DECISION SCIENCES INSTITUTE

Preventing Inference Attacks in Online Social Networks: A Data-Driven Framework

(Full Paper Submission)

Xiaoyun He
Auburn University at Montgomery
xhe@aum.edu

Haibing Lu
Santa Clara University
hlu@scu.edu

ABSTRACT

In recent years, the growth of online social networking sites has been meteoric. With the increasing user involvement, social networks now serve as a repository of all kinds of information. While there have been studies demonstrating private information can be inferred from social networks, few has taken a holistic view on designing mechanisms to prevent the inference attacks. In this study, we present a framework that leverages the data from both social networking service provider and user profiles to proactively find possible inference attacks against users, and suggest minimal modifications to the user profiles to eliminate such attacks.

KEYWORDS: Privacy, inference attacks, online social networks, collaborative approach

INTRODUCTION

Recent years have witnessed unprecedented growth of online social networking sites (OSNs) such as Facebook, LinkedIn, XING, etc. For example, as of December 2015, Facebook reports 1.59 billion monthly active users, with a growth rate as high as 3% per week (Facebook 2015); and LinkedIn boasts over 400 million members in over 200 countries, with a new member joining LinkedIn approximately every second (LinkedIn 2015). The emergence of OSNs has essentially transformed the way people express themselves and establish social connections in the digital era (boyd and Ellison 2007).

Individual users of an online social network can create profiles containing various personal attributes such as age, group affiliation, lists of personal interests, contact information, and so on. Some OSNs offer the ability to create and join groups that share common interests or affiliations, upload live videos, and hold discussions in forums. To alleviate privacy concerns (Barnes 2006), OSNs usually have controls that allow users to choose their own privacy settings, for instance, whether to make your profile public or private, who is allowed to view the profile, contact you, add you to their list of contacts, and so on.

Despite the availability of the privacy setting controls to the users, such controls are not sufficient in enforcing privacy of users (Singh et al. 2009). Indeed, privacy issues arising in

OSNs are of real problem. This has been evidenced by both the media exposure and the studies from the academia (e.g., Dwyer et al. 2007; Gross and Acquisti 2005; Zeller 2006). OSNs users may be vulnerable to various privacy attacks, such as automated user profiling (Balduzzi et al. 2010; Dougnon et al. 2015), identity attacks via user de-anonymization techniques (e.g., Narayanan and Shmatikov 2009; Wondracek et al. 2010), and inference attacks (e.g., He et al. 2006; Lindamood et al. 2009; Zheleva and Getoor 2009), and so on.

In order to prevent some of the above privacy attacks, technical solutions have been proposed in the literature. For instance, several anonymization solutions have been proposed to protect against identity attack in OSNs (e.g., Backstrom et al. 2007; Zou et al. 2009). While a recent study has proposed sanitization techniques that could be used to prevent inference attacks (Heatherly et al. 2013), there lacks a holistic view that can help to enforce privacy requirements or give any assurance on the level of protection in the presence of possible inferences. In this study, we focus on inference attacks, and present a data-driven framework that can be used to alleviate such attacks through the collaboration from both sides of social networks providers and users.

RELATED LITERATURE

Along with the increasing popularity of OSNs, OSNs users encounter growing risks of privacy attacks. Online Social Networking site provides a rich digital forum for social interactions. While some OSNs users are open to sharing and posting personal information, others may not. Often, the privacy risks posed to OSNs users are unexpected and unaware of. The most immediate danger of posting on OSNs is that it may leave a permanent fingerprint of whatever being posted (Rosenblum 2007). With the click of a button, what a user just typed could be instantly disseminated, and stored in countless independent permanent places. Next-minute damage recovering is almost impossible. Worse, the power of a search engine makes it searchable within a few seconds.

Some concrete examples of the privacy risks to OSNs users include stalking, spamming, and the possible damage to their future educational and career opportunities. With the information from a user profile, a potential adversary can determine the likely physical location of the user and thus incur real-world stalking (Gross and Acquisti 2005). In addition, most OSNs, such as Facebook, currently provide the third-party applications with full access to user profile data, which further deteriorate privacy control. For example, a popular Facebook application, Compare Friends, that promised users' privacy in exchange for opinions on their friends later started sell this information (Singh et al. 2009).

Inference Problem

Several studies show that, even OSNs users take advantage of the available privacy setting control to make their profiles or sensitive attributes private (i.e., no one but their friends can see their profile detail), it is still possible to infer private information that a user is not willing to disclose (e.g., He et al. 2006; Heatherly et al. 2013; Lindamood et al. 2009; Zheleva and Getoor 2009). In fact, they demonstrate that a surprisingly large amount of information can be leaked by just exploiting friendship links and/or group affiliations in the real-world online social network data.

In particular, the problem of sensitive attribute inference is to infer the hidden sensitive attribute values of private profiles that are conditioned on the observed sensitive values, links and group memberships in a social network (Zheleva and Getoor 2009). It is assumed that an adversary can apply a probabilistic model for predicting the hidden sensitive values. By combining the given information of a social network in difference ways, the adversary can launch various inference attacks (i.e., attacks without links and groups, attacks using links, attacks using groups, or a mixed of them).

As a simplified example, suppose users have attributes of gender and self-declared political views in a collected Facebook data (Zheleva and Getoor 2009). Assume that user Alice has set her profile as private. From the group information available in Facebook, Alice belongs to four groups: Crochet Mastery, the National Coalition on Health Care, Health Care for America Now, and Hand Embroidery. Based on these public information, the group-based classification model can accurately predict that Alice is female and a liberal. With assumption of 50% private profiles, the attack accuracy for gender attribute reaches at 73.4% by using group-based classification model; and the accuracy is 72.5% by using both links and groups.

The above inference problem does have significant privacy implications. Since fewer users in OSNs hide their friendship links and even if they do, their friendship links can still be constructed through the backlinks from their public profile friends. Group participation information availability is similar - even if a user keeps his / her profile private, his / her participation in a group is displayed on the group's membership list. Currently, most OSNs (e.g., Facebook and MySpace) do not allow users to hide their group membership from public groups.

In addition, privacy attacks to OSNs users may be motivated by the various self-interests of the attackers, such as targeted marketing, insurance screening, email phishing, or political monitoring (Bonneau and Preibusch 2010). These attacks would have negative effects on the credibility and attractiveness of OSNs. Therefore, it is critical for OSNs operators to protect their users against undesired private information digging and inform them of possible privacy breaches. Further, research has shown that simply removing some attribute information or friendship links may not be enough to prevent inference attacks, and thus comprehensive techniques are needed in practice (Heatherly et al. 2013; Lindamood et al. 2009).

PROPOSED FRAMEWORK FOR INFERENCE PREVENTION

As discussed earlier, the various data from different sources including user profiles and postings can be used to initiate inference attacks. On the other hand, some of these data can be leveraged to incorporate privacy protection mechanism into online social network services. Here, we present a data-driven framework addressing the inference problem identified earlier. Figure 1 depicts the framework.

At the service provider side, the online social network data encompasses the profile (personal information, friends list, group memberships, etc.) of all users and their interaction (posts on the walls, status updates, etc.). This data is modeled as a semantic graph and is used to develop a knowledge base consisting of an inference attribution repository and a contradictory information repository.

Figure 1: The Conceptual Framework for Inference Prevention

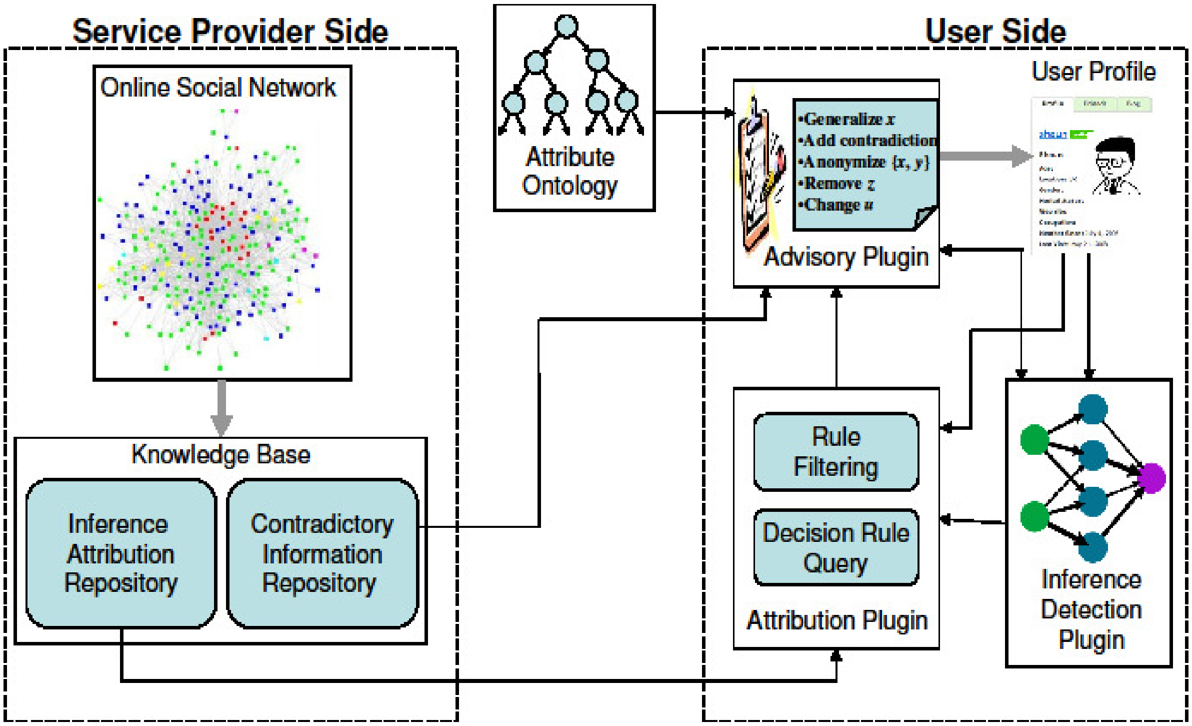
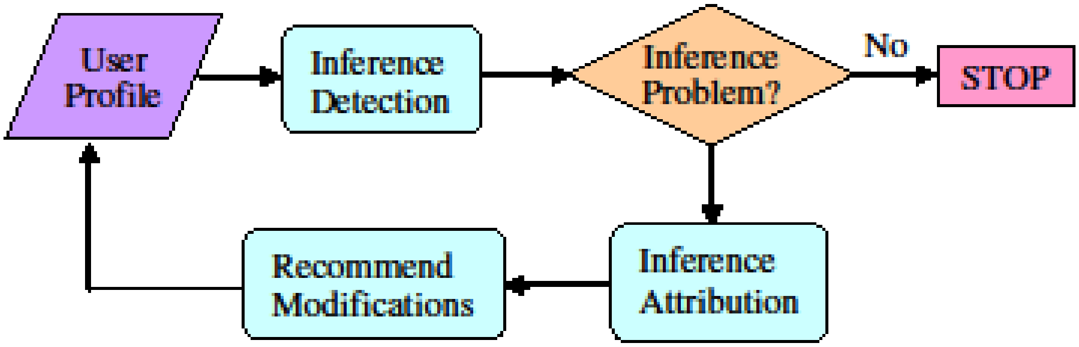


Figure 2: The Flow Chart for Inference Prevention



The inference attribution repository is created by applying classification algorithms such as decision tree learners, for example, C4.5 (Quinlan 2014), on the online social network data to discover the attribution rules for sensitive attributes. The contradictory information repository at the service provider site is also a set of rules where the attributes/attribute values in the conditional part of the rule contra-indicate (inversely relate to) the sensitive attribute value.

The service provider also provides a set of plugins including inference detection plugin, attribution plugin, and advisory plugin which can be downloaded by a user and run at his/her site to check any privacy breach or change the profile to avoid such breach. The inference detection plugin is primarily a classifier that given a user profile predicts what sensitive attributes of the user can potentially be leaked. This can be based on any combination of the state of the art inference techniques that have been developed in the literature.

The process flow for inference prevention is shown in Figure 2. This process is initiated by the user whenever the user creates a new online profile or updates an existing profile. This profile is taken as an input to the inference detection plugin running locally at the user site. Based on the privacy preferences of the user, the inference detection plugin detects if the user profile is consistent with the privacy preferences or could possibly leak sensitive attribute values. As a running example, suppose that on January 4, 2010, user Alice has the following attribute values in her profile: Location=`Sunshine State', Group Affiliation = `NewYearBornParty', and Recent Activity = `Just qualified for full driver license'. Since `Sunshine State' implies that she is in Florida, `NewYearBornParty' is a group for people who are born on January 1st, and the age limit for obtaining full driver's license in Florida is 18, thus it can be inferred that Alice's birthdate is January 1, 1992. The inference detection plugin detects the following inference: Location(`Sunshine State') and GroupAffiliation(`NewYearBornParty') and RecentActivity(`Just qualified for full driver license') infers BirthDate(`January 1, 1992').

Once an inference problem is detected, the attribution plugin is activated to determine the source of the problem (i.e., what causes the detected inference). Specifically, the attribution plugin interacts with the inference attribution repository in the knowledge base to identify the possible attribute value(s) contributing to the detected inference problem. Based on the relevant inference rules retrieved from the inference attribution repository, the advisory plugin makes suggestions to the user on what needs to be modified in the profile to break the detected inference.

Note that online social network is a dynamic environment, thus the contents of the knowledge base as well as the inference detection plugin should evolve over time. The OSN service provider needs to keep updating and adding rules to the knowledge base, so that it reflects the most current state of potential inferences. This evolving process may also involve learning from the past experience. The goal is to make it better capture the potential inferences as much as possible.

Functionality Description

We now discuss the detailed functionality of each component of the inference prevention framework and the algorithms employed to support such functionality.

Inference Attribution Repository

The inference attribution repository stores a set of rules for all the attributes that are declared private/sensitive in at least of the user's profile. These rules are derived from the online social network data by applying classification algorithms such as decision tree learners C4.5 for the given set of sensitive attributes. Specifically, for each sensitive attribute, a decision tree can be learned from the data that provides the combination of attribution values leading to the inference of the private data. For example, given a sensitive attribute value y, a standard decision tree generation algorithm C4.5 could potentially come up with an attribution rule that if attribute values x and z are present in the profile or if w and q are present in the profile, y could be inferred.

Contradictory Information Repository

The knowledge base also has a Contradictory Information Repository. This component is used to record inference rules that directly contradict the detected sensitive information. Such contradictory information can also be used to break the detected inferences, since its presence may directly contra-indicate the presence of the inferred information. This can be a very appealing option if the user would like to leave their current profile unchanged but still create the illusion of possessing the contradictory attribute.

The question that then arises is how we can find such contra-indicative information. Such information can actually be derived from the social network data itself, by using a clever trick. The key idea is that we are looking for anti-correlations. To detect these, we thus can reverse or modify the target attribute value in the original data based on which an initial new inference is detected. Then, the same module used for the inference detection (say, decision tree learner) can be run against the newly reversed data. This will give us a new set of inference rules that indicate the given target value. Since we have purposely reversed the original target value, comparing the affecting attribute values in these new rules with those in the initial inference would tell us the contradictory information existing between them.

Inference Detection Plugin

The Inference detection plugin is simply a classifier model made available to users by the service provider to support inference detection. This could be based on any machine learning or data mining technique, or their combination. For example, the service provider could train a neural network based classifier using the online social network data. When the trained model is provided to the user, the user can easily run their profile through the model to identify possible inferences. One requirement of such an approach would be that the underlying models have to be scalable as well as incrementally updatable to incorporate the ever-expanding data accrued in the social network.

Attribution Plugin

The attribution plugin interacts with the knowledge base at the service provider site to determine the possible attributes or attribute values in the user profile that leads to disclosure of sensitive attributes as determined by the inference detection plugin. Specifically, the attribution plugin queries the inference attribution repository for all the rules that includes the given sensitive attribute in the target value or the target value of such rules directly or indirectly influence the conditional part of the rule with the given sensitive attribute in the target value.

Since the attribution repository includes rules created from the entire social network, some of the rules in the query result may not be relevant, as the given user profile may not include the attributes specified in those rules. Therefore, a filtering process is employed to purge out the irrelevant rules.

Advisory Plugin

Advisory plugin is run locally at the user site but it interacts with the contradictory information repository at service provider site, attribute ontology that may be hosted by a third party, and attribution plugin running at user site to suggest possible modifications in the user profile for breaking inference. Some possible modifications can be the following. (1) Generalize attribute values. This option can utilize an attribute ontology for generalization of the relevant attribute values. (2) Perturb attribute values. To break the inference, an attribute value can be distorted. (3) Add contradictory information. Another possibility is to actually introduce new data within the profile that directly contra-indicates the presence of the sensitive value. (4) Remove certain information. Instead of directly adding contradictory information, another possibility is to simply eliminate some of the information causing the inference, thus breaking the linkage. (5) Make some more information private. Instead of making any modifications to the user profile, an alternative is to just make the attributing information itself private. In this fashion, it cannot be used to make the sensitive inferences. Of course, this only applies as long as the newly private made information itself cannot be inferred, thus possibly requiring multiple runs of the inference prevention process. (6) or a combination of all of the above. In certain cases, a single modification of any of the above kinds may not be sufficient to ensure privacy, in which case, we can use a combination of the above to enforce it.

In order to guarantee that the above instrumented modifications have made the user free of inferences, it may be necessary to run inference detection module again. This is due to the possibility of creating new inferences by the modifications.

Ontology

As discussed above, breaking inference may require generalizing some of the attribute values in user profile. For this, the advisory plugin can utilize a relevant ontology which may be provided by a third party, such as VIVO ontology repositories (VIVO 2015) and DAML ontology (DAML 2015). The generalizable attributes in a user profile can be categorized into multiple types based on their semantics and the domains they take their value from. There can be different attribute categories such as address and locations, organizational affiliations, etc. that are generalizable and are commonly used in the online profiles of users. For each category, ontology from appropriate domains needs to be used for attribute value generalization.

Minimization of Modification

In order to break the detected inferences, we may need to modify several attribute values in the user profile. Ideally, we would like to minimize this modification while still breaking the inferences. That is, usability should be enhanced as much as possible while satisfying privacy protection requirement.

Recall that, to determine the inference source for a user, the Attribution plugin may get a set of candidate rules from the knowledge base. Based on these rules, modification suggestions are given. However, among the candidate rules, there may be overlaps on the affecting attribute values. That is, some of the attribute values may appear in multiple candidate rules. Given our rule-based approach, minimizing the modification of user profile is thus equivalent to finding a minimal set of affecting attribute values such that all the candidate rules contains at least one of these attribute values. In other words, we need to find a set of attribute values of the smallest size which would still break all the candidate inferences. This can be solved by finding a minimal set of attribute values.

CONCLUSION AND FUTURE WORK

In this study, we have focused the inference problem arising in the popular online social networking sites. We propose a framework to eliminate the rule-based inference problem by detecting and breaking the inferences that are represented as rules of attributes and / or attribute values. To prevent privacy inferences, user profiles may need to be modified. We briefly show that how this modification can minimized so that usability can be enhanced. We plan to implement and release a privacy plugin for the users of social networking platforms such as Facebook in the future. While our framework protects users from common privacy inferences, background knowledge when added to the user's posts and interactions may still suffice to breach privacy. We plan to examine this issue in the future.

REFERENCES

References available upon request.