계형 데이터에서 개인 식별을 위한 공격법은 크게 결합공격(linkage attack)과 배경지식공격(background attack)이 있다. 결합공격은 한 데이터 테이블과 다른 데이터 테이블의 결합(join)을 통해 생성된 데이터 테이블을 사용하여 개인을 식별함으로써 민감한 정보를 알아내는 방법이다.

k-익명화성 모델 : k-익명성 모델은 데이터 테이블에서 모든 레코드가 서로 구분이 가지 않는 k-1개 이상의 레코드를 가지도록 한다[2].

ℓ-다양성 모델 : 그러나 결합공격에 배경지식공격이 더해질 경우 k-익명성 모델은 개인 정보를 완벽히 보호하지 못함을 Machanavajjahala 등이 밝혀, 결합공격과 배경지식공격을 동시에 방어하기 위한 ℓ-다양성 모델을 제시하였다[3]. ℓ-다양성 모델은 서로 구분이 가지 않는 레코드들 사이에서 민감한 정보는 최소한 ℓ개 이상이어야 개인 정보가 보호된다는 조건을 가진다. 즉, 서로 구분이 가지 않는 레코드들이 민감한 정보까지 같은 값을 가진다면 서로 구분이 가지 않는 것이 의미가 없음을 보이고 이것을 동질성 공격(homogeneity attack)이라 하였다. 동질성 공격은 결합공격을 시도했을 때 나타날 수 있는 공격의 한 종류이다. 또한 ℓ-다양성 모델은 공격자가 배경지식을 이용하여 특정 레코드가 특정 민감한 값을 가질 수 없음을 추측하여 해당 특정 레코드 및 다른 레코드의 민감한 값을 추론하는 공격법까지 방지한다. 이러한 추론 공격법을 배경지식공격 이라고 한다. 즉, ℓ-다양성 모델은 k-익명성 모델의 문제점인 동질성 공격과 배경지식공격에 대한 방어책을 제시하였다.

m-불변성 (m-invariance) 모델 : m-불변성 모델은 테이블 데이터에 레코드가 추가/삭제되는 상황에서 개인정보 보호를 할 수 있는 기법으로, 모조 데이터(counterfeit)추가를 통해 시간에 따라 배포되는 여러 데이터를 결합하여도 개인이 용이하게 식별되지 않게 하였다[7]. 이 외에도 [8,9] 등이 동적인 데이터 배포에서 프라이버시를 고려하였다.

2.2 소셜 네트워크 데이터 관련연구

관계형 데이터 관련연구는 테이블로 표현된 데이터를

다루는 기법이므로 그래프로 표현되는 소셜 네트워크 데이터에 바로 적용하는 것은 불가능하다. 이런 이유로 최근 소셜 네트워크 데이터와 관련된 개인정보 보호 기법에 관한 연구가 많은 주목을 받고 있다. 소셜 네트워크의 가장 큰 특징은 구조적 정보를 가지고 있다는 것이므로 구조적 정보에 초점을 맞춘 기법들이 제안되었다[10-14]. 한 정점과 그와 바로 연결된 정점들의 구조를 이용한 공격법인 1-이웃 그래프(1-neighborhood graph) 공격을 방지하기 위한 기법이 연구 되었으며[11], 정점의 차수를 이용한 공격법을 방지하기 위한 기법인 k-차수 익명성(k-degree anonymity) 기법도 연구되었다[12]. 1-이웃 그래프 공격과 정점의 차수를 이용한 공격을 동시에 방지하는 기법으로 k-후보자 익명성(k-candidate anonymity) 기법도 제안되었다[13]. 이 기법은 어떠한 질의 그래프에 대해서도 k개 이상의 후보 부분그래프가 존재하여, 구조적 공격으로 인한 소셜 네트워크에서 개인의 용이한 식별을 방지한다. 최근에는 소셜 네트워크가 업데이트되는 상황에서 개인정보 보호가 가능한 기법인 k-자기 동형(k-automorphism) 모델이 제안되었다[14].

이상의 기법들은 소셜 네트워크에서 공격자의 구조적 공격만을 고려한 반면, 구조적 공격과 함께 정점이 가진 정보를 이용한 공격을 방지하기 위한 기법들이 연구되었다[4-6]. Zheleva 등은 소셜 네트워크에서 각 간선의 정보를 보호하기 위한 기법을 제시하였다. 이 기법은 클러스터링 방법을 사용하여 다섯 가지 간선 정보 보호 기법을 제시하여 각 방법에서 간선 정보 노출 정도를 실험을 통해 보여주었다[4]. 정점의 정보를 보호하는 기법으로 [5]가 제시되어 클러스터링 방법을 이용하여 정점의 정보가 k-익명성 모델을 만족하도록 함으로써 소셜 네트워크에서 개인의 용이한 식별을 방지하였다. Wei 등은 전체 소셜 네트워크를 k개의 부분그래프로 나누어 각 부분 그래프 내에서는 어떤 정점도 서로 구분되지 않게 하여 개인의 식별을 방지하였다[6].

3. 문제 정의

3.1 데이터 모델

정의 1. 소셜 네트워크 G(V, E) 소셜 네트워크에 속한 개인의 집합을 $V=\{v_1, v_2 \ldots, v_n\}$, 개인 간의 관계의 집합을 $E=\{e_1, e_2, \ldots, e_m\}$라 할 때, 소셜 네트워크 G는 G(V, E)로 나타낸다. □

정점은 각 개인을 나타내며, 개인과 관련된 내용 정보를 가지고 있다. 예를 들어 나이, 성별, 주소, 질병, 연봉 등이 이에 속한다. 내용 정보는 식별자, 준식별자, 민감한 속성의 세 가지로 나눌 수 있다.

*식별자(Identifier)*는 ID로 나타내며 이름, 주민등록번호와 같이 개인을 식별할 수 있는 데이터를 뜻한다.

준식별자(Quasi-Identifier : QI)는 직접적으로 개인을 식별할 수 없지만 외부정보와 합쳐져 개인을 판별할 수도 있는 정보를 뜻한다. 주소, 나이, 성별 등이 이에 속한다. QI는 다시 *수치적 준식별자*(Numerical Quasi-Identifier : NQI)와 *계층적 준식별자*(Hierarchical Quasi-Identifier : HQI)로 나누어진다. 수치적 준식별자는 데이터의 도메인(domain)이 연속된 숫자인 것으로 나이와 같은 데이터가 이에 속한다. 계층적 준식별자는 데이터의 도메인이 계층적 구조를 가진 것으로 주소와 같은 데이터가 이에 속한다. 즉, 상위 개념은 더 추상적이고 하위 개념은 더 구체적인 데이터다.

민감한 속성(Sensitive Attribute : SA)은 개인과 대응되어 추측되면 안 되는 데이터이다. 예를 들어, 질병, 재산 등이 이에 속한다. 설명의 편의를 위해 본 논문에서는 이름, 주민등록번호는 ID, 주소, 나이는 QI, 질병은 SA라고 가정한다.

소셜 네트워크를 배포할 때 개인이 식별되어 SA가 노출되는 것을 방지하기 위해 ID는 삭제하고 QI는 수정하고 SA는 그대로 남겨두는 상황을 가정한다. ID를 삭제하는 이유는 소셜 네트워크 배포만으로 개인의 SA가 노출되는 것을 방지하기 위함이고, QI를 수정하는 이유는 외부 데이터와 결합되어 개인이 식별됨으로써 SA가 노출될 확률을 줄이기 위함이다. SA를 그대로 남겨두는 이유는 데이터 활용도를 높이기 위해서이다.

3.2 DA 공격

본 논문은 공격자가 ID와 QI가 함께 있는 테이블(예를 들어, 표 1. 주소록), QI와 SA가 함께 있는 테이블(예를 들어, 배포된 병원데이터) 그리고 ID와 차수가 함께 있는 테이블(예를 들어, 표 2. 소셜 네트워크에서 개인별 이웃의 수)을 정보로 가지고 있다고 가정한다. 이 때 공격자는 차수(degree)까지 포함된 결합공격을 시도할 수 있으며, 이 새로운 공격 타입을 DA(Degree +Attribute) 공격이라 한다.

예를 들어, 그림 1(c)가 배포되었을 때, 공격자가 표 1 뿐 아니라 구조 정보인 표 2까지 이용한 결합 공격을 DA 공격이라 한다. 공격자는 DA 공격을 통해 개인을 식별할 수 있을 뿐 아니라 민감한 정보까지 추측할 수 있다. DA 공격을 방어하기 위해선 배포된 소셜 네트워가 ℓ-다양성 모델도 만족해야 한다. 즉, QI와 차수에 의해 서로 구분이 가지 않는 정점집합에서 서로 다른 SA 값이 최소한 ℓ개 이상 있어야 한다.

4. ℓ-차수 다양성(ℓ-degree diversity)

DA 공격을 막기 위해 소셜 네트워크의 서로 구분되지 않는 정점들 사이에서 서로 다른 SA 값은 ℓ개 이

상이어야 한다. QI에 의해 서로 구분되지 않는 정점들의 집합을 동질클래스(Equivalence Class : EC)라 한다. 이때 서로 구분되지 않는 정점의 수는 당연히 ℓ과 같거나 그보다 커야한다. 즉, $|EC| \geq \ell$

본 논문에서 제안하는 ℓ-차수 다양성 모델은 소셜 네트워크에서 각 정점이 내용 정보(QI)와 구조 정보(차수)에 의해 식별되지 않아야 하며, QI에 의해 서로 구분되지 않는 EC내의 정점들의 서로 다른 SA는 최소 ℓ개 이상이어야 한다.

정의 2. ℓ-차수 다양성 주어진 소셜 네트워크 G에 대하여, 각 EC의 서로 다른 SA가 ℓ개 이상이어야 한다.

정리 1. ℓ-차수 다양성 모델에 의해 생성된 소셜 네트워크 G_T는 k-익명성 모델과 ℓ-다양성 모델을 만족한다.

증명. ℓ-차수 다양성 모델은 QI와 차수에 의해 서로 구분 되지 않는 EC를 생성하고 각 EC에 속한 정점의 서로 구분되지 않는 SA는 최소 ℓ개 이상 가지게 한다. 이것은 QI에 의해 서로 구분 되지 않는 EC를 생성하고 각 EC에 속한 개체가 서로 구분되지 않는 SA를 최소 ℓ개 이상 가지게 하는 ℓ-다양성 모델을 만족한다. 또한 ℓ-다양성 모델은 k-익명성 모델을 만족하므로, ℓ-다양성 모델을 만족하는 ℓ-차수 다양성 모델은 k-익명성 모델을 만족한다. □

예제 1. 표 3은 배포된 병원 데이터이다. 나이와 성별은 QI이며 질병은 SA라 하자. 각 환자의 이름과 주민등록 번호는 나와 있지 않고 환자 번호로 구분되어 있다. $|N|=6$인 표 3은 k=3에 대해 3-익명화를 다음과 같이 만족한다. $EC_1=\{1, 2, 3\}$, $EC_2=\{4, 5, 6\}$. 그러나 표 3은 $\ell=3$에 대해 3-다양성 모델을 만족하지 못한다. □

표 3 배포된 병원 데이터

환자번호	나이	성별	질병
1	27	여	독감
2	27	여	독감
3	27	여	독감
4	27	여	독감
5	27	여	골절
6	27	여	소화불량

주어진 데이터가 ℓ에 대하여 ℓ-다양성 모델을 만족하는 지 판별하기 위해서 본 논문에서는 실현가능한 ℓ 판별 기법을 쓴다.

정리 2. 실현가능한 ℓ 판별 주어진 데이터에 대하여 서로 구분이 되는 SA 값을 세어서 그 수를 큰 순으로 각각 $b_1, b_2, \ldots, b_n$ 이라 하자($b_1 \geq b_2 \geq \ldots \geq b_n$). 수열 $B=\{b_1, b_2, \ldots, b_n\}$에 대하여 수열의 첫 번째 원소에서 ℓ 번째 원소까지 1씩 빼서 수열 $\{b_1', b_2', \ldots b_n'\}$을 얻는다. 수열 $\{b_1', b_2', \ldots b_n'\}$에서 값이 0인 것은 수열에서 제외하고 새로운 수열 $B'=\{b_1'', b_2'', \ldots b_m\}$을 얻는다($n \geq m$). 새로운 수열 $B'=\{b_1'', b_2'', \ldots b_m\}$에 대하여 $m \geq \ell$ 이라면 즉, $|B'| \geq \ell$ 이라면 주어진 데이터는 ℓ에 대하여 ℓ-다양성 모델을 만족하는 EC를 생성 가능하다.

증명. 증명 생략.

예제 2. 예제 1의 표 3에 대해서 $b_1=4$, $b_2=1$, $b_3=1$이다(b_1은 독감, b_2는 골절, b_3는 소화불량). 수열 $B=\{4, 1, 1\}$에서 $\ell=3$이므로 첫 번째부터 3번째 원소까지 1씩 빼서 $\{3, 0, 0\}$을 얻고, 수열의 원소에서 0인 것을 빼서 $B'=\{3\}$을 얻는다. 결과 수열인 B'은 $|B'| \geq 3$을 만족하지 못하므로 주어진 표 3은 $\ell=3$에 대해서 3-다양성 모델을 만족하지 못한다.

5. 소셜 네트워크 익명화 알고리즘

5.1 EC 구성

주어진 데이터가 ℓ에 대해서 ℓ-다양성 모델을 만족하는 EC를 생성할 수 있다면 몇 개의 EC를 생성해야 할지 고려해야 한다. EC 생성을 통해 같은 EC에 속한 정점들은 QI와 차수 수정을 통해 익명화 된다. 이 때, EC에 속한 정점의 수가 많아질수록 EC에 속한 모든 정점이 같은 QI와 차수를 가지게 되므로 정보손실이 늘어난다.

같은 EC에 속한 정점들은 ℓ-다양성 모델을 만족하며, 그 수가 최소가 되어야 수정되는 정보의 양이 적어 정보손실이 줄어든다는 것을 알 수 있다. 이것은 EC의 수가 많을수록 정보손실이 줄어든다는 것을 의미한다. k-익명성 모델을 만족하는 최대 EC의 수는 전체 정점의 수를 $|N|$이라 했을 때 $\left\lfloor \frac{|N|}{k} \right\rfloor$ 이지만 ℓ-다양성 모델을 위해서는 더욱 세밀한 조건이 필요하다.

예제 3. 표 3에서 k=2일 때, 2-익명성 모델을 만족하는 최대 EC는 3개이다($EC_1=\{1, 2\}$, $EC_2=\{3, 4\}$, $EC_3=\{5, 6\}$). 그러나 표 3은 $\ell=2$일 때, 최대 EC의 개수는 2개이다.($EC_1=\{1, 2, 5\}$, $EC_2=\{3, 4, 6\}$) □

주어진 데이터와 ℓ에 대하여 EC의 최대 개수를 구하기 위해 본 논문에서는 최대 EC 판별 기법을 쓴다.

정리 3. 최대 EC 판별 주어진 데이터에 대하여 서로 구분이 되는 SA 값을 세어서 그 수를 큰 순으로 각각 $b_1, b_2, \ldots, b_n$ 이라 하자($b_1 \geq b_2 \geq \ldots \geq b_n$). 최대 EC의 개수를 구하기 위해 ECC(EC Count)는 0으로 설정한다. 수열 $B=\{b_1, b_2, \ldots, b_n\}$에 대하여

i) 수열의 첫 번째 원소에서 ℓ 번째 원소까지 1씩 빼서 수열 $\{b_1', b_2', \ldots, b_n'\}$을 얻는다.

ii) 수열 $\{b_1', b_2', \ldots b_n'\}$에서 0인 것은 수열에서 제

외하고 새로운 수열 B′={b_1'', b_2'', ...b_m}을 얻는다($n \geq m$).

iii) 새로운 수열 B′= {b_1'', b_2'', ...b_m}이 생성되면 ECC를 1 증가시킨다.

B′에 대하여 $m \geq \ell$이라면 즉, $|B'| \geq \ell$이라면 B′을 B로 설정하고 i ~ iii 과정을 반복한다. 더 이상 i 의 과정을 실행할 수 없을 때 ECC가 주어진 데이터와 ℓ에 대해 최대 EC 개수가 된다.

증명. 증명 생략.

예제 4. 예제 3의 표 3에서 b_1=4, b_2=1, b_3=1 이다. 수열 B={4, 1, 1}에서 ℓ=2 이므로 첫 번째 원소부터 2번째 원소까지 1씩 빼서 B={3,0,1}을 얻고 0인 것을 제외해서 B′={3, 1}을 얻으며 ECC는 1이 된다. $|B'| \geq 2$ 이므로 B={3, 1}로 해서 B′={2}을 얻고 ECC는 2가 된다. 이때, $|B'| \geq 2$를 만족하지 못하므로 표 3과 ℓ=2에 대한 최대 EC의 개수는 2가 된다. □

최대 EC의 개수를 구하면 소셜 네트워크의 각 정점들로 EC를 구성한다. 정보손실을 줄이기 위해 유사한 정점끼리 EC를 구성한다. ℓ-다양성 모델을 만족하는 EC를 구성하기 위해서는 두 가지 방법이 있다. 첫 번째 방법은 EC를 구성할 때 SA를 고려해서 각 정점을 나누는 것으로 많은 선행 연구에서 이용한 기법이다 [15,16]. 두 번째 방법은 EC를 구성한 후 ℓ-다양성 모델을 만족하지 못하는 EC를 정점 교환(swap)을 통해 ℓ-다양성 모델을 만족하게 하는 것이다. 본 논문에서 제안하는 ℓ-차수 다양성 모델은 두 번째 방법을 이용하여, 정보손실을 최소화하는 EC를 구성한 후 ℓ-다양성 모델을 만족하지 못하는 EC에 대해 효과적인 정점 교환 기법을 이용한다.

EC를 구성할 때 유사한 정점끼리 하나의 EC를 구성하면 정보손실을 줄일 수 있다. 유사한 정점이란 QI(내용 정보)와 차수가 비슷한 정점을 뜻한다. 유사한 정점끼리 EC를 구성하면 익명화를 위해 QI와 차수를 수정할 때 수정되는 정보의 양이 줄어들어 정보손실을 줄일 수 있다.

QI 수정은 같은 EC내에 속한 정점들이 모두 같은 QI 속성 값을 가지도록 수정한다. NQI 속성 데이터는 같은 EC내의 데이터 중 가장 큰 값과 가장 작은 값의 범위로 수정된다. HQI 속성 수정을 위해서는 각 HQI 속성에 대한 의미 분류 트리(taxonomy tree)가 필요하다. HQI는 계층적 정보를 가지고 있는 속성이므로 트리를 이용하여 계층 간의 포함 관계를 나타낸다. 그림 2는 주소에 대한 의미 분류 트리이다. HQI 속성 데이터는 같은 EC내의 값들이 의미 분류 트리의 최소 공통 조상 노드로 수정된다.

차수는 EC에 속한 정점들이 해당 EC에서 가장 큰 차수를 가지는 정점과 같은 차수를 가지게 수정한다. 이것은 간선 추가를 통한 방법을 쓰기 때문이며, 간선 삭제를 통한 방법을 위해서는 가장 작은 차수를 가지는 정점과 같은 차수를 가지게 하면 된다.

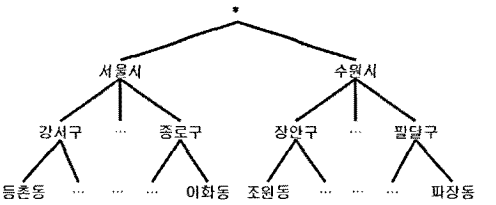

그림 2 주소에 대한 의미 분류 트리

예제 5. 그림 1에서 서현과 연아가 같은 EC에 속했다면 NQI는 <20-26>으로, HQI는 <서울시 종로구>으로, 차수는 2로 수정된다.

5.2 정점 유사도

유사한 정점을 구하기 위한 정점 유사도(inter Node Similarity : NS)는 정점 구조 유사도(inter Node Structure Similarity : N_SS)와 정점 내용 유사도(inter Node Content Similarity : N_CS)로 구성된다. N_CS는 [17]에서 제안한 정보손실공식을 본 논문에 맞는 N_CS로 수정한 것이다. [17]에서 제안한 공식이 두 개체 간 차이를 이용하여 손실되는 정보의 양을 측정한 것에 비해 본 논문의 N_CS는 두 개체 간 유사성을 이용하여 유사도를 측정하였다.

정의 3. 정점 구조 유사도(inter Node Structure Similarity : N_SS) 소셜 네트워크에서 두 정점 N_1과 N_2의 차수를 각각 $N_1.d$, $N_2.d$라 했을 때, N_1과 N_2간의 N_SS는 다음과 같다.

$$\frac{\max(N_1.d, N_2.d) - |N_1.d - N_2.d|}{\max(N_1.d, N_2.d)}$$ □

N_SS는 두 정점 간 차수의 유사성을 두 정점의 차수 중 큰 수로 나누어 일반화 한다. 즉, 차수가 작은 정점과 차수가 큰 정점이 같은 차수를 가지기 위해 차수가 작은 정점에 추가되는 차수의 수가 작을수록 N_SS는 큰 값이 나온다.

정의 4. 정점 내용 유사도(inter Node Content Similarity : N_CS) 소셜 네트워크의 두 정점 N_1, N_2에 대해 각각의 준식별자 NQI, HQI를 N_1.NQI, N_1.HQI, N_2.NQI, N_2.HQI 라 하자. 두 정점 간 NQI 유사도는 NQI의 속성의 도메인 크기를 |NQI|라 할 때,

$$\frac{|NQI| - |N_1.NQI - N_2.NQI|}{|NQI|}$$

이다. NQI 속성이 f 개일 때, 각 NQI 유사도를 구해 합을 f로 나누어 준다. 두 정점 간 HQI 유사도는 HQI 속성의 도메인 높이(HQI속성에 대한 의미 분류 트리 높

이)를 Height(HQI), N_1.HQI와 N_2.HQI의 최소공통조상(least common ancestor)노드를 루트로 하는 트리를 LCAT(N_1.HQI, N_2.HQI)라 했을 때,

$$\frac{Height(HQI) - Height(LCAT(N_1.HQI, N_2.HQI))}{Height(HQI)}$$

이다. HQI 속성이 g개 일 때, 각 HQI 유사도를 구해 합을 g로 나누어 준다. N_CS는 각각 구한 NQI와 HQI의 합을 2로 나누어 일반화 한다. □

정의 5. 정점 유사도(inter Node Similarity : NS) 두 정점간의 NS는 N_SS와 N_CS의 합으로 구한다. 즉, NS = N_SS + N_CS 이다.

예제 8. 그림 1에서 지성과 지연의 NS는 $\frac{2}{3}+\frac{32}{57}=\frac{70}{57}$ 이다. □

유사한 정점끼리 EC를 구성하기 위해 정점 간 NS 값 비교를 통해 NS 값이 큰 정점끼리 최대 EC의 개수에 맞게 EC를 구성한다.

예제 9. 그림 1에서 ℓ=2 일 때, EC를 구성한다고 해보자.(각 EC에 두 개씩 정점이 들어감. SA 값은 고려 안함) 우선 최대 EC의 개수를 구하면 3개가 나온다. NS를 이용하여 각 정점을 3개의 EC에 배분하면 각각 EC_1 = {서현, 연아}, EC_2 = {지연, 지성}, EC_3 = {승엽, 재석}이 된다.

5.3 SA 교환 (SA Swapping)

예제 9에서 구성된 3개의 EC는 정보손실을 줄이기 위해 유사한 정점끼리 뭉쳤지만 아직 ℓ-다양성 모델 조건을 만족하지 못한다. EC_1의 서현, 연아는 모두 SA 값으로 빈혈을 가지고 있어 한 EC에서 서로 구분되는 SA값이 적어도 2개 이상이어야 하는 2-다양성 모델 조건에 맞지 않는다. 이를 위해 본 논문에서는 SAS(SA Swapping)기법을 통해 ℓ-다양성 모델 조건을 만족하는 EC를 구성한다. SAS는 ℓ-다양성 모델 조건을 만족하지 못하는 EC의 SA 값을 다른 EC의 SA 값과 교환하여 모든 EC가 ℓ-다양성 모델 조건을 만족하게 한다. 또한 SAS는 ℓ개의 축을 가지는 ℓ차원의 공간을 생성한다. 각 축은 같은 순서의 SA 값 좌표를 가진다.

예제 10. 그림 1에 대해서 2-다양성 모델을 만족하는 SAS 공간은 그림 3과 같다. □

생성된 SAS 공간에 각 EC를 SA 값에 맞추어 공간상에 투영한다. EC에 속한 SA 값이 ℓ개를 초과하면 중복되는 SA 값은 제외하고 ℓ개의 값으로만 공간에 투영한다. 중복되는 값을 제외하고도 ℓ개 이상의 서로 구분되는 SA 값이 있으면 해당 EC는 공간에 투영하지 않고 SwapList에 저장한다.

예제 11. 그림 1에 대해 예제 9와 같이 EC가 구성되었다고 했을 때, 예제 10의 공간에 각 EC를 투영하면

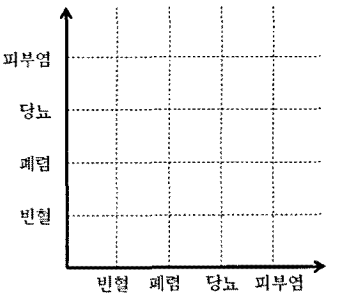

그림 3 그림 1의 2-다양성 모델 SAS

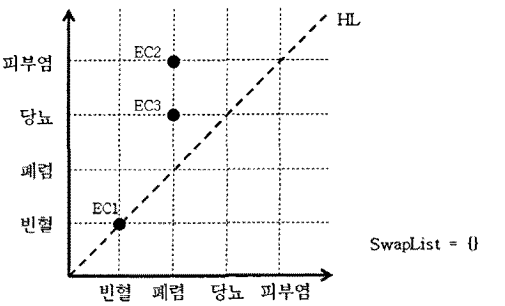

그림 4 각 EC를 투영한 모습

그림 4와 같다. EC_2는 SA 값으로 폐렴, 피부염을 가지고 있으므로 (폐렴, 피부염) 또는 (피부염, 폐렴)에 투영된다. SA 값의 순서는 관계가 없다. □

EC가 투영된 공간에서 각 축이 같은 값을 가지는 EC는 ℓ-다양성 모델을 만족하지 못한다. 예를 들어, 그림 4에서 점선으로 표시된 대각선 HL(Homogeneous Line)위에 있는 점은 각 축이 같은 값을 가지는 것이며, EC_1이 이에 속한다. SAS는 HL위에 있는 EC를 다른 EC와의 SA 값 교환을 통해 HL위에 EC가 존재하지 않게 하여 ℓ-다양성 모델을 만족하게 한다.

정리 4. HL(Homogeneous Line) SAS 공간에서 각 축이 같은 값을 가지는 점을 연결한 선. HL위에 있는 점은 ℓ-다양성 모델 조건을 만족하지 못한다.

HL위에 점을 교환할 때, 교환을 통해 다른 EC가 HL위에 속하게 된다면 의미가 없어진다. 교환을 통해 모든 EC가 HL위에 존재하지 않기 위해서는 HL위에 속한 EC와 교환 가능한 EC 후보를 선택하는 것이 중요하다.

교환 가능한 EC 후보를 구하면 HL위에 있는 EC와 하나의 정점 교환을 통해 모든 EC가 ℓ-다양성 모델을 만족하게 한다. 정점을 교환할 때 여러 가지 경우의 수가 발생할 수 있는데 교환 후 각 EC의 NS합이 가장 큰 값을 가지게 교환하면 된다. 교환 가능한 EC 후보가 두 개 이상일 때도 교환 후 각 EC의 NS합이 가장 큰 값을 가지게 교환하여 유사한 정점끼리 EC를 구성하게

한다. 여기서 중요한 점은 SwapList에 있는 EC도 교환 가능한 후보가 될 수 있으므로 SA 값 비교를 통해 가능 여부를 판단해야 한다.

SAS를 통해 각 EC가 ℓ-다양성 모델을 만족하면 QI 와 차수 수정을 통해 EC가 QI로 인해 서로 구분이 되지 않게 한다.

예제 13. 그림 1에서 SAS를 통해 2-다양성 모델을 만족하는 EC는 EC_1 = {서현, 승엽}, EC_2 = {지연, 지성}, EC_3 = {재석, 연아}이다. QI와 차수 수정을 통한 각 EC는 다음과 같다. □

EC_1

주소	나이	질병	차수
서울시	20-29	빈혈	2
서울시	20-29	폐렴	2

EC_2

주소	나이	질병	차수
수원시	24-28	폐렴	5
수원시	24-28	피부염	5

EC_3

주소	나이	질병	차수
서울시	26-39	당뇨	1
서울시	26-39	빈혈	1

ℓ-다양성 모델을 만족하는 EC가 생성되면 새로운 차수를 만족하는 실현 가능한 그래프가 존재하는지 판별해야 한다. 실현가능한 그래프 판별을 위해 [18]에서 제안한 기법을 이용한다. [18]의 기법은 주어진 차수의 내림차순 수열을 만들어 해당 수열이 제안 공식을 만족하는지 판별하고 만족하지 않는다면 차수 추가를 통해 제안 공식을 만족하는지 판별하는 과정을 반복적으로 수행하여 실현 가능한 그래프의 차수 수열을 생성한다.

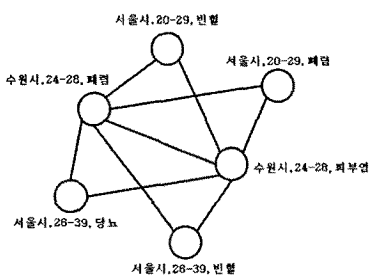

그림 5 그림 1에 대한 2-차수 다양성 모델을 만족하는 소셜 네트워크 G_T

Input original social network G, integer ℓ
Output ℓ-degree diverse social network G_T
Procedure
01: Discriminate a possibility of making G_T by G and ℓ
02: Calculate maximum number of EC
03: Form each EC by using NS
04: Get EC which do not satisfy ℓ-diversity
05: Swap SA of EC to satisfy ℓ-degree diversity
End of Procedure

그림 6 제안하는 소셜 네트워크 익명화 알고리즘

예제 14. 예제 13에서 생성된 차수의 내림차순 <5, 5, 2, 2, 1, 1>은 [18]의 기법에 따라 실현 가능한 그래프가 아니다. 따라서 [18]의 기법에 따라 실현 가능한 그래프의 차수 수열을 생성하면 <5, 5, 2, 2, 2, 2>가 된다. □

새롭게 생성된 차수 수열을 이용하여 추가되는 차수정보를 각 EC에 추가한다. 새롭게 수정된 EC를 이용하여 ℓ-차수 다양성 모델을 통해 DA 공격을 막을 수 있는 수정된 소셜 네트워크 GT를 만들고 이것을 배포한다.

예제 15. 그림 1의 소셜 네트워크 G 대해 DA공격을 방어하는 모델인 2-차수 다양성 모델을 만족하는 소셜 네트워크 G_T는 그림 5와 같다.

6. 평 가

6.1 실험 환경

본 장에서는 다양한 실험을 통해 제안 기법에 대한 성능 및 효율성에 대해 알아본다. 실험의 주목적은 제안 기법으로 생성된 소셜 네트워크 G_T가 원시 소셜 네트워크와 얼마나 차이가 나는지 측정하는 것이다. 구조 정보의 차이를 측정하기 위해 우리는 실험을 통해 추가된 간선의 수를 알아보았다. 내용 정보의 차이를 측정하기 위해 우리는 내용정보 보존율을 측정하였다. 내용 정보 보존율이란 원시 소셜 네트워크의 데이터의 각 QI 도메인의 전체 범위를 1로 봤을 때, 수정된 내용 정보를 측정하여 원시 내용 데이터의 보존 정도를 나타낸 것이다.

실험을 위해 UC Irvine Machine Learning Repository로부터 실제 성인 데이터를 추출하여 데이터 소실 등 오류가 있는 부분을 제외하고 실험 집합을 구축하였다[19]. 모든 실험에서 QI는 나이, 성별, 인종, 결혼 상태, 국가의 5가지로 하였으며 SA는 임의의 50가지 값을 생성하였다. QI에서 나이는 NQI이며 나머지는 HQI이다. 개인에 대한 구조 정보는 멱함수 분포 법칙을 따르는 척도 없는 그래프(scale-free-graph)로 구성하였다[20]. 이에 따라 본 실험에서는 멱함수 분포 법칙을 따르는 그래프를 GTGraph를 통해 생성했다[21].

6.2 실험 결과

소셜 네트워크의 확장성에 대해 알아보기 위해 정점의 수를 다양하게 변화시켜 가면서 실험을 했으며, 정점당 평균 간선수를 변화시켜 소셜 네트워크의 응집도가 높은 상황과 낮은 상황에 대한 비교를 하였다. 또한 ℓ

값의 변화를 통해 프라이버시 조건이 엄격할 때와 비교적 덜 엄격할 때에 따른 데이터 변화를 살펴보았다. 정점의 수는 1000, 5000, 25000개로 변화 시켰으며, 정점 당 평균 간선 수는 1, 5, 10, 20으로 변화 시켰다. ℓ 값은 2, 5, 10, 15, 20, 25로 변화시켰다. 각 실험 인자를 변화시키면서 실험을 수행한 결과는 그림 7-13과 같다.

그림 7은 정점 당 평균 간선 수가 1개일 때 제안 기법에 의해 수정된 소셜 네트워크의 내용 정보 보존율을 나타낸다. 소셜 네트워크 구성 시 연결되어 있지 않은 부분이 없도록 만들었기 때문에 정점 당 평균 간선 수가 1이면 모든 정점이 정확히 1개의 간선을 가진다. 때문에 간선의 추가는 없었으며, 유사한 정점으로 EC를 구성할 시 내용 정보만 고려한 결과 내용 정보 보존율은 거의 100%에 가까웠다.

그림 8과 그림 9는 정점 당 평균 간선 수가 5개 일 때 제안 기법의 결과를 보여준다. 그림 8에서 보듯 ℓ 이 증가하면 추가되는 간선의 수도 증가하였다. 특히 ℓ 이 10 이하일 경우 정점의 수에 관계없이 거의 일정한 간선이 추가 되었지만 ℓ 이 15 이상이 되면 정점의 수가 적을수록 추가되는 간선의 수가 적었다. 또한 ℓ =20 까지는 정점이 5000개 일 때와 25000개 일 때 거의 같은 성능을 보였다. 그림 9는 정점의 수가 적을수록 ℓ 이 커지면 내용 정보 보존율이 낮음을 보여준다. 이는 정점의 수가 많을수록 유사한 내용 정보를 가진 정점이 늘어나 내용 유사도가 높은 EC를 구성하기 용이하기 때문이다.

그림 10은 그림 8과 유사한 형태의 증가를 보이지만 추가되는 간선의 수가 더욱 많음을 알 수 있다. 정점 당 평균 간선의 수가 증가하면 비록 소셜 네트워크 그래프가 멱함수 법칙을 따르는 일정한 형태로 구성되었지만 정점 수의 분포가 커져서 더욱 많은 간선의 추가가 필요함을 알 수 있다. 그림 11은 정점의 수가 많을수록 내용 정보 보존율이 높음 확실히 보여준다. 그러나 그림 8과 비교하면 정점의 수가 적을 때 내용 정보 손실의 폭은 더욱 증가했음을 알 수 있다.

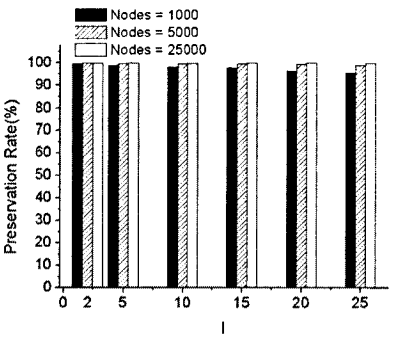

그림 7 정점 당 평균 간선 수 = 1

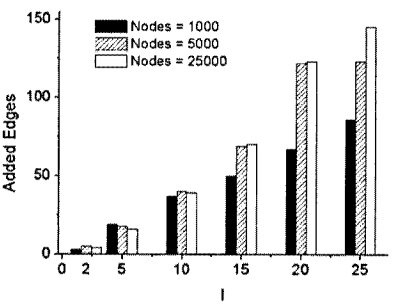

그림 8 정점 당 평균 간선 수 = 5

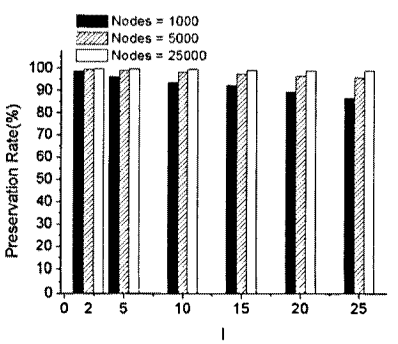

그림 9 정점 당 평균 간선 수 = 5

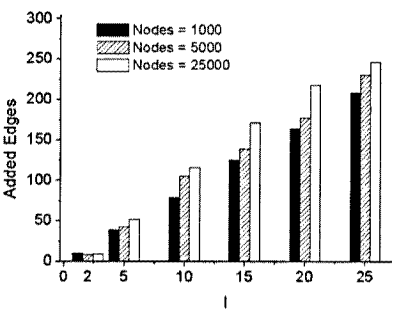

그림 10 정점 당 평균 간선 수 = 10

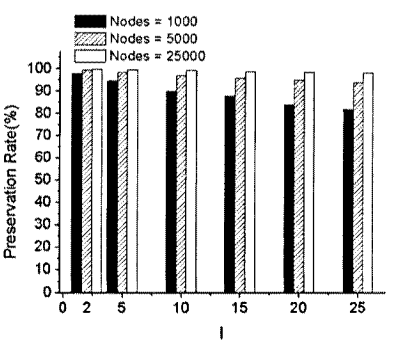

그림 11 정점 당 평균 간선 수 = 10

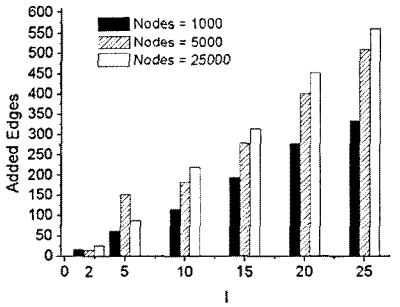

그림 12 정점 당 평균 간선 수 = 20

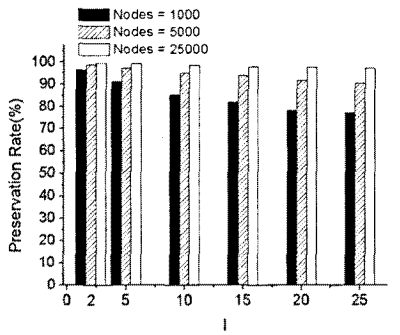

그림 13 정점 당 평균 간선 수 = 20

그림 12와 그림 13은 정점 당 평균 간선의 수가 20일 때의 결과를 보여준다. 정점 당 평균 간선의 수가 늘어나면 같은 ℓ 값에서 더욱 많은 간선의 추가가 됨을 다시 확인할 수 있다. 또한 정점 당 평균 간선의 수 증가는 정점의 수가 적을 때 더욱 많은 내용 정보 보존율 감소를 가져온다.

결과적으로 실험은 ℓ이 클수록 즉, 프라이버시 조건이 강해질수록 더욱 많은 간선의 추가가 발생하며 내용 정보 보존율은 낮아짐을 보였다. 정점 당 평균 간선의 수 가 증가하면 총 정점의 수가 많지 않을 때 내용 정보 보존율의 급격한 감소를 가져온다. 정점의 수가 많으면 유사한 내용 정보를 가지는 정점이 많아 NS가 높은 EC를 구성하기 쉬워 높은 내용 정보 보존율을 보이지만 추가되는 간선의 수는 정점의 수가 적을 때 보다 많음을 알 수 있었다.

7. 결 론

본 논문에서는 기존의 소셜 네트워크 프라이버시 보호 기법이 다루지 못했던 새로운 공격타입인 DA 공격을 제시하고 그것에 대한 프라이버시 보호 기법을 제시하였다. 제안 기법은 소셜 네트워크에 포함된 내용 정보에 의해 프라이버시가 노출되는 것을 막기 위해 ℓ-다양성 모델을 만족하게 내용 정보를 수정하였으며, 소셜 네트워크 차수 구조 정보로 인해 프라이버시가 노출되는 것을 막기 위해 간선 추가를 통해 정점을 익명화 하였다. 기존 기법이 내용 정보를 고려하지 않았거나 동질성 공격에 취약하였거나 클러스터링 기법을 이용하여 전체 구조 정보 손실을 초래하였지만 본 제안기법은 동질성 공격을 방어할 수 있는 내용 정보 고려와 함께 간선 추가 기법으로 구조 정보 손실을 줄일 수 있었다.

본 논문에서 제안한 기법은 소셜 네트워크의 또 다른 구조 정보 중 하나인 부분그래프(subgraph)를 고려하지 못한 단점이 있다. 이를 보완하기 위해 향후 연구에서는 부분그래프를 이용한 공격도 방어할 수 있는 기법에 대한 고려가 필요하며, 지속적으로 소셜 네트워크가 업데이트 되는 상황에서 프라이버시 보호 및 정보손실을 더욱 줄일 수 있는 기법에 대한 고려도 필요하다.

참 고 문 헌

[1] L. Sweeney, "Uniqueness of Simple Demographics in the U.S. Population," *Carnegie Mellon University, Laboratory for International Data Privacy*, 2000.

[2] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based System*, vol.10, no.3, pp.557-570, 2002.

[3] A. Machanavajjhala, J.Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *In Proceedings of International Conference on Data Engineering*, p.24, 2006.

[4] E. Zheleva, and L. Getoor, "Preserving the privacy of sensitive relationships in graph data," In *Proceedings of the 1st ACM SIGKDD international conference on Privacy, security, and trust in KDD*, pp.153-171, 2007.

[5] A. Campan, and T.M. Truta, "A clustering approach for data and structural anonymity in social networks." In Proceedings of the 2nd ACM SIGKDD international conference on Privacy, security, and trust in KDD, pp.33-54, 2008.

[6] Q. Wei, and Y. Lu, "Preservation of Privacy in Publishing Social Network Data," In *Proceedings of the 2008 International Symposium on Electronic Commerce and Security*, pp.421-425, 2008.

[7] X. Xiao, and Y. Tao, "m-invariance: Towards privacy preserving re-publication of dynamic datasets," In *Proceedings of the ACM SIGMOD*, pp.689-700, 2007.

[8] J.W. Byun, Y. Sohn, E. Bertino, and N. Li, "Secure anonymization for incremental datasets," In *Proceedings of the VLDB Workshop on Secure Data Management*, pp.48-63, 2006.

[9] B.C.M. Fung, K. Wang, A.W.C. Fu, and J. Pei,

"Anonymity for continuous data publishing," In *Proceedings of the 11th international conference on Extending database technology*, pp.264-275, 2008.
[10] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore art thou R3579X? anonymized social networks, hidden patterns, and structural steganography," In *Proceedings of the international conference on World Wide Web*, pp.181-190, 2007.
[11] B. Zou, and J. Pei, "Preserving privacy in social networks against neighbourhood attacks," In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pp.506-515, 2008.
[12] K. Liu, and E. Terzi, "Towards identity anonymization on graphs," In *Proceedings of the ACM SIGMOD*, pp.93-106, 2008.
[13] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, "Resisting structural re-identification in anonymized social networks," In *Proceedings of the VLDB Endowment*, vol.1, no.1, pp.102-114, 2008.
[14] L. Zou, L. Chen, and M.T. Ozsu, "k-automorphism: A general framework for privacy preserving network publication," *In Proceedings of the VLDB Endowment*, vol.2, no.1, pp.946-957, 2009.
[15] X. Xiao, and Y. Tao, "Anatomy : Simple and Effective Privacy Preservation," In *Proceedings of International Conference on Very Large Data Bases*, pp.139-150, 2006.
[16] Y. Ye, Q. Deng, C. Wang, D. Lv, Y. Liu and J. Feng, "BSGI : An effective algorithm towards stronger l-diversity," In *Proceedings of the 19th International conference on Database and Expert Systems Applications*, pp.19-32, 2008.
[17] J.W. Byun, A. Kamra, E. Bertino, and N.Li, "Efficient k-anonymization using clustering techniques," In *Proceedings of the 12th international conference on Database and systems for Advanced applications*, pp.188-200, 2007.
[18] P. Erdos, and T. Gallai, "Graphs with prescribed degrees of vertices," *Mat.Lapok*, 1960.
[19] D.J. Newman, S.Hettich, C.L. Blake, and C.J. Merz. "UCI Repository of Machine Learning Databases," *http://archive.ics.uci.edu/ml/*
[20] A.L. Barabasi, "Linked: The new science of networks," *Basic Book*, 2002.
[21] D.A. Bader and K. Madduri, "GTGraph: A synthetic graph generator suite," http://sdm.lbl.gov/~kamesh/software/GTgraph/

성 민 경

2009년 고려대학교 정보통신대학 컴퓨터과학 학사. 2009년~현재 고려대학교 정보통신대학 컴퓨터전파통신공학과 석박사 통합과정 재학 중. 관심분야는 데이터 프라이버시, 대용량 데이터 처리

정 연 돈

1994년, 고려대학교 전산과학과 졸업(학사). 1996년 한국과학기술원 전산학과 졸업(석사). 2000년 한국과학기술원 전산학 전공 졸업(박사). 2000년~2003년 한국과학기술원 전산학전공 Post-Doc. 연구원 및 연구교수. 2003년~2006년 동국대학교 컴퓨터공학과 교수. 2009년~2009년 Georgia Tech. 방문교수. 2006년~현재 고려대학교 컴퓨터·통신공학부 교수. 관심분야는 Database Privacy, Spatial Databases, Mobile Databases, Graph Databases, Data-Intensive Systems, Database Systems