Unintentional and Involuntary Personal Information Leakage on Facebook from User Interactions

Po-Ching Lin[1] and Pei-Ying Lin[2]
[1]Department of Computer Science and Information Engineering,
National Chung Cheng University
Chiayi, Taiwan 621
[e-mail: pclin@cs.ccu.edu.tw]
[2]HTC Corporation
Taoyuan, Taiwan 330
[e-mail: easter6236@gmail.com]
*Corresponding author: Po-Ching Lin

Received November 20, 2015; revised May 5, 2016; accepted June 4, 2016; published July 31, 2016

Abstract

Online social networks (OSNs) have changed the way people communicate with each other. An OSN usually encourages the participants to provide personal information such as real names, birthdays and educational background to look for and establish friendships among them. Some users are unwilling to reveal personal information on their personal pages due to potential privacy concerns, but their friends may inadvertently reveal that. In this work, we investigate the possibility of leaking personal information on Facebook *in an unintentional and involuntary manner*. The revealed information may be useful to malicious users for social engineering and spear phishing. We design the inference methods to find birthdays and educational background of Facebook users based on *the interactions among friends* on Facebook pages and groups, and also leverage *J-measure* to find the inference rules. The inference improves the finding rate of birthdays from 71.2% to 87.0% with the accuracy of 92.0%, and that of educational background from 75.2% to 91.7% with the accuracy of 86.3%. We also suggest the sanitization strategies to avoid the private information leakage.

Keywords: Online social networks, Facebook, user interactions, privacy leakage, J-measure

http://dx.doi.org/10.3837/tiis.2016.07.024 **ISSN : 1976-7277**

1. Introduction

Since the rise of online social networks (OSNs), people are used to sharing their daily life with friends on websites. It is common to see millions of registered users or even more on a popular OSN. A well known example is Facebook, which reached an achievement with 1 billion users in a single day on August 24, 2015 [1]. The *Like* and *Share* buttons are viewed over 22 billion times daily across more than 7.5 million websites [2]. A recent study revealed that 30% of Americans get their news on Facebook [3]. In Facebook, one can post photos, messages and links on one's wall, and friends can comment on or even just "like" a post.

Facebook was criticized for being indifferent to user privacy in the past decade [4]. Disclosing personal information is a double-edged sword. When it comes to OSNs, information exposure is usually a plus or even a must for users to join a new community [5]. An OSN usually encourages users to expose personal information because self-information disclosure can build trust, strengthen the ties between people, and bind romantic relationships or friendships [6,7]. The usage of OSNs usually raises serious concern about the overall privacy, and it is essential to protect personal identifiable information (PII) [8], which alone or combined with other public information, can be used to distinguish or trace an individual's identity [9]. Furthermore, a user's personal information may be also leaked by his or her friends unintentionally and involuntarily during their interactions or information sharing. *The unintentional and involuntary information leakage from user interactions is relatively inconspicuous and uncontrollable*, but increases the chances of successful targeted attacks or spear phishing by social engineering. For example, an attacker will know a person better from an OSN and easily impersonate a familiar person to cheat him or her.

In this work, we study the possibility of inferring personal information that users do not intend to provide *from user interactions* (i.e., unintentional and involuntary information leakage). Related studies include measuring privacy risks on a single OSN [10-12] and elucidating online social footprints on multiple OSNs [9,13-17]. Some researchers studied how to avoid revealing private information in blogs [5,18], and some used graph models to protect sensitive labels in OSNs [19-21]. Most prior studies were based on known facts to judge what happened before, i.e., a posteriori. For example, if a user *likes* something on a Facebook page, we can infer that the user is interested in a specific issue on that page. In another example, if a user does not reveal a private attribute, one may infer it from the groups to which the user belongs or by selecting the most popular value of this attribute among the user's friends. In this work, we want to know a user's previous experiences or facts to decide what the likely result or effect of something will be, i.e., *a priori*. In other words, this work suggests an extended logical process, which involves more than one attribute, and it is able to deduce further information from the logical overlap between the given data.

We evaluate the inference of educational background and birthday as the personal information. The user interactions are retrieved by the Facebook API instead of a crawling program because it gets user information such as the posts on the walls with user permission, whereas a crawling program gets only publicly available data. The experiments show that even though the Facebook users do not provide the two attributes in public, it is still possible to infer the attributes. This work also suggests an ontology to recognize synonyms and adopts J-measure to evaluate the highest probability of users' educational background. We consider not only the most popular value of each attribute like the previous work in [11], but also the attributes of gender and hometown of the users' friends. Intuitively, inference depending

solely on the most popular value may be imprecise, and this work involves more attributes for the inference.

The main contributions of this paper are as follows:

1. We infer user privacy based on the *interactions between users instead of just the user profiles* because the former is likely to give rise to unintentional and involuntary personal information leakage. We use the Facebook Query Language (FQL) to obtain the user interactions, including the *posts* and *comments* on a user's wall. The Facebook API allows us to collect personal information as much as possible when compared with a crawler.
2. We analyze the *like pages* of users with Wikipedia as an ontology to find out the terms (e.g., school names) of the same meaning. We also use J-measure to evaluate the quality of possible inference rules to correctly find unintentional and involuntary personal information. The educational background and other attributes such as living place and gender are also evaluated together to improve the results.

The remainder of this paper is organized as follows. In Section 2, we review related work and the background of this work. In Section 3, we describe the inference methodology based on the interactions between users rather than just on the user profiles, which are used for the inference in almost all of the prior studies. The data sets for evaluation and validations are described in Section 4. We conclude this work and present future work in Section 5.

2. Related Work and Background

In this section, we will review related work and introduce the theories used in this work.

2.1 Related work

The related work covers the studies dedicated to privacy protection and information leakage on OSNs. We emphasize their limitations in the review below to justify the significance of this work.

Several studies [9, 13-16] presented the methods to grasp more privacy across multiple OSNs. In these studies, the profiles of the same person on different OSNs are identified and joined to reveal that person's privacy. The methods rely on that a user has accounts on multiple OSNs, or they will not work. The studies also do not analyze the content of a user's page, in which the interactions with his/her friends can serve as a good source of rich information for grasping the user's privacy in addition to his/her profile. Each of the studies is detailed below.

Krishnamurthy et al. [9] studied the availability and accessibility of personal identifiable information (PII) on 12 popular OSNs, and mined the PII across the OSNs for third-party web servers to track the browsing behavior of a specific user. However, that work did not analyze the content in the users' pages. Talukder et al. [13] developed a privacy protection tool that measures the amount of sensitive information leakage in a user profile, and suggested self-sanitization actions to regulate the amount of leakage. The inference depends only on the attributes of a user and his/her friends, rather than the information in their interactions. Creese et al. [14] built a data-reachability model for elucidating privacy and security risks related to the usage of OSNs. The model clarifies the reachability of the target information by inferring from a user's publicly available attributes. The discovered information is from the voluntary data revealed by the users, but the involuntary data revealed by their friends are not included. That work did not evaluate the inference accuracy from the experiments. Chen et al. [15] studied the online social footprints and the consistency of attribute revelation patterns across multiple OSNs. In addition, they limited the number of target profiles because of the complexity involved in the crawling process. Irani et al. [16] also studied users' online social

footprints, and concluded that an attacker can reconstruct 10% to 35% of an individual's social footprint by using the person's name with few false positives. However, they did not verify the correctness of personal information, but simply combined it.

The following three studies [19-21] rebuilt social links based on graph theory and inferred sensitive attributes from the graph structure. Yuan et al. [19] considered a graph model where each vertex in the graph is associated with a sensitive label. Hay et al. [20] presented an approach to anonymizing network data. The approach models aggregate network structure and then allowed samples to be drawn from that model. Zhou et al. [21] used a graph model to identify an essential type of privacy attacks called neighborhood attacks.

The two studies [5,18] attempted to infer private information in blogs. We will introduce [5] later. Watanabe et al. [18] describes a new disclosure control system with natural language information analysis, but they do not mention any sensitive attributes they have found.

There are also several recent studies dedicated to privacy inference on Facebook [22-25]. These studies used static user profiles on Facebook and other publicly available sources of personal information for privacy inference according to certain social relationships such as friends, schoolmates and parents (except [24], which studied the leakage from Facebook applications). Compared with them, this work features inferring user privacy based on the user interactions, which were relatively less explored in the literature.

We are also particularly interested in the following studies [5,10-12,17,22-23,25] because they are close to this work. The inference methods on different OSNs are not interchangeable because the personal information policy and the provided attributes are totally different. The related studies are briefly summarized in **Table 1**, and compared with this work in **Table 2**.

Table 1. Studies of private information attributes on OSNs

Paper	Social network	Data collection	Personal information found
[5]	Wretch	crawler program	full name, age, educational background
[10]	Facebook	Facebook API	friendship link
[11]	Facebook	crawler program	age, country, high school name, high school grade year, state
[12]	Facebook	crawler program	music interest
[17]	Foursquare, Google+, Twitter	crawler program	home location
[22]	Facebook	crawler program	age
[23]	Facebook	crawler program	school, city, name, birth year, etc.
[25]	Facebook	crawler program	photo, name, birthdate, etc.

Lam et al. [5] studied involuntary information leakage in a social network service on Wretch [26]. The authors identified some typical patterns for revealing personal information through the one-line annotations from the users' friends, and inferred full names, ages and educational background of the users. However, the inference is hard in general due to the difficulty in precisely grasping the semantics of the free-form annotations. In contrast, we obtain a user's information from Facebook in a well-formed JSON file by the Facebook API.

Becker et al. [11] studied how to measure privacy risk on OSNs. They use a simple, intuitive algorithm: for each attribute, the algorithm selects the most popular value (must exceed a given threshold) of this attribute among a user's friends based on the principle from an old adage: birds of a feather flock together. It is useful and intuitive to find unexposed personal information, but an attribute cannot be inferred if the number of friends sharing the most common attribute value is under the threshold.

As to social relationships, Lindamood et al. [10] used naive Bayes classification to determine a

user's attribute given his or her friendship links. They modified the Naive Bayes algorithm to predict private attributes using both node attributes and friendship link structures. They also modified both attributes (e.g., deleting some information from a user's online profile) and link details (e.g., deleting links between friends) to protect privacy, and also explored the effect of the protection on combating potential inference attacks.

Table 2. Comparison with prior studies

Paper	Pros	Comparison with this work
[5]	Representative study in unintentional and involuntary information leakage	The data on Wretch are in a free form, whereas we collect well-formed data on Facebook
[10]	Find out the relationship in Facebook link without knowing the users' friend list	We gather the users' friend list by Facebook Query Language (FQL) with the consent of the users. The purpose of this work is different.
[11]	Inferring many attributes based on the principle of "birds-of-a-feather-flock-together"	That work selects the most popular value of an attribute whose number exceeds a threshold. We do not need to set a threshold.
[12]	Inferring music interests of users by *Like* data without the consent of the users	This work infers information from the *Like* data with the consent of the users, so we can collect the users' data not available to the public.
[17]	Inferring users' home location by using public available attributes	In this work, we infer information with the consent of the users, so we can collect the users' data not available to the public.
[22]	Inferring users' ages primarily from friends or the users with some friendship links in the social network	In this work, we infer user privacy from their interactions, rather than just the static user profiles.
[23]	Inferring users' privacy from the school information in the user profiles and the friend lists	The same as that in [22].
[25]	Inferring kids' privacy from their parents due to over-sharing	The same as that in [22].

Chaabane et al. [12] studied users' interests on Facebook with an ontology based on Wikipedia because many interests are ambiguous and drawing semantic links between different interests is difficult. Their work presents a semantics-driven inference technique to predict private user attributes. Using only the interest in music often disclosed by users, they extracted unobservable interest topics by analyzing the corpus of interests, and derived a probabilistic model to associate the users with each of the topics. Pontes et al. [17] studied location-based social networks, and focused on inferring a user's home location. They analyzed a simple method to infer the user's home location using publicly available attributes and also the geographic information associated with the locatable friends. Their study was conducted on three popular OSNs: Foursquare, Google+ and Twitter, and the results showed that it is possible to infer a user's home city with high accuracy of around 67%, 72% and 82% on the three social networks.

To the best of our knowledge, there is little or no research (besides [5]) on finding private attributes not provided by a user from the interactions with the user's friends and the content on the user's pages to see if additional personal information can be inferred.

2.2 Ontology

The terms found in the user profiles and interactions may be synonyms or semantically similar, so it is important to identify these terms in the inference process to grasp their meanings. In this work, we use ontology to find out the synonyms and terms similar to the school names. It is noted that using ontology is not a requirement in this work, since only the synonyms and terms similar to the school names are to be identified. However, we still present the usage of ontology, considering that a formal and extensible method such as ontology is necessary in general cases if much more terms in various domains are to be identified in the inference.

We construct a two-layered ontology [27] to organize the terms elicited from Wikipedia. The first layer in the two-layered ontology forms a domain hierarchy presented in **Fig. 1**. This domain ontologies have been mapped to the WordNet [28] domains based on [29]. Each domain contains a term lattice in the second layer presented in **Fig. 2**. The term lattice includes keywords which refer to a school or university. A sub-domain inherits the term lattice from its super-domain, adds new terms specific to the sub-domain itself, and overrides (i.e., redefines) similarity between the terms in the super-domain.

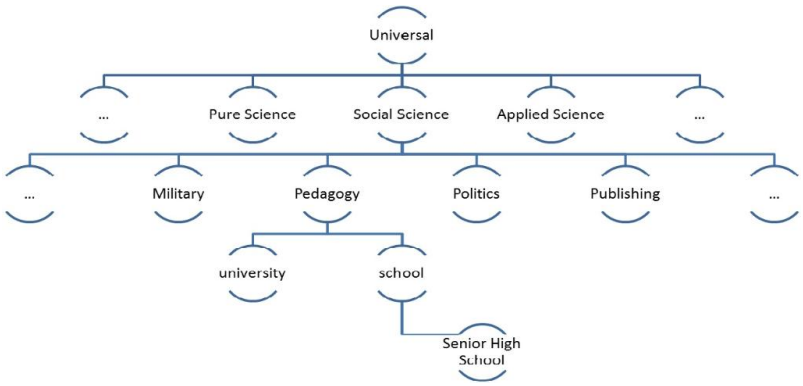

Fig. 1. Domain hierarchy.

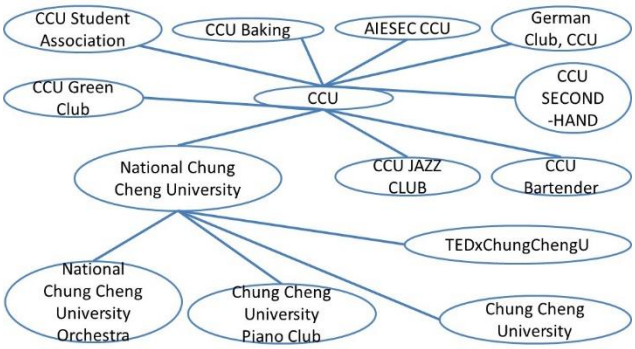

Fig. 2. The term lattice for National Chung Cheng University (CCU)

2.3 J-measure

We use J-measure to measure the quality of the rules to infer the attributes (e.g., the educational background) in the user profiles. Smyth and Goodman [30] introduced the J-measure as an information theoretic means of quantifying the information content of a rule. According to the notation in [30], suppose a rule is in the form *IF* $Y = y$ *THEN* $X = x$, where X

and *Y* are two attributes and *x* and *y* are their values. The information content of the rule is measured in bits and denoted by

$$\mathrm{J}(\mathrm{X};\mathrm{Y}=\mathrm{y}) = p(y) * j(X;Y=y) \quad (1)$$

J(X; Y=y) is a product of the following two terms:

- $p(y)$, the probability that the left-hand side of the rule will occur.

- $j(X;Y=y)$, called the j-measure (with a lower case j) to measure the goodness-of-fit of a rule. Also known as the *cross-entropy*, it is defined as

$$\mathrm{j}(\mathrm{X};\mathrm{Y}=\mathrm{y}) = \mathrm{p}(\mathrm{x}|\mathrm{y}) * \log\big(\mathrm{p}(\mathrm{x}|\mathrm{y})\mathrm{p}(\mathrm{x})\big) + \big(1-\mathrm{p}(\mathrm{x}|\mathrm{y})\big) * \log\Big(\big(1-\mathrm{p}(\mathrm{x}|\mathrm{y})\big)\big(1-\mathrm{p}(\mathrm{x})\big)\Big). \quad (2)$$

The cross-entropy depends on the two values:

- $p(x)$: the probability that the rule consequence (right-hand side) is matched if no other information is available. It is known as *a priori* probability of the rule consequence.

- $p(x|y)$: the probability that the rule consequence is matched if the given antecedents are satisfied. This is also known as *a posterior* probability of *x* given *y*.

3. Inference Methodology

Facebook basic privacy settings and tools have an audience option to specify who can see a user's post on the wall. Several options are available: *Public*, *Friends*, *Friends of friends*, *Only me* and *Custom*. The option *Public* means anyone on and off Facebook can see the post. The option *Friends* or *Friends of friends* means only the user's friends or their friends (i.e., friends of the user's friends) on Facebook can see the post. The option *Only me* means only the user can see the post, and *Custom* allows the user to specify a list of other users who can or cannot see the post. The default audience of new posts (e.g., updating status or adding photos/video) and clicked *Likes* on pages is *friends*. Thus, one can track the posts and profile of a friend, as well as the responsive posts and *likes* from another who is not a friend. As a result, even though those who care about personal privacy are not willing to reveal their privacy, their personal information may be still inferred through such user interactions. In this section, we will present the methods to infer the birthdays and educational background of the Facebook users.

3.1 Involuntary Leakage of Birthday

A user can post messages to a friend's wall on Facebook, which uses *Ticker* [31] to remind users to keep up with the latest news as it happens, such as a friend's birthday. Facebook encourages users to write a birthday wish on a friend's *Timeline*. The birthday attribute must be provided when a user applies for a Facebook account, and it is open to the user's friends and their friends by default. In other words, it means someone unexpected can easily find a user's birthday on Facebook. A user can change the default setting to that *only me* can see the birthday attribute if he/she does not want to open his/her birthday.

3.2 Inferring Birthday

Since Facebook encourages users to say a greeting to a friend's birthday on the wall, we are interested in knowing whether there is a way or not to infer a user's birthday even if his/her birthday is not open to the public. For this purpose, we use a simple, intuitive method to find out a user's birthday. We look for the posts to a user containing the keyword "happy birthday".

Looking for the information is simple because the posts from a user's friends are readily available by downloading using the Facebook API. The date which accompanies most such keywords in the posts is the user's birthday with the highest possibility.

Facebook Query Language (FQL) is a query language that allows to query user data using a SQL-style interface, and we leverage it to systematically search the posts to a user for the keywords to infer birthday. **Stream** is an FQL table used to return a list of stream posts. Several columns, including **source_id**, **message** and **created_time**, in the **Stream** table are used in the inference. **Source_id** is the identifier of a user, a page, a group, or an event whose wall the posts are on. The **message** column contains the messages written in the post. The **created_time** is the time a post was published, expressed in the UNIX timestamp, i.e., UTC. We record the **created_time** of the posts with the keyword "happy birthday", and then choose the most frequent date among the dates of the created time to be the user's real birthday.

3.3 Involuntary Leakage of Educational Background

The previous study on Wretch [5] showed that the connections between users on OSNs are usually a duplicate of their *offline relationships*. That is, the users who have connections on OSNs are likely to have *common attributes*. Unlike Wretch, there are no such relationship tags that describe two users as "classmates" on Facebook. Instead, Facebook has only relationship tags to describe two users as "family" (e.g., sisters, cousins) or "in relation". However, such relationships are rarely relevant to educational background, and it is more difficult to infer users' educational background on Facebook than on Wretch.

The method to infer educational background involves two phases. In the first phase, we reveal the educational background on the basis of ontology because many terms are synonyms of the same school. We build up a domain hierarchy of school names. In every school domain, we have a term lattice composed of keywords learned from Facebook. The learning method will be described in detail in Section 3.3.1. These keywords include the names of pages, groups and locations related to individual schools. Then, we pull out the *Like* data of the users (i.e., who clicked *Likes*), and compare the names of pages and groups liked by the users with the keywords in the term lattice. The *Like* is useful to infer to which school a user belongs with the help of ontology. The usage of ontology will be described in Section 3.3.1.

If the educational background cannot be inferred in the first phase, we infer it by resorting to the assistance of J-measure in the second phase. We focus on college students because the subjects in our study are mostly college students. If the users were juveniles, for example, the educational background would be up to high schools. In that case, only the target of the inference, instead of the inference method, will be different. The previous study [11] selects the most popular value of an attribute from a user's friends, and inferred that the user should have that value in the attribute if the number of friends who share the value exceeds a threshold. That inference method relies heavily on the empirical threshold, which lacks the statistical basis to support it. In contrast, the J-measure is an information theoretic means of quantifying the information content of a rule. In other words, the quality of the rules to infer the target from the attributes can be quantified.

We use three attribute values collected from friends for the inference with J-measure, including living place, gender and educational background. We choose gender in the joint probability calculation because its distribution varies with different colleges. A college of science and technology usually has more male students than female ones; a college of education is in the opposite [32]. Living place usually refers to the location of a school, so it is also considered. With the assistance of J-measure, the inference of educational background can reach a high level of accuracy without a user-specified threshold.

3.3.1 Usage of Ontology

Ontology is used to identify the synonyms of school names, as well as all the pages and groups related to a school. Although we cover only the schools in Taiwan in the experiments, the strategy will be the same for schools elsewhere. We acquire the university/college school name list in Taiwan [33] from Wikipedia, and then search the *groups* and *pages* on Facebook for the school names on the list. After the search, we can get the *groups* and *pages* related to the school names for the inference. The detail is described below.

Groups/pages related to schools on Facebook:

In order to link a user's *Like* data to a school by ontology, finding the connection between the *Like* data and the school name is crucial. We search for the keywords on Facebook with the full name and the abbreviation of a school. Then, we find the groups and the pages with the full name or the abbreviation of that school in the following ways:

- Find the groups with the full school name.
- Find the groups with the abbreviation of school name.
- Find the pages with the full school name.
- Find the pages with the abbreviation of school name.

Taking *National Chung Cheng University* as an example, we learn that CCU is the abbreviation of National Chung Cheng University from Wikipedia. Thus, searching Facebook for the keywords will find the pages/groups named "*National Chung Cheng University*". Then we will find all the pages and groups with the keywords *CCU* and *National Chung Cheng University* in their names.

We reveal the school names by comparing the *Like* data (including the *Likes* on pages and groups) with the ontology term lattice, which can resolve to which school the pages or groups belong. We then observe the pages and groups for the information of *likes*. If a user *likes* some pages or groups, it usually means he or she is likely to have connections with that school. In this work, we assume that if a user *likes* a school, it has a high probability that the user is or used to be a student of that school.

We also need to obtain the attributes of gender and living place to calculate the J-measure in Section 3.3.2. We have almost all the users' genders when pulling out data by FQL. However, around 35% of the users do not provide living places, so we need to infer more living places of the users first. The inference is also based on ontology. Since users like to check into their positions on Facebook, we can build up ontology that resolves the living places of the users from the check-in positions in this work. If a user recently checked in two or more places, we choose the most frequently visited place as the living place.

3.3.2 Use of J-measure

If there are still missing attribute values after the inference based on ontology, we apply the principle of "the most common value of an attribute restricted to a concept" [34] to fill out the missing values, where the concept is the educational background in the inference. We use the example in **Table 3** to explain the filling process. In this example, Taichung is the most common value among the users associated with CCU, so the missing attribute value in the last record of **Table 3** is filled with Taichung.

In this work, inferring educational background by J-measure takes not only the most popular educational background among a user's friends but also living place and gender into consideration. We use ontology to resolve the synonyms of school names, as well as the names of *pages* and *groups*. J-measure is used to evaluate the quality of the inference rules. In the J-measure formula defined in Section 2.3, X represents educational background and Y is gender and living place. The detail is described below.

We refer to the three attributes in the inference: *gender*, *living place* and *school*. Gender is either male or female. The statistics of living place and schools is gathered by counting the number of friends in each living place and associated with each school. Based on the statistics of these attributes, we will build a joint probability table, and infer a user's living place and school with the highest probability according to the J-measure. **Table 4** presents an example for illustrating the computation of the J-measure with the three attributes and their joint probabilities. Assume we have a data set with the attributes: gender (male and female), living place (Taipei, Taichung, Chiayi), and educational background (NCU, NCUE, CCU), so totally 18 rows are in the joint probability table.

Table 3. An example with missing attribute values

Instance	Gender	Living place	Educational background
1	Female	Taichung	NCUE
2	Female	Taipei	NCUE
3	Male	Taichung	CCU
4	Female	Taipei	NCU
5	Male	Taichung	CCU
6	Female	Taipei	CCU
7	Male	Taichung	NCU
8	Female	**Missing**	CCU

Table 4. An example of joint probability for educational background

Gender	Living place	Educational background	Joint probability
Male	Taipei	NCU	0.11
Male	Taipei	NCUE	0.01
Male	Taipei	CCU	0.02
Male	Taichung	NCU	0.00
Male	Taichung	NCUE	0.16
Male	Taichung	CCU	0.01
Male	Chiayi	NCU	0.00
Male	Chiayi	NCUE	0.01
Male	Chiayi	CCU	0.17
Female	Taipei	NCU	0.11
Female	Taipei	NCUE	0.01
Female	Taipei	CCU	0.02
Female	Taichung	NCU	0.00
Female	Taichung	NCUE	0.15
Female	Taichung	CCU	0.01
Female	Chiayi	NCU	0.02
Female	Chiayi	NCUE	0.01
Female	Chiayi	CCU	0.18

Among all the possible rules, we choose only the rules whose antecedents are consistent with a user's attribute values. For example, if a user is a female, then we will consider only the rules whose antecedents indicate a female user. Then we select the consequence of the rule with the maximum J-measure among the rules in consideration as the inference result.

In **Table 5**, we list the conditional probability $p(x|y)$, the left-hand side probability $p(y)$, the cross entropy $j(X;Y=y)$, and their products for possible rules, among which Rule 1 has the highest J-measure among the 10 rules in **Table 5**. The rule implies that if a person lives in

Chiayi, then he/she is more likely to study in CCU. We also found that gender has little influence on inferring educational background in this example.

Table 5. Possible rules to infer educational background

Rule	Rule Description	$p(x\|y)$	$p(y)$	$j(X;y)$	$J(X;y)$
1	if Chiayi then CCU	0.8750	0.40	0.6771	0.2078
2	if male and Chiayi then CCU	0.9444	0.18	0.9476	0.1706
3	if Taichung then CCU	0.0606	0.33	0.4632	0.1529
4	if Female and Chiayi then CCU	0.8571	0.21	0.6196	0.1301
5	if Taipei then CCU	0.1429	0.28	0.2446	0.0685
6	if Male then CCU	0.4082	0.49	0.0000101	0.0000049
7	if Female then CCU	0.4118	0.51	0.0000093	0.0000047
8	if Female and Taichung then NCUE	0.9375	0.16	1.1215	0.1794
9	if Chiayi then NCUE	0.0513	0.39	0.3775	0.1464
10	if Female then NCU	0.2549	0.51	0.0009	0.0004

It is noted that even if some users provide only one or two attributes, we still have a chance to reduce the number of rows in the joint probability table because some combinations of attribute values may be inconsistent with the provided attribute values and the joint probabilities of them can be dropped. The number of possible rules to be considered can be also reduced in the inference.

4. Data Collection and Validation

4.1 Data Collection

We developed a Facebook application using FQL to collect the interactions and profiles of Facebook users. The application can acquire whatever information available to the user who runs it. We began the collection with an initial group of eight users who volunteered to install and use our application for collecting the information available to them. The interactions in the information may include the posts on the walls, the *Like* data, the friends' posts on their walls and their friends' *Like* data. The friend lists, educational background, living places and genders of the users and their friends were also collected if they are available. In the collection, we acquired the available interactions and profiles of totally 3,648 users who have posts on their Facebook wall and had at least one *Like* of groups or pages. Note that although the initial group involves only eight users, this study involves the available information of totally 3,648 users. This number is manageable for us to manually verify the correct rates of the inference given our limited human resources, and has been able to demonstrate the feasibility of inferring private personal information from the user interactions. We agree that involving more users will make the evaluation more rigorous. Nonetheless, that will result only in different leakage rates, but the feasibility of such privacy inference from user interactions still holds.

4.2 Inference Results

In this subsection, we will study the unintentional or involuntary leakage of birthdays and educational background due to the inference from the user interactions. According to the

evaluation, the birthdays of 15.8% more users and the educational background of 16.5% more users can be inferred from the user interactions. The correctness of both inferences are also evaluated for users of different genders, education levels and college majors. The numerical results will be detailed in the following.

4.2.1 Inference Results of Birthday

Among the profiles we have collected, 71.2% of the users open their birthdays to friends. After the inference process, 15.8% more users' birthdays can be identified. The increase does not involve incorrect inference results. The correct rate of inferring birthdays is 85.5% (see Section 4.3.1).

Table 6 presents the rates of birthday leakage. It is surprising that more than half of the people made their birthdays public to friends. The default setting of birthday in Facebook is public to *friend's friend*. If a user does not want to expose his/her birthday, they can change the default setting of birthday into private. However, even if the user does so, we can still find his/her birthday, meaning his/her attempt to keep the birthday unknown fails. In this table, the statistics of self-disclosed birthday, inferred birthday, correct rate of self-disclosed birthday, correctly inferred birthday and totally correct rate are defined below.

Table 6. Rates of birthday leakage

Statistics	Leakage rate
Self-disclosed birthday	71.2%
Inferred birthday	15.8%
Correct rate of self-disclosed birthday	97.5%
Correctly inferred birthday	85.5%
Totally correct rate	92.0%

- **Self-disclosed birthday**: The users make his/her birthday public to friends.
- **Inferred birthday**: The users do not make his/her birthday public, but we still find the correct birthdays.
- **Correct rate of self-disclosed birthday**: We also apply the inference process to those users who open his/her birthday to verify the accuracy of the inference. If the open dates and the inferred ones are the same, then we assume that it is the user's real birthday. If the dates are different, there are two possibilities: 1) The user offers a fake birthday, but we find the real one. 2) The inference finds an incorrect birthday.
- **Correctly inferred birthday**: The rate of correctly inferred birthdays by manual validation.
- **Totally correct rate**: This rate indicates the correct rate among all the birthdays we have found, including self-disclosed and inferred birthdays.

We are also interested in knowing whether or not males and females have a different rate for birthday *leakage*, no matter whether it is self-disclosed or inferred. According to the inference result in **Table 7**, we found the females are more likely to disclose their birthdays to friends.

Table 7. Rates of birthday leakage based on genders

Statistics	Female	Male
Self-disclosed birthday	75.4%	65.8%
Inferred birthday	16.1%	13.8%
Correct rate of self-disclosed birthday	98.0%	97.2%
Correctly inferred birthday	90.3%	89.6%
Totally correct rate	94.1%	93.7%

We also summarize the self-disclosure and the inference results based on ages, education levels and college majors. We did not infer the ages, education levels and college majors of the users, but chose the results from those who disclose the three attributes for the analysis. Among all the users in this study, 46% of the users open their ages to friends. The self-disclosure rate of education level is 75%, and that of college major is 52%.

Fig. 3 presents that the correct rate of birthday inference is higher for users at ages 15-18. **Fig. 4** indicates that senior high school students are more likely to open their birthdays to friends. The numbers in **Fig. 3** and **Fig. 4** are consistent, because senior high school students are in the age group of 15-18 years old.

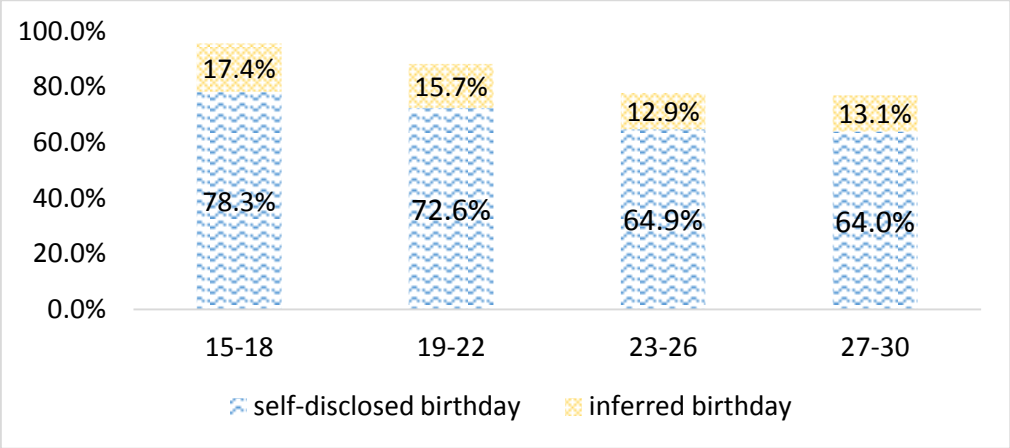

Fig. 3. Correct rates of birthday inference based on ages

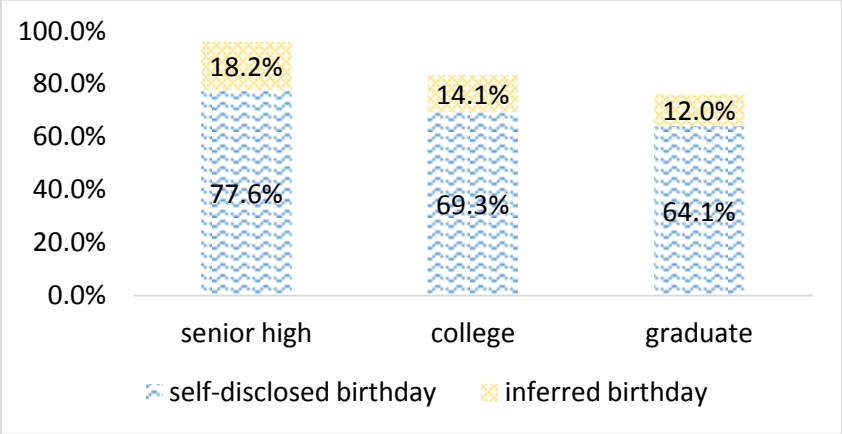

Fig. 4. Correct rates of birthday inference based on education levels

In **Fig. 5**, the correct rates of the colleges of humanities and social science are slightly higher than those of the colleges of science and engineering. The rates of the colleges of law and management are in between.

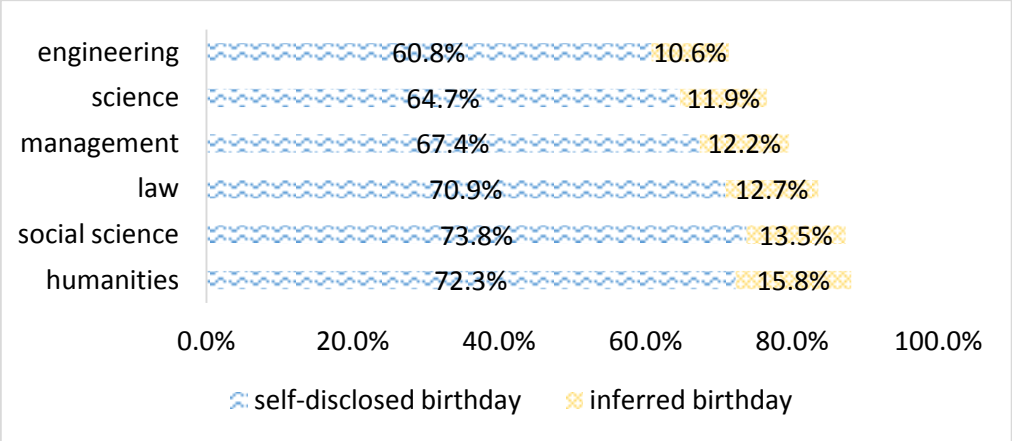

Fig. 5. Correct rates of birthday inference based on college majors.

4.2.2 Inference Results of Educational Background

Inferring educational background involves two phases. The experimental result is shown in **Table 8**. The definitions of the statistics in this table are similar to those in **Table 6** and are self-explanatory, so we do not repeat them herein. In the first phase, we compare users' *Like* data with the ontology term lattice. In the second phase, we use the J-measure to measure the rules to infer the educational background. The correct rate in this phase is 75.5%. In the two phases in total, the educational background of 16.5% more users can be inferred, and the educational background of totally 86.3% of the users can be correctly derived.

Table 8. Rates of educational background leakage

Statistics	Leakage rate
Self-disclosed educational background	75.2%
Inferred educational background in phase 1	12.5%
Inferred educational background in phase 2	4.0%
Correctly inferred educational background in phase 1	80.0%
Correctly inferred educational background in phase 2	75.5%
Totally correct rate	86.3%

The method in this work is compared with the previous study [11], which selects the most popular value of educational background among a user's friends. Since it is illogical to identify the birthday based on the concept of "birds of a feather flock together", we compare only the inference of educational background. The comparison is listed in **Table 9**. Our method can correctly infer more educational background than the work in [11].

Table 9. Comparison with [11] in terms of the rates of inferred educational background

Statistics	This work	[11]
Self-disclosed educational background	75.2%	
The rate of inferred educational background	16.5%	8.8%
Correctly inferred educational background	78.8%	73.2%
Totally correct rate	86.3%	77.8%

Table 10 presents that males are more likely to self-disclose their eduction records than females, but the involuntary leakages of educational background are insignificant for both

genders in either phase 1 or phase 2.

Table 10. Correct rates of educational background leakage based on genders

Statistics	Female	Male
Self-disclosed educational background	73.1%	77.9%
Inferred educational background in phase 1	11.9%	12.8%
Inferred educational background in phase 2	3.2%	4.3%
Correct inferred educational background in phase 1	84.4%	85.8%
Correct inferred educational background in phase 2	76.5%	74.1.%
Totally correct rate	85.7%	86.9%

In **Fig. 6**, the numbers between different majors are also insignificant. The self-disclosure rates are over 70%. The percentages of inferred educational background after phase 1 and after phase 2 are over 12% and around 4%, respectively.

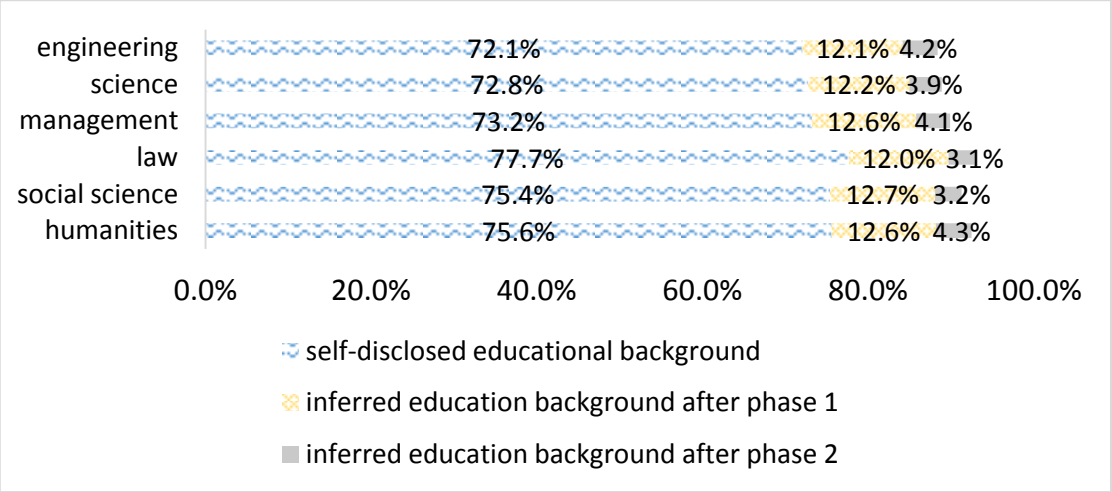

Fig. 6. Rates of educational background leakage based on college majors

4.3 Verification of Inference Results

4.3.1 Verification of Inferring Birthdays

We have a complete verification of the birthday inference results by manual checking, and examine the causes behind incorrect inference. The majority of incorrectly inferred birthdays are caused by mistaking a few days before or after the real birthday as the real birthday. Two causes can result in this imprecision. 1) A user provides a fake birthday a few days before or after his/her real birthday, and his/her friends gives their blessing on the wrong day. 2) A user does not provide his/her birthday, but some of his/her friends roughly know the user's birthday. They still give a blessing, but the date may be a few days different from the real birthday. Thankfully, there are still some cases in which even though a user provides an incorrect birthday, we can still find the correct one. In this case, the user's friends know his/her birthday, and they give a blessing on the correct date.

4.3.2 Verification of Inferring Educational Background

The complete verification of inferring educational background is also manually conducted.

The majority of incorrectly inferred educational background is due to two causes: 1) If a student transferred from one school to another, the inference result may be the previous school he or she used to study at. 2) If two schools are closely located, the students at the two schools are likely to be familiar with one another and perhaps participate in common activities. In this case, a student at one school may be mistaken for one at the other in the inference.

5. Conclusion and Future work

This work studies the possibility of involuntary privacy leakage on real-life OSNs, and demonstrates the degree of involuntary leakage is significant. The study was conducted on Facebook, one of the most popular OSN service. Facebook encourages the users to provide personal profiles as complete as possible. Some users prefer not to disclose their profiles online because they realize it is unsafe to do so. However, Facebook users are allowed to write posts on their friends' walls, so personal information, such as a person's birthday, can be still disclosed easily. We show that 71.2% of users disclose their birthdays and 15.8% more can be inferred. About 75.2% of users disclose their educational background and 16.5% more can be inferred. Although it is true that a stranger cannot conduct such privacy inference unless a Facebook user explicitly makes his/her posts public, we want to emphasize in this work that adding a friend without care will leak more personal privacy easily through user interactions. Even a stranger still has a chance to approach a user with social engineering techniques and infers his/her privacy. Thus, we suggest users carefully control the private setting of posted messages and actively remove all private information on the walls, such as the birthday blessings from friends, to protect their privacy. Moreover, a user should pay attention to not accepting requests of making friends from unfamiliar users, or his or her interactions on Facebook will be likely to be available to strangers who may want to infer the private information.

In future work, we can apply the ITRULE algorithm [35] to speed up rule searching, and thus we can add more attributes into consideration. Second, since we have users' posts on the walls, we can collect the check-in places to infer their living places, and further reduce the number of missing attributes of living places. Furthermore, the candidates of educational background in the joint probabilities for computing J-measure can be reduced by choosing only the educational background of a user's close friends. The close friends are those who have interactions with the user in the past period of time. Thus, the result of inferred educational background will be refined.

Acknowledgment

This work was supported in part by Ministry of Science and Technology, Taiwan, also in part conducted under the "III Innovative and Prospective Technologies Project (1/1)" of the Institute for Information Industry which is subsidized by the Ministry of Economic Affairs of the Republic of China.

References

[1] C. Matthews, "1 billion people used Facebook on Monday." Article (CrossRef Link).
[2] R. C. He, "Introducing new Like and Share buttons," Article (CrossRef Link).
[3] A. Mitchell, J. Kiley, J. Gottfried and E. Guskin, The Role of News on Facebook. Article (CrossRef Link).
[4] T. Risen, Happy Birthday: Facebook Celebrates Its 10th Birthday, Article (CrossRef Link).

[5] F. Lam, K. T. Chen and L. J. Chen, "Involuntary Information Leakage in Social Network Services," *Third International Workshop on Security (IWSEC)*, Nov. 2008. Article (CrossRef Link).
[6] A. N. Joinson and C. B. Paine, "Self-disclosure," *privacy and the Internet, In the Oxford Handbook of Internet Psychology*, pp. 237-252, 2007. Article (CrossRef Link).
[7] K. R. Goldner, "Self Disclosure on Social Networking Websites and Relationship Quality in Late Adolescence," ETD Collection for Pace University, Jan. 2008. Article (CrossRef Link).
[8] E. McCollister, T. Grance and K. A. Scarfone, "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)," In: NIST SP - 800-122. pp 58, Apr. 2010. Article (CrossRef Link).
[9] B. Krishnomurthy and C. E. Wills, "On the Leakage of Personally Identifiable Information via Online Social Networks" in *Proc. of the 2nd ACM Workshop on Online Social Networks (WOSN)*, Aug. 2009. Article (CrossRef Link).
[10] J. Lindamood, R. Heatherly, M. Kantarcioglu and B. Thuraisingham, "Inferring Private Information Using Social Network Data" in *Proc. of* the 18th International Conference on World Wide Web (WWW), Apr. 2009. Article (CrossRef Link).
[11] J. Becker and H. Chen, "Measuring Privacy Risk in Online Social Networks" in *Proc. of Workshop of Web 2.0 Security and Privacy (W2SP)*, May 2009. Article (CrossRef Link).
[12] A. Chaabane, G. Acs and M. A. Kaafar, "You Are What You Like! Information Leakage Through Users' Interests," in *Proc. of Annual Network and Distributed System Security Symposium (NDSS)*, Feb. 2012. Article (CrossRef Link).
[13] N. Talukder, M. Ouzzani, A. K. Elmagarmid, H. Elmeleegy and M. Yakout, "Privometer: Privacy Protection in Social Networks," in *Proc. of IEEE International Conference on Data Engineering (ICDE) Workshops*, Mar. 2010. Article (CrossRef Link).
[14] S. Creese, M. Goldsmith, J.R.C. Nurse and E. Phillips "A Data-Reachability Model for Elucidating Privacy and Security Risks Related to the Use of Online Social Networks," in *Proc. of IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*, June 2012. Article (CrossRef Link).
[15] T. Chen, M. A. Kaafar, A. Friedman and R. Boreli, "Is More Always Merrier? A Deep Dive Into Online Social Footprints," in *Proc. of the ACM Workshop on Online Social Networks (WOSN)*, Aug. 2012. Article (CrossRef Link).
[16] D. Irani, S. Webb, K. Li and C. Pu, "Large Online Social Footprints - An Emerging Threat," in *Proc. of the International Conference on Computational Science and Engineering* - Volume 03, Aug. 2009. Article (CrossRef Link).
[17] T. Pontes, G. Magno, M. A. Vasconcelos, A. Gupta, J.M. Almeida, P. Kumaraguru and V. Almeida, "Beware of What You Share: Inferring Home Location in Social Networks," in *Proc. of The IEEE International Conference on Data Mining series (ICDM)*, Dec. 2012. Article (CrossRef Link).
[18] N. Watanabe and H. Yoshiura, "Detecting Revelation of Private Information on Online Social Networks" *IIH-MSP*, page 502-505. IEEE Computer Society, Oct. 2010. Article (CrossRef Link).
[19] M. Yuan, L. Chen, P. S. Yu and T. Yu, "Protecting Sensitive Labels in Social Network Data Anonymization" *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 25, no. 3, pp. 633-647, Mar. 2013. Article (CrossRef Link).
[20] M. Hay, G. Miklau, D. Jensen, D. Towsley and P. Weis, "Resisting Structural Re-Identification in Anonymized Social Networks" in *Proc. of the Very Large Database (VLDB) Endowment*, vol. 1, pp. 102-114, Aug. 2008. Article (CrossRef Link).
[21] B. Zhou and J. Pei, "Preserving Privacy in Social Networks against Neighborhood Attacks," in *Proc. of IEEE 24th International Conference Data Engineering (ICDE)*, Apr. 2008. Article (CrossRef Link).
[22] R. Dey, C. Tang, K. Ross and N. Saxena, "Estimating Age Privacy Leakage in Online Social Networks," *IEEE INFOCOM*, Mar. 2012. Article (CrossRef Link).

[23] R. Dey, Y. Ding and K. W. Ross, "Profiling High-School Students with Facebook: How Online Privacy Laws Can Actually Increase Minors' Risk," in *Proc. of ACM Internet Measurement Conference (IMC)*, Oct. 2013. Article (CrossRef Link).
[24] A. Chaabane, Y. Ding, R. Dey, M. A. Kaafar and K. W. Ross, "A Closer Look at Third-Party OSN Applications: Are They Leaking Your Personal Information?" in *Proc. of Passive and Active Measurement Conference (PAM)*, Mar. 2014. Article (CrossRef Link).
[25] T. Minkus, K. Liu and K. W. Ross, "Children Seen But Not Heard: When Parents Compromise Children's Online Privacy," in *Proc. of International World Wide Web (WWW) Conference*, May 2015. Article (CrossRef Link).
[26] Wikipedia Wretch page, Article (CrossRef Link).
[27] L. F. Lai, C. C. Wu, P. Y. Lin and L. T. Huang, "Developing a fuzzy search engine based on fuzzy ontology and semantic search," in *Proc. of IEEE International Conference on Fuzzy Systems (FUZZ)*, June 2011. Article (CrossRef Link).
[28] WordNet Domains Hierarchy, Article (CrossRef Link).
[29] L. Bentivogli, P. Forner, B. Magnini and E. Pianta, "Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing," in *Proc. of Workshop on Multilingual Linguistic Resources*, Aug. 2004. Article (CrossRef Link).
[30] P. Smyth and R. M. Goodman, "Rule Induction Using Information Theory," In: G. Piatetsky-Shapiro and W.J. Frawley (eds.), Knowledge Discovery in Databases. AAAI Press, pp. 159-176, 1991. Article (CrossRef Link).
[31] Facebook Ticker introduction, Article (CrossRef Link).
[32] S. L. Ewert, "The Determinants of Gender Inequality in Higher Education," *Ewert, Stephanie L., ProQuest LLC, Ph.D. Dissertation*, University of Washington, 2010. Article (CrossRef Link).
[33] Wikipedia university/collage school list in Taiwan. Article (CrossRef Link).
[34] O. Maimon and L. Rokach, "The Data Mining and Knowledge Discovery Handbook," *Tel-Aviv (eds.)*, University Israel, pp. 40-41. Article (CrossRef Link).
[35] P. Smyth and R. M. Goodman, "An Information Theoretic Approach to Rule Induction from Databases," *IEEE Trans. Knowledge and Data Engineering*, vol. 4 no. 4, pp. 301-316, Aug. 1992. Article (CrossRef Link).

Po-Ching Lin received the M.S. and Ph.D. degrees in computer science from National Chiao Tung University, Hsinchu, Taiwan, in 2001 and 2008, respectively. He joined the faculty of the Department of Computer and Information Science, National Chung Cheng University (CCU), Chiayi, Taiwan, in August 2009. He is currently an Associate Professor. His research interests include network security, network traffic analysis, and performance evaluation of network systems.

Pei-Ying Lin received her master's degree in Computer Science and Information Engineering from National Chung Cheng University in 2014. She works at HTC Corporation after graduation. Her research interests include social networking and mobile systems.