韓國컴퓨터情報學會 論文誌
第15卷 第1號, 2010. 1.

2010-15-1-3-2

소셜 네트워크에서 구조정보와 내용정보를 고려한 프라이버시 보호 기법

성 민 경*, 이 기 용**, 정 연 돈*

A Privacy Protection Method in Social Networks Considering Structure and Content Information

Minh Kyoung Sung*, Ki Yong Lee**, Yon Dohn Chung*

요 약

지난 몇 년간 소셜 네트워크(Social network) 서비스는 급속도로 성장해 왔으며 향후 수년간 이러한 추세는 지속될 전망이다. 이에 따라 해당 기업, 공공기관에서는 다량의 소셜 네트워크 데이터를 보유하게 되었으며, 이 데이터를 배포하여 각종 연구 기관에서 인구통계, 통계분석 등의 연구 목적에 사용할 수 있다. 그러나 배포되는 소셜 네트워크 데이터는 외부정보와 결합되어 개인프라이버시 노출의 문제를 초래할 수 있다. 소셜 네트워크 데이터 소유자는 데이터를 배포하기 전 개인을 식별할 수 있는 명시적 정보를 삭제하거나 암호화해야 함은 물론 외부정보와 결합되어 개인프라이버시 노출의 문제를 발생시킬 가능성이 있는 데이터 또한 수정해야 한다. 데이터 수정 과정에서 수정되는 데이터의 양이 적을수록 데이터의 유용성은 높아진다. 본 논문에서는 소셜 네트워크 프라이버시 보호 기법과 관련된 기존 연구가 고려하지 않은 내용정보 고려 및 구조정보 왜곡을 보완하는 새로운 기법을 제안한다. 또한 다양한 실험결과를 통해 소셜 네트워크의 여러 환경에서 제안 기법의 확장성 및 타당성을 알아본다.

Abstract

Recently, social network services are rapidly growing and it is estimated that this trend will continue in the future. Social network data can be published for various purposes such as statistical analysis and population studies. When data publication, however, it may disclose the personal privacy of some people, since it can be combined with external information. Therefore, a social network data holder has to remove the identifiers of persons and modify data which have the potential to disclose the privacy of the persons by combining it with external information. The utility of data is maximized when the modification of data is minimized. In this paper, we propose a privacy protection method for social network data that considers both structural and content information. Previous work did not consider content information in the social network or distorted too much structural information. We also verify the effectiveness and applicability of the proposed method under various experimental conditions.

▸ Keyword : 프라이버시(privacy), 소셜 네트워크(social network), 데이터 배포(data publishing), 익명화(anonymity)

• 제1저자 : 성민경 교신저자 : 정연돈
• 투고일 : 2009. 12. 03, 심사일 : 2009. 12. 10, 게재확정일 : 2010. 01. 26.
* 고려대학교 정보통신대학 컴퓨터•전파통신학과 ** 한국과학기술원 전산학과
※ 이 연구에 참여한 연구자의 일부는 2단계 BK21 사업의 지원을 받았음

Ⅰ. 서 론

최근 들어 LinkedIn, Facebook, Twitter 등과 같은 온라인 소셜 네트워크 서비스 제공 사이트들이 늘어나고 있으며, 많은 사용자들이 서비스에 가입하여 온라인 소셜 네트워크 활동을 하고 있다. 이에 따라 소셜 네트워크와 관련된 개인정보데이터의 양은 늘어나며, 이러한 데이터는 마케팅, 인구 통계, 전염병 연구 등 많은 분야에서 유용한 정보를 추출하는데 사용될 수 있다. 그러나 데이터를 배포하고 사용하는 과정에서 민감한 개인정보가 노출되어 개인 프라이버시(privacy) 문제가 생길 수 있다. 예를 들어, 미국 인구의 87%는 성별, 생년월일, 5자리 ZIP 코드 정보를 바탕으로 개인이 유일하게 판별되었다는 연구 결과가 있다[1]. 또한 어떤 개인이 유전병을 가지고 있으면 소셜 네트워크를 이용하여 그의 가족을 알아내어 그의 가족들도 유사한 유전병을 가지고 있다고 높은 확률로 추측할 수 있다.

이러한 상황을 방지하기 위해 소셜 네트워크에서 프라이버시를 보호하기 위한 연구들이 최근 진행되고 있다[2-8]. 소셜 네트워크의 정보는 각 개인 간의 관계를 정점(node)과 이들을 연결하는 간선(edge)으로 표현되는 구조정보와 각 정점과 간선에 대해 나이, 직업, 소속, 관계 등과 같은 정보를 표현하는 내용정보로 나뉜다. 하지만 일부 연구[3, 5]를 제외한 대부분의 연구는 소셜 네트워크에서 프라이버시 보호를 위해 단지 구조정보만을 고려하고 있다.

Campan 등은 구조정보와 정점의 내용정보 모두를 고려한 소셜 네트워크에서의 프라이버시 보호 기법을 제안하였다[3]. 이 기법은 클러스터링을 기반으로 한 방식으로, 프라이버시 보호를 위해 정점이나 간선을 부가적으로 추가하지는 않지만 구조정보의 일부분만을 나타낼 수 있다는 단점이 있다. Zheleva 등은 구조정보와 간선의 내용정보를 고려한 기법을 제안하였다[5]. 이 기법 역시 클러스터링 기반으로 한 방식을 제안하였으며, 정점의 내용정보에 대한 고려가 없다.

본 논문에서는 소셜 네트워크 데이터에서 구조정보와 정점의 내용정보를 동시에 고려한 그래프 수정 기반 방식의 기법을 제안한다. 기존 연구와 우리의 연구의 차별성은 1) 기존 연구[3, 5]는 클러스터링 기반 기법을 이용하여 구조정보의 일부만 나타내어 전체 구조정보가 왜곡되는 것에 반해 우리의 연구는 그래프 수정 기반 기법을 사용하여 전체 구조정보의 왜곡을 최소화 할 수 있으며 2) 기존연구 [5]는 간선의 내용정보만을 고려하였으나 우리의 연구는 보다 다양한 정보를 포함하는 정점의 내용정보를 고려한 것이다.

본 논문에서는 프라이버시 보호를 위해 여러 논문에서 널리 사용된 k-anonymity 개념을 이용한다[9]. 이 개념은 정보를 추측할 때, 후보를 최소 k개 이상 두어 1/k보다 큰 확률을 가지고 추측을 할 수 없게 만드는 것이다. 제안 기법은 내용정보를 일반화하고 간선 추가를 통해 구조 정보를 수정하여 프라이버시 노출 확률을 1/k 이하로 줄이는 것을 목적으로 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구를 살펴본다. 3장에서는 연구배경을 통해 본 논문에서 다루고자 하는 데이터 모델과 공격 모델을 정의할 것이다. 4장에서는 제안기법에 수반되는 정보손실을 측정하고, 알고리즘을 설명한다. 5장은 실험을 통해 제안기법의 성능을 살펴보고, 끝으로 6장에서 본 논문을 요약하고 결론을 내린다.

Ⅱ. 관련 연구

데이터 소유자는 데이터를 배포하기 전, 데이터에 포함된 개인 정보 보호를 위해 데이터를 수정하여 익명화해야 한다. 이때, 데이터 소유자는 다음과 같은 두 가지 요소를 고려하여야 한다. 첫째로 수정된 데이터가 배포 되었을 때 개인의 프라이버시가 보호되어야 하며, 둘째는 원시 데이터와 수정된 데이터의 차이를 최소화 하여 데이터의 유용성을 최대화 하는 것이다.

Sweeney 등은 이름이나 주민등록번호 등의 식별자를 제거하고 데이터를 제공하는 것만으로는 프라이버시 보호에 충분하지 않다고 지적했다[9]. 그들은 k-anonymity 모델을 제시하여 데이터 일반화를 통한 프라이버시 보호 개념을 만들었다.

Machanavajjhala 등은 k-anonymity 모델의 단점인 민감한 데이터의 다양성 부족을 지적하고 이를 방지하기 위한 새로운 모델인 l-diversity를 제안하였다[10]. Liu 등은 t-closeness를 통해 민감한 데이터의 문자적 다양성뿐만 아니라 의미적 다양성도 함께 고려한 모델을 제시하였다[11].

Xiao 등은 정적 데이터(static data)뿐만 아니라 시간이 지남에 따라 지속적으로 변화하는 데이터(dynamic data)에서 프라이버시 보호를 위한 기법인 m-invariance를 제안하였다[12]. 이 기법은 데이터가 추가/삭제되는 상황에서도 프라이버시 보호를 위한 기법으로 데이터 일반화와 모조 데이터(counterfeit) 추가를 통해 프라이버시를 보호하였다. 그러나 이러한 기법들은 그래프로 표현되는 소셜 네트워크에 직접적으로 적용될 수 없다. 때문에 최근 들어 소셜 네트워크 데이터에서 프라이버시 보호를 위한 기법이 많이 연구 되고 있다[1, 3-8].

Backstrom 등은 익명화된 소셜 네트워크 그래프에서 프라이버시 노출 문제를 처음으로 지적하였다[1]. 그래프는 구조적 특성을 가지기 때문에 구조정보를 고려한 기법이 필요하다고 하였으며 능동적 공격(active attack)과 수동적 공격(passive

attack)의 새로운 공격 모델을 제시하였다.

구조정보 중 부분그래프(subgraph)를 공격자가 알고 있을 때 프라이버시를 보호하기 위한 기법으로 Zou 등은 이웃 공격(neighborhood attack)모델을 제시하고 이에 대한 보호 기법을 제안하였다[4]. 그들은 공격자가 특정 정점과 그와 연결된 이웃들로 이루어진 부분 그래프를 알고 있을 때, 전체 그래프에서 해당 부분 그래프와 동형 이성(isomorphism)인 부분 그래프가 최소 k-1개 이상 나타나야 한다는 모델을 제시하였다.

또 하나의 구조정보로 Liu 등은 공격자가 정점의 차수(degree)를 알고 있을 때, 프라이버시 보호를 위한 기법을 제안하였다[6]. 그들은 전체 그래프에서 같은 차수를 가지는 정점이 최소 k개 이상이 되어야 하는 k-degree anonymity를 제시하고 이를 위한 간선 추가 알고리즘을 개발하였다.

Hay 등은 정점의 이웃그래프와 차수를 동시에 고려한 모델을 제안하였다. 그들은 클러스터링 기반 기법을 이용하여 유사한 정점과 간선을 통합하여 개별 정점과 간선의 프라이버시가 드러나지 않게 하였다. 그들은 k-candidate anonymity을 통해 어떤 질의에도 k개의 후보 부분그래프가 나타나게 하는 기법을 제시하였다[7].

Zou 등도 정점의 이웃그래프와 차수를 동시에 고려한 기법을 제안하였다. 그들은 클러스터링 기반의 기법이 너무 많은 데이터의 구조정보 손실을 초래한다고 주장하며 그래프 수정 기반의 프라이버시 보호 기법을 제시하였다[8]. 또한 이 기법은 일시적 데이터뿐만 아니라 지속적으로 변화하는 데이터 까지 처리할 수 있는 특징이 있다.

위에 소개된 관련연구는 소셜 네트워크 데이터의 구조정보만 고려하여 공격자가 구조적 정보만 알고 있을 때의 프라이버시 보호만 할 수 있는 단점이 있다. 하지만 공격자가 소셜 네트워크의 구조적 정보만 가지고 있다는 가정은 현실적이지 않다. 이에 반해 다음 소개되는 관련 연구는 소셜 네트워크의 구조정보뿐만이 아니라 내용정보까지 고려한 기법을 제시하고 있다.

Zheleva 등은 소셜 네트워크에서 각 간선이 가지는 내용정보를 보호하기 위한 기법을 제안하였다[5]. 클러스터링 기반 기법을 사용하여 다섯 가지 간선 내용정보 보호 방법 제시 및 실험을 통해 각 상황에서 프라이버시 노출 정도를 측정하였다.

Campan 등은 각 정점이 가지는 내용정보를 보호하기 위한 클러스터링 기반 기법을 제안하였다[3]. 그들은 유사한 내용정보와 구조정보를 가진 정점들을 하나로 묶어 각 값들을 대표하는 하나의 슈퍼 정점(super node)을 만든 후, 슈퍼 정점 간 관계만을 나타내 원시 데이터의 축약된 구조정보만을 나타내는 방식을 제안하였다.

III. 문제 정의

1. 데이터 모델

소셜 네트워크는 그래프로 나타낼 수 있으며 그래프는 G(V, E)로 나타내어 V는 정점의 집합을, E는 간선의 집합을 각각 뜻한다. 소셜 네트워크에서 각 개인은 하나의 정점에 대응되며, 개인 사이의 관계는 간선으로 나타낸다. 또한 각 정점은 정점을 나타내는 내용정보를 가지고 있다. 예를 들어 이름, 나이, 성별, 직업 등의 내용정보가 정점에 포함되어 있다. 본 논문에서 간선은 내용정보를 가지고 있지 않다고 가정한다. 그러나 이 가정은 제안기법의 간단한 수정을 통해 보완될 수 있다.

정점에 포함된 내용정보는 다음의 세 가지로 분류된다. 첫째로 식별자(identifier : ID)는 개인의 이름이나 주민등록번호 등 개개인을 판별할 수 있는 데이터다.

둘째로 준식별자 (quasi-Identifier : QI) 는 직접적으로 개개인을 판별할 수는 없지만 다른 외부 정보와 합쳐져 개개인을 판별할 수도 있는 정보를 뜻한다. 예를 들어 나이, 성별, 직업 등이 이에 해당한다.

셋째로 민감한 데이터(sensitive data : SD)는 식별자와 유일하게 대응되어 특정 개인에 대한 정보를 드러내면 안 되는 데이터 즉, 타인에게 개인이 가지고 있다고 알려지면 안 되는 데이터이다. 예를 들어 질병명칭, 재산 등이 이에 속한다.

본 제안 기법에서 데이터를 배포할 때, 개인 프라이버시를 보호하며 유용성을 최대로 유지하기 위해 식별자는 삭제되며 준식별자는 최소한으로 수정되고 민감한 데이터는 그대로 남겨진다.

2. 공격 모델

공격자의 수나 그들이 가진 지식을 정확히 알 수 없기 때문에 공격자가 가진 정보를 정확하게 아는 것은 어렵다. 공격자가 가진 정보가 많다고 가정하면 엄격한 프라이버시 보존 기법이 적용되어야 하는 반면 공격자가 가진 정보가 많지 않다고 가정하면 보다 덜 엄격한 프라이버시 보존 기법이 적용될 수 있다. 본 논문에서는 기존 연구들이 했던 방식과 같이 공격자가 특정 정보를 가지고 있다고 가정하고 이에 대한 공격 모델을 설정한다. 본 논문에서 공격자는 소셜 네트워크 데이터 그래프에서 특정 개인의 데이터가 있는지 없는지 알고, 특정 개인의 민감한 정보를 제외한 모든 정보(식별자와 준식별자)를 알며, 그래프에서 특정 개인의 차수를 알고 있다고 가

정한다. 공격자의 정보에 의하여 특정 개인에 대한 민감한 정보를 추측할 때, 1/k 확률보다 큰 확률을 가지고 결정할 수 있을 때 프라이버시가 노출 되었다고 한다.

정의 1 . 공격자의 정보 (adversary's knowledge) 공격자는 공격의 대상이 되는 개인이 배포된 소셜 네트워크 데이터에 포함되어 있는지의 여부 및 개인의 차수와 준식별자 값을 알고 있다. □

정의 2 . 프라이버시 노출 (privacy disclosure) 배포된 소셜 네트워크와 공격자의 정보를 이용하여 1/k 보다 큰 확률로 개인에 대한 민감한 정보를 추측 할 수 있을 때, 프라이버시 노출이 발생하였다고 한다. □

예제 1. 그림 1을 소셜 네트워크 G라고 하자. 각 정점은 개인을 나타내며 식별자 대신 각 정점을 구분하는 번호로 구별되어 있다. 즉, 이름과 같은 식별자는 데이터에 나타나 있지 않다. 각 정점은 내용정보를 포함하고 있으며, 각각 나이, 성별, ZIP, 연봉을 뜻한다. 여기서 나이, 성별, ZIP 는 준식별자이고, 연봉은 민감한 정보라고하자. 공격자가 영희라는 사람이 소셜 네트워크 G에 포함되어 있는 23세의 여성이고 20437의 ZIP을 가졌으며, 소셜 네트워크 상에서 차수가 1이라는 것을 알면 그래프 G 에서 해당 정보와 정점 1이 유일하게 대응되어, 정점 1에 대응하는 것이 영희라는 것을 알게 된다. 비록 영희에 대한 식별자가 그래프 G에 나타나 있지 않더라도 영희의 연봉이 2000만원 이라는 것을 알게 되어 개인 프라이버시가 노출된다. □

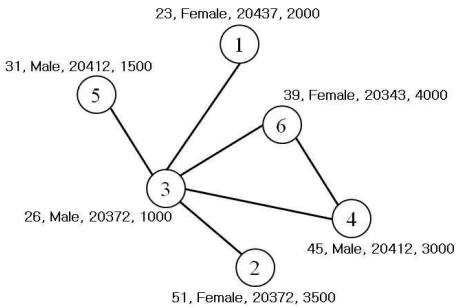

그림 1. 소셜 네트워크 그래프 G
Fig. 1. Social Network Graph G

이와 같이 내용정보와 구조정보가 같이 표현된 소셜 네트워크 그래프 G 자체로는 프라이버시가 드러나지 않더라도 공격자가 가진 정보에 의해 프라이버시 노출이 가능하다. 데이터 제공자는 공격자가 자신의 정보를 가지고 특정인의 민감한 정보를 추측했을 때, 1/k보다 큰 확률로 추측할 수 없도록 데이터를 수정한 후 데이터를 제공해야 한다.

IV. 제안 기법

본 장에서는 3.2절에서 설명한 공격 모델을 방어하기 위한 기법을 제안한다. 제안 기법은 공격 모델을 방어하기 위해 동일한 준식별자(내용정보)와 동일한 차수(구조정보)를 갖는 정점이 소셜 네트워크 그래프에서 최소한 k개 이상 나타나도록 원시 데이터를 수정한다. 원시 데이터 수정 과정에서 필연적으로 발생하는 정보 손실에 대해 우선 설명하고, 정보 손실을 최소화하는 알고리즘과 예제를 알아본다.

1. 정보손실

내용정보를 포함한 소셜 네트워크 데이터에서 정보손실은 내용정보 손실과 구조정보 손실로 구분된다. 내용정보 손실은 나이, 직업과 같은 준식별자를 일반화하여 수정하는 과정에서 발생하는 손실이며, 구조정보 손실은 소셜 네트워크 그래프에서 같은 차수를 가진 정점을 만들기 위해 간선을 수정하는 과정에서 생기는 손실이다. 구조정보 수정을 위해 정점을 추가/삭제, 간선을 추가/삭제 할 수 있지만, 본 논문에서는 구조정보 수정 시 간선을 추가하는 상황만 고려한다. 직관적으로 이 두 가지 손실은 같은 범주에서 생각하기 어려우므로 이를 함께 고려하여 계산할 수 있는 측정법을 본 절 후반부에 소개한다.

같은 내용정보와 구조정보를 갖는 정점을 k개 이상 만들기 위해서는 k개 이상의 정점을 한 그룹으로 묶어 해당 그룹의 준식별자와 차수를 각각 같게 만들면 된다. 이때 그래프상의 모든 정점을 한 그룹으로 묶어 준식별자와 차수를 각각 같게 만든다면 프라이버시 보호는 되지만 정보 유용성은 매우 떨어진다. 즉, k개 이상의 정점을 한 그룹으로 묶을 때 해당 그룹에 속하는 정점의 수는 k이상의 최소 정수가 되어 수정되는 정보의 양을 최대한 줄이는 것이 좋다. 또한 한 그룹에 속하는 정점들을 최대한 유사한 내용정보와 구조정보를 지닌 정점끼리 묶으면 수정되는 정보가 적어져 정보 손실의 양을 최소화할 수 있다. 정점의 그룹화에 따르는 내용정보의 손실 측정과 구조정보의 손실 측정은 아래와 같다.

내용정보 손실 측정을 위해 본 논문에서는 [13]에서 사용했던 측정법을 이용한다. 내용정보는 두 가지로 구성된다. 숫자 데이터(numerical data)와 계층적 데이터(hierarchical data). 숫자 데이터는 나이와 같이 데이터의 범위가 연속된 숫자에 한정된 데이터이다. 계층적 데이터는 직업, 국가와 같이 상위에는 추상적인 개념이 들어가고 하위로 갈수록 구체적 개념이 들어가는 데이터이다. 계층 데이터는 의미 분류 트리를 사용

하여 표현될 수 있다. 그림 2 는 ZIP과 성별에 대한 의미 분류 트리의 예시이다.

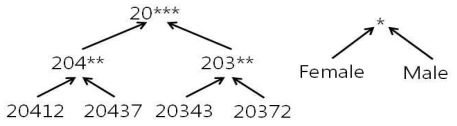

그림 2. 의미 분류 트리
Fig. 2. Taxonomy Tree

그룹화를 했을 때 각 그룹에서 발생하는 숫자 데이터의 손실과 계층적 데이터의 손실을 모두 합한 것을 내용정보의 손실이라 한다. 그룹화를 했을 때 연속된 숫자 정보는 각 그룹에서 가장 작은 숫자와 가장 큰 숫자의 범위로 수정되며, 계층적 정보는 각 데이터의 최소공통조상으로 수정된다.

정의 3. 내용정보 손실(TLC : Total Loss of Content information) 정점들을 통합 하여 그룹에 속한 정점들의 준식별자를 같은 값으로 만들 때, 원시정보와 변화된 정보의 차이를 내용정보 손실이라 하며 다음과 같이 정의된다.

$$TLC(G, G_T) = \frac{1}{|QI|}\sum_{h=1}^{q} CL(p_h)$$

G는 원시 그래프, GT는 수정된 그래프, $|QI|$는 준식별자 속성(attribute)의 개수, p_h는 각 그룹, q는 그룹의 개수, $CL(p_h)$는 그룹 p_h에서 정보손실을 각각 뜻한다. 그룹 p_h 에서 내용정보 손실(Content information Loss of the group ph) $CL(p_h)$는

$$CL(p_h) = |p_h| \cdot \left(\sum_{i=1}^{m}\frac{MAX_{N_i} - MIN_{N_i}}{|N_i|} + \sum_{j=1}^{n}\frac{H(\wedge(\cup_{C_j}))}{H(T_{C_j})}\right)$$

이다. 여기서 준식별자는 숫자 데이터의 속성 $N_1, N_2, ..., N_m$과 계층적 데이터의 속성 $C_1, C_2, ..., C_n$으로 구성된다. $|p_h|$는 그룹 p_h 에 속한 정점의 개수, MAX_{N_i}, MIN_{N_i}는 그룹 p_h에 속한 정점 중 속성 N_i가 가장 큰 값과 작은 값을 각각 뜻한다. m은 숫자 데이터 속성의 개수를, $|N_i|$는 N_i 속성의 크기를 뜻한다. T_{C_j}는 계층 데이터 C_j의 의미 분류 트리(taxonomy tree)이며, $\wedge(\cup_{C_j})$는 그룹 p_h에 속한 정점들의 속성C_j에 대한 계층 데이터의 최소공통조상(least common ancestor)이다. n은 계층 데이터 속성의 개수를 뜻한다. 즉, |QI|=m+n 이다. $H(T_{C_j})$는 의미 분류 트리 T_{C_j}의 높이이다. □

예제 2. 그림 1에서 정점ID {1,5}, {3,6}, {2,4}가 각각 그룹을 형성하고 차례대로 그룹 p_1, p_2, p_3 이라고 각각 했을 때, 내용정보 손실은 다음과 같다. □

$CL(p_1)$	$\|2\| \cdot \left(\frac{31-23}{51-23}+\frac{0}{1}+\frac{1}{2}\right)=\frac{22}{14}$	$TCL=(G, G_T)=\frac{1}{3}\cdot\left(\frac{22}{14}+\frac{55}{14}+\frac{62}{14}\right)=\frac{139}{42}$
$CL(p_2)$	$\|2\| \cdot \left(\frac{39-26}{51-23}+\frac{1}{1}+\frac{1}{2}\right)=\frac{55}{14}$	
$CL(p_3)$	$\|2\| \cdot \left(\frac{51-45}{51-23}+\frac{1}{1}+\frac{2}{2}\right)=\frac{62}{14}$	

구조정보 손실 측정을 위해서는 추가된 간선의 개수를 측정한다. 즉, 간선의 추가가 많을수록 구조정보 손실이 크다고 하며 간선의 추가가 적을수록 구조정보 손실이 적다고 한다.

정의 4. 구조정보 손실(TLS : Total Loss of Structure information) 정점들을 그룹화 하여 각 그룹에 속한 정점들의 차수를 같은 값으로 만들 때, 원시정보와 변화된 정보의 차이를 구조정보 손실이라 하며 다음과 같이 정의된다.

$$TLS(G, G_T) = \sum_{p=1}^{q} SL(p)$$

그룹 p에서 구조정보 손실(Structure information Loss of the group p) $SL(p)$는

$$SL(p) = \sum_{i=1}^{|p|}\frac{MAX_D - D_{p_i}}{MAX_D - MIN_D}$$

으로 MAX_D는 그룹 p에서 가장 차수가 큰 정점의 차수이고 MIN_D는 그룹 p에서 가장 차수가 작은 정점의 차수이며, D_{p_i}는 그룹 p 에서의 각 정점의 차수들을 뜻한다. 즉, 한 그룹에서 모든 정점의 차수를 가장 차수가 큰 정점의 차수와 같게 만드는데 추가되는 차수의 수를 측정한다. □

예제 3. 그림 1에서 정점ID {1,5}, {3,6}, {2,4}가 각각 그룹을 형성하고 차례대로 그룹 1, 2, 3 이라고 각각 했을 때, 구조 정보 손실은 다음과 같다. □

$SL(1)$	$0+0=0$	$TLS(G, G_T)=\left(0+\frac{3}{3}+\frac{1}{1}\right)=2$
$SL(2)$	$\frac{0}{5-2}+\frac{3}{5-2}=\frac{3}{3}$	
$SL(3)$	$\frac{0}{2-1}+\frac{1}{2-1}=\frac{1}{1}$	

내용정보 손실과 구조정보 손실을 이용하여 정보 손실의 양을 측정한다. 내용정보와 구조정보 관계에 관한 여러 정보 손실 측정법이 있지만 본 논문에서는 두 손실의 합으로 정보 손실을 측정한다.

정의 5. 정보 손실(TL : Total Loss of information) 배포된 데이터의 프라이버시 보호를 위해 정점들을 그룹화 하여 그룹에 속한 정점들의 준식별자와 차수를 같은 값으로 만들 때 , 원시정보와 변화된 정보의 차이를 정보 손실이라 하며 다음과 같이 정의된다. (γ은 가중치 값 이며 $(0 \le \gamma \le 1)$)

$$TL(G, G_T) = \gamma \cdot TLS(G, G_T) + (1-\gamma) \cdot TLC(G, G_T)$$

□

예를 들어, 그림 1에서 정점ID {1,5}, {3,6}, {2,4}가 각각 그룹을 형성하고, $\gamma = 0.5$, k=2 라고 하면 TL은 다음과 같다.

$$TL(G, G_T) = 0.5 \cdot 2 + (1-0.5) \cdot \frac{139}{42} = \frac{223}{84}$$

2. 알고리즘

제안된 알고리즘은 프라이버시 보호를 위해 원시그래프데이터와 프라이버시 강도를 나타내는 k, 구조정보와 내용정보의 상대적 중요도를 나타내는 γ를 이용하여 수정된 그래프를 만든다. 4.1절에서 언급한 바와 같이 유사한 정점들을 그룹화하기 위해, [13]에서 제안했던 k-member clustering 기법을 이용하지만 우리의 알고리즘은 [13] 에서는 고려하지 않은 구조정보까지 고려한다. 알고리즘은 다섯 단계로 이루어진다.

단계 1: 소셜 네트워크 그래프를 테이블 데이터로 변환한다.
단계 2: 정보 손실 수식을 고려하여 각 정점들을 그룹화 한다.
단계 3: 정보가 수정된 테이블을 만든다.
단계 4: 실현 가능한 그래프를 만들기 위해 차수를 수정한다.
단계 5: 수정된 소셜 네트워크 그래프를 생성한다.

알고리즘에 대한 예제는 다음과 같다.

첫 번째로 단계 1을 통해 소셜 네트워크 그래프를 각 정점이 하나의 레코드 (record)로 대응되는 테이블 데이터로 만든다. 본 단계의 목적은 그룹화 과정을 손쉽게 하기 위해서다. 각 정점의 준식별자, 민감한 데이터, 차수가 각각 속성을 이룬다. 그림 1의 소셜 네트워크 그래프에 대해 표1은 각 속성 값과 차수를 나타낸 테이블이다.

표 1. 그림1의 그래프 G에 대한 테이블 데이터
Table 1. Table Data of Graph G in Figure 1

ID	나이	성별	ZIP	연봉	차수
1	23	Female	20437	2000	1
2	51	Female	20372	3500	1
3	26	Male	20372	1000	5
4	45	Male	20412	3000	2
5	31	Female	20412	1500	1
6	39	Male	20343	4000	2

테이블이 생성되면 단계 2를 통해 그룹화 작업을 한다. 우선 테이블에서 가장 차수가 큰 정점(레코드)를 뽑아서 그룹을 만들고 그 정점은 테이블에서 삭제한다. 남은 정점 중 처음에 뽑힌 정점과 가장 유사한 즉, 정보 손실 값이 가장 작은 정점을 뽑아서 같은 그룹에 넣고 테이블에서 삭제한다. 뽑힌 두 정점의 준식별자에 따라 각 속성별로 평균값을 구하고 그 값은 해당 그룹의 대표 값이 된다. 다음 정점을 선택할 때, 그룹의 대표 값과 비교해서 정보 손실 값이 가장 작은 정점을 선택한다. 이 과정을 한 그룹이 k개의 정점을 가질 때까지 수행한다. 한 그룹이 k개의 정점을 가지면 새로운 정점으로 새로운 그룹을 만든다. 이러한 그룹 생성 과정을 테이블에서 k개 미만의 정점이 남을 때까지 수행한다. k개 미만의 정점으로는 새로운 그룹을 생성할 수 없으므로 각 정점은 정보 손실을 최소화 하는 그룹에 각각 포함된다. 예를 들어 표 1에서 k=2, γ=0.5라고 하면 {3,6}, {1,5}, {2,4} 로 각각 그룹이 만들어 진다.

단계3은 정보가 수정된 테이블을 만든다. 같은 그룹에 속한 정점들은 모두 같은 준식별자와 차수를 가져야 한다. 숫자 데이터는 해당 그룹에서 가장 작은 값과 큰 값의 범위로 일반화 된다. 계층적 데이터는 최소공통조상 값으로 일반화되며, 차수 데이터는 해당 그룹에서 가장 큰 차수 값과 모두 같게 수정된다. 표2는 단계3에 대한 결과 값을 보여준다.

표 2. 단계3에 대한 결과
Table 2. Result of Step 3

ID	나이	성별	ZIP	연봉	차수
1	23-31	Female	204**	2000	1
2	45-51	*	20***	3500	2
3	26-39	Male	203**	1000	5
4	45-51	*	20***	3000	2
5	23-31	Female	204**	1500	1
6	26-39	Male	203**	4000	5

단계 4에서는 단계3에서 만들어진 데이터가 실현 가능한(realizable)소셜 네트워크 그래프를 생성하지 못 할 수 있으므로 실현 가능한 소셜 네트워크 그래프로 만들기 위한 작업을 수행한다. 실현 가능한 그래프가 되기 위해서는 우선 모든 정점의 차수의 합이 짝수이어야 한다. 소셜 네트워크 그래프에서 각 간선은 두 정점에 연결되어 있으므로 차수를 구할 때 간선은 두 번씩 세어진다. 모든 수의 곱하기 2는 짝수가 되어야 하므로 모든 정점의 차수의 합은 짝수가 되어야 한다. 만약 짝수가 아니라면 차수의 합이 홀수인 그룹의 각 정점에 1씩 더해서 전체 합이 짝수가 되게 하여야 한다. 이 때, 두 개 이상의 그룹의 차수의 합이 홀수라면 그 합이 가장 작은 그룹의 정점에 1씩 더한다. 차수의 합이 짝수가 되면 Erdös 등 이 제시한 기법을 이용하여 실현 가능한 그래프인지 판별 한다[14]. Erdös 등은 주어진 내림차순 정렬된 양수의 차수 집합에서 실현 가능한 그래프 생성 판별을 위한 필요충분조건을 제시하였

다. 그래프가 실현 가능하지 않다고 판별되면 차수의 합이 가장 작은 그룹의 정점에 차수를 1씩 더해서 다시 모든 정점의 차수의 합이 짝수 인지와 실현 가능한 그래프인지 판별한다. 이 과정을 그래프가 실현 가능하다고 판별될 때까지 반복한다. 표3은 단계4에 대한 결과이다.

표 3. 단계4에 대한 결과
Table 3. Result of Step 4

ID	나이	성별	ZIP	연봉	차수
1	23-31	Female	204**	2000	2
2	45-51	*	20***	3500	2
3	26-39	Male	203**	1000	5
4	45-51	*	20***	3000	2
5	23-31	Female	204**	1500	2
6	26-39	Male	203**	4000	5

마지막으로 표3과 원시 그래프 G 를 이용하여 프라이버시가 보호되는 수정된 그래프를 만든다. 원시 그래프의 차수 집합 So 와 단계4에 의한 결과 차수 집합 Sr을 구할 수 있다. So와 Sr은 각각 ID 순서에 의해 정렬이 되어 있다. 그러면 우리는 Sr에서 So를 뺀 새로운 차수 집합 Ss를 구할 수 있다. 즉, Ss는 원시 그래프에 비해 새롭게 추가된 차수를 나타내며 이 차수들에 맞게 간선을 추가하면 된다. 우선 Ss에서 가장 큰 원소를 찾고 그 원소를 제외하고 두 번째로 큰 원소를 찾는다. 원시 그래프를 참고하여 두 원소 사이에 간선이 없으면 간선을 추가하고, 간선이 있으면 다음으로 큰 원소와 비교해서 간선 추가 여부를 결정한다. 간선을 추가할 때는 두 정점에 대응되는 Ss의 원소에서 각각 1씩 빼서 간선이 추가되었음을 확인한다. 이 과정을 모든 원소가 0이 될 때까지 수행한다. 그림 3은 알고리즘의 결과로 만들어진 프라이버시가 보호되는 소셜 네트워크 그래프를 나타낸다. 그림 3에서 공격자가 특정 개인의 대한 정보를 가지고 있다 하더라도 최소 2개의 후보 정점이 생겨 민감한 정보에 대해 정확한 예측을 할 수 없다.

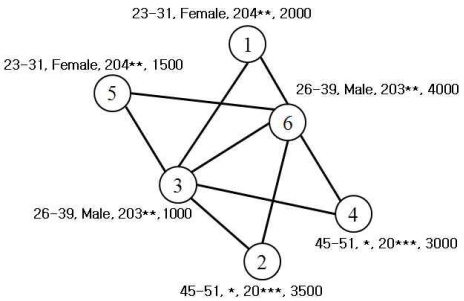

그림 3. 수정된 소셜 네트워크 그래프 GT
Fig. 3. Modified Social Network Graph GT

V. 성능 평가

본 장에서는 다양한 실험에 의한 성능평가 및 분석을 통해 제안 기법의 타당성을 검증하고자 한다. 이를 위해 UC Irvine Machine Learning Repository로부터 실제 성인 데이터를 추출하여 실험을 수행하였다[15]. 모든 실험에서 준식별자로 나이, 성별, 인종, 결혼 상태, 국가의 5가지 속성을 고려하였다. 나이는 연속된 숫자 데이터며, 성별, 인종, 결혼 상태, 국가는 계층적 데이터다. 성별, 인종, 결혼유무는 예/아니오 의 1계층이며, 국가에 대한 계층 정보는 그림 4와 같다. 또한 [15]에는 내용정보만 있고 구조정보는 없기 때문에 각 개인을 정점으로 하여 그래프를 구성해야 하였다. R.Barabasi에 따르면 소셜 네트워크는 척도 없는 그래프 (scale-free graph)에 따른다고 하였으며 척도 없는 그래프는 멱함수 법칙(power law)을 따른다[16]. 이에 따라 멱함수분포에 따르는 그래프를 생성하여 실험을 수행하였다. 다양한 정점의 수에 따른 실험을 통해 확장성에 관한 평가를 하였으며 전체 그래프에서 간선의 수 변화를 통해 밀도가 높은 그래프(정점 당 간선의 수가 많은)와 낮은 그래프(정점 당 간선의 수가 적은)에서 차이점을 살펴보았다. 정점의 수는 1000개, 2000개, 5000개, 10000개로 변화 시켜 보았으며 간선의 수는 총 정점의 수의 10배, 20배, 50배에 대하여 각각 실험하였다. 프라이버시 강도를 나타내는 실험인자 k의 값은 2, 5, 10, 25, 50 으로 변화시켜 가면서 정보손실의 정도를 측정하였다. 구조정보와 내용정보의 중요도를 결정하는 실험인자 γ은 0.5로 고정 시켰다. 정점의 수, 간선의 수, 실험인자 k의 변화를 통해 실험을 수행한 결과는 그림 5 -10과 같다.

그림 5와 6은 정점의 수가 증가할 때 정보손실의 양을 측정한 그래프이다. 정보손실은 각 정점 당 발생하는 내용정보 손실과 구조정보 손실의 합으로 표준화 되었다. 그래프에서 보듯이 정점의 수가 늘어날수록 각 정점에서 발생하는 정보손실의 양은 줄어든다. 이것은 한정된 도메인에서 정점의 수가 늘어날수록 유사한 정보를 가진 정점이 늘어나게 되어 더욱 유사한 정점끼리 그룹을 형성할 수 있기 때문이다. 또한 프라이버시의 강도를 나타내는 k가 커질수록 정보손실은 더욱 크게 나타났다. 이것은 k가 커지면 더욱 많은 정점들이 한 그룹을 형성해야 하므로 정보 수정의 양이 늘어나게 되면서 생기게 된 결과라 할 수 있다. 그러나 정점의 수가 늘어날수록 k값에 따른 정보손실의 양의 차이는 줄어들었다. 이와 같이 정점의 수가 늘어남에 따라 정보손실의 양이 줄어들게 되므로 본 제안기법은 정점의 수에 대해 확장성을 가지고 있다고 할 수 있다.

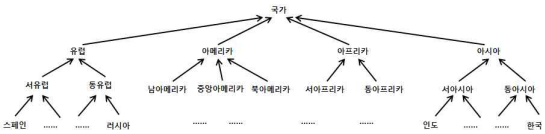

그림 4. 국가 정보에 대한 의미 분류 트리
Fig. 4. Taxonomy Tree of Country Information

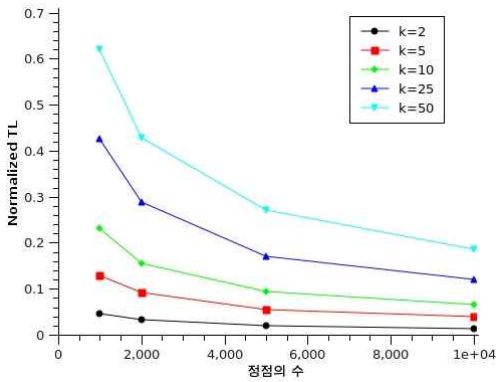

그림 5. 총 간선의 수 = 총 정점의 수의 10배
Fig. 5. Total Number of Edges = Total Number of Nodes x 10

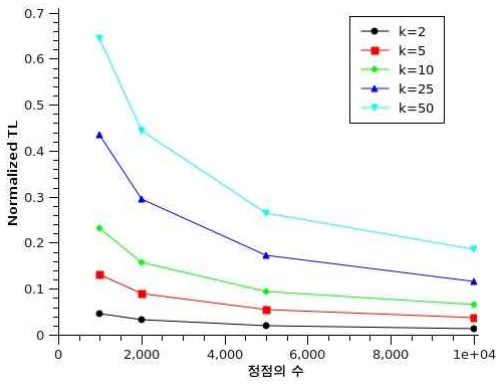

그림 6. 총 간선의 수 = 총 정점의 수의 25배
Fig. 6. Total Number of Edges = Total Number of Nodes x 25

그림 7과 8은 k의 증가에 따른 표준화된 정보손실그래프를 나타내고 있다. 그래프에서 보듯이 k가 커질수록 정보손실의 양도 늘어나고 있다. 앞에서 설명했듯 이것은 프라이버시의 강도가 강해질수록 정보손실의 양이 늘어난 것이다. 또한 정점의 수가 많으면 같은 k에서 정보손실의 양은 더 적은 것을 볼 수 있다. k=2 일 때는 정점의 수에 따른 정보손실의 차이가 크지 않으나 k가 커짐에 따라 정점의 수에 따른 정보손실의 차이가 점점 증가함을 볼 수 있다. 이것은 정점의 수가 많지 않을 때, k를 크게 설정하면 정보손실의 양이 늘어남을 보여준다. 즉, 소셜 네트워크 그래프에 표현된 개인들이 많지 않을 때 높은 프라이버시 보호 조건을 제약하면 정보손실이 더욱 늘어남을 뜻한다. 이와 같이 소셜 네트워크 프라이버시 보호를 위해 k를 설정할 때, 소셜 네트워크에 속한 총 정점의 수를 고려해서 적절한 k를 설정하는 것이 필요하다.

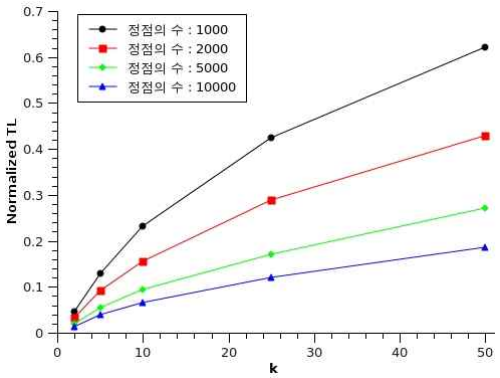

그림 7. 총 간선의 수 = 총 정점의 수의 10배
Fig. 7. Total Number of Edges = Total Number of Nodes x 10

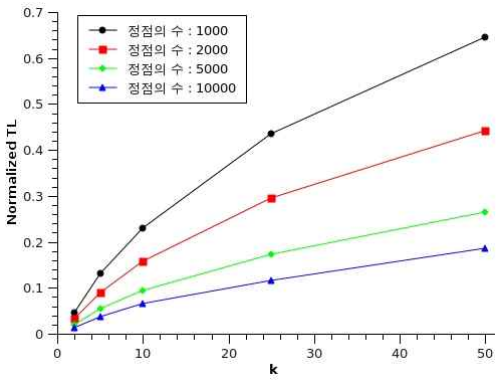

그림 8. 총 간선의 수 = 총 정점의 수의 25배
Fig. 8. Total Number of Edges = Total Number of Nodes x 25

직관적으로, 간선의 수 변화는 내용정보에는 영향을 미치지 않고 구조정보에만 영향을 미칠 것이다. 그림 9, 10은 소셜 네트워크 그래프에 있는 총 정점의 수에 대해 그래프에 있는 총 간선의 수의 변화에 따른 구조정보 손실을 보여주고 있다. k가 증가함에 따라 구조정보의 손실은 증가하지만 간선의 수 증가에 의한 구조정보의 손실은 증가하지 않는다. 즉, 소셜 네트워크 그래프에서 간선의 밀도가 높고 낮음에 관계없이 구조정보 손실은 일정하다. 이것은 본 실험에서 그래프 생성 시 멱함수 분포에 따라 정점이 간선을 가지게 한 결과로 볼 수 있

다. 멱함수 분포에 따라 그래프 생성 시 간선의 밀도에 관계없이 높은 차수를 가지는 소수의 정점과 다수의 낮은 차수를 가지는 정점으로 구성된 그래프를 만들게 되므로 간선의 수는 구조정보에 큰 영향을 미치지 않는다는 결론을 낼 수 있다.

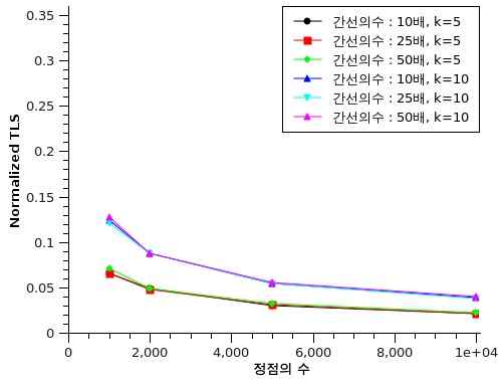

그림 9. k=5, k=10
Fig. 9. k=5, k=10

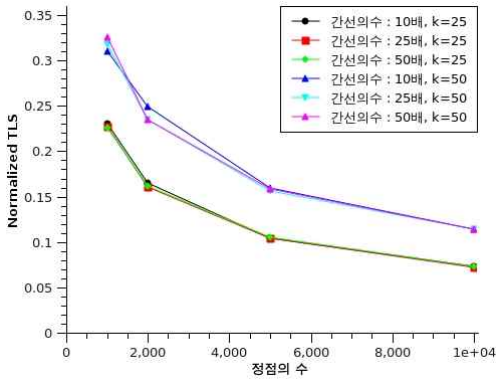

그림 10. k=25, k=50
Fig. 10. k=25, k=50

본 실험의 결과로 정점의 수가 많아질수록 각 정점에 대한 정보손실의 양은 줄어들며, k가 커질수록 프라이버시 강도가 강해져서 정보손실의 양은 늘어나며, 그래프의 총 간선의 수는 정보손실에 거의 영향을 주지 않음을 알 수 있었다.

VI. 결론

본 논문에서는 기존의 소셜 네트워크 프라이버시 보호 기법이 다루지 못한 새로운 공격형태에 대한 프라이버시 보호 기법을 제시하였다. 기존의 기법은 구조정보를 가진 소셜 네트워크에 적합하지 않거나 소셜 네트워크에서 중요한 정보인 내용정보를 다루지 않았다. 또한 내용정보까지 다룬 기존기법은 너무 많은 구조정보 왜곡을 초래하였다.

본 논문에서는 소셜 네트워크에서 구조정보와 내용정보의 유용성을 최대화하면서 프라이버시를 보호하는 효과적인 기법을 제안하였다. 제안기법을 통해 프라이버시 노출의 문제없이 소셜 네트워크 그래프에서 중요한 정보인 구조정보의 이용과 동시에 그 안에 포함된 내용정보까지 이용할 수 있게 되었다.

본 논문에서 제안한 기법은 구조정보에 대한 공격자의 정보가 한정되어 있다는 단점이 있다. 본 제안기법은 공격자가 구조정보 중 정점의 차수만 알고 있다고 가정하였는데 부분그래프 또한 구조정보에서 널리 이용되는 정보이다. 이를 보완하기 위해 향후 연구에는 공격자의 정보에 정점의 차수 뿐 아니라 부분그래프의 구조정보까지 포함되어 있다고 가정하고 프라이버시를 보호하는 기법에 관한 연구수행을 할 것이며, 소셜 네트워크 데이터가 변하는 상황 고려 및 정보손실을 더욱 줄일 수 있는 기법 등도 함께 수행할 계획이다.

참고문헌

[1] L. Backstrom, C. Dwork, and J. Kleinberg, “Wherefore art thou R3579X? anonymized social networks, hidden patterns, and structural steganography” in WWW’07, 2007.

[2] A. Korolova, R. Motwani, and S. U. Nabar, “Link privacy in social networks,” in CIKM’08, 2008.

[3] A. Campan, and T. M. Truta, “A clustering approach for data and structural anonymity in social networks,” in PinKDD’08, 2008.

[4] B. Zou, and J. Pei, “Preserving privacy in social networks against neighbourhood attacks,” in ICDE’08, 2008.

[5] E. Zheleva, and L. Getoor, “Preserving the privacy of sensitive relationships in graph data,” in PinKDD’07, 2007.

[6] K. Liu, and E. Terzi, “Towards identity anonymization on graphs,” in SIGMOD’08, 2008.

[7] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, “Resisting structural re-identification in anonymized social networks,” in PVLDB’08, 2008.

[8] L. Zou, L. Chen, and M. T. Ozsu, “K-automorphism: A general framework for privacy preserving network publication,” in VLDB’09, 2009.

[9] L. Sweeney, “K-anonymity: A model for protecting privacy,” International Journal on Uncertainty, Fuzziness and Knowledge-based System, 2002.
[10] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity,” in ICDE, 2006.
[11] N. Li, T. Li, and S. Venkatasubramanian, “t-Closeness: Privacy beyond k-anonymityand l-diversity,” in ICDE, 2007.
[12] X. Xiao, and Y. Tao, “m-Invariance: Towards privacy preserving re-publication of dynamic datasets,” in SIGMOD'07, 2007.
[13] J. W. Byun, A. Kamra, E. Bertino, and N. Li, “Efficient k-anonymization using clustering techniques,” in DASFAA, 2007.
[14] P. Erdos, and T. Gallai, “Graphs with prescribed degrees of vertices, Mat.Lapok, 1960.
[15] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI Repository of Machine Learning Databases. *http://archive.ics.uci.edu/ml/*
[16] R. Barabasi, "Linked; The new science of networks," 2002.
[17] 김선태, “내부 정보유출 방지,” 한국컴퓨터정보학회지, 제 16권, 제 1호, 37-51쪽, 2008년. 6월.
[18] 박태희, “데이터베이스 암호화 정책,” 한국컴퓨터정보학회지, 제 16권, 제 1호, 61-72쪽, 2008년 6월.

저 자 소 개

성 민 경

2009 : 고려대학교 정보통신대학 컴퓨터과학 학사
현재 : 고려대학교 정보통신대학 컴퓨터전파통신공학과 석사 재학 중
관심분야 : 데이터 프라이버시, 대용량 데이터 처리

이 기 용

1998 : 한국과학기술원 전산학과 학사.
2000 : 한국과학기술원 전산학과 석사.
2006 : 한국과학기술원 전자전산학과 전산학전공 박사.
2008 : 삼성전자 기술총괄 소프트웨어 연구소 책임연구원.
현재 : 한국과학기술원 전산학전공 연구조교수
관심분야 : 데이터베이스, 임베디드 DBMS, 데이터 웨어하우스, OLAP

정 연 돈

1994 : 고려대학교 전산과학과 학사.
1996 : 한국과학기술원 전산학과 석사.
2000 : 한국과학기술원 전산학과 박사.
2000 : 한국과학기술원 정보전자연구소 Post-Doc. 연구원.
2001 : 한국과학기술원 전자전산학과 전산학전공 연구교수.
2003 : 동국대학교 컴퓨터공학과 교수
현재 : 고려대학교 컴퓨터·통신공학부 교수
관심분야 : 모바일/브로드캐스트 데이터베이스, 센서 네트워크, 공간 데이터베이스, 데이터 프라이버시 등