Privacy Protection in Social Networks via Inference Technique

Jianming He, *Student Member, IEEE,* and Wesley W. Chu, *Fellow, IEEE*

Abstract

Since privacy information can be inferred via social relations, the privacy confidentiality problem becomes increasingly challenging as online social network services gain more popularity. Using a Bayesian network approach to model the causal relations among people in social networks, we find that personal attributes can be inferred with high accuracy especially when people are connected with strong relations. We then propose schemes to protect privacy by selectively hiding or falsifying information based on the characteristics of the social network. We demonstrate through experiments and analysis that these methods are much more effective than randomly altering information.

I. Introduction

With the increasing popularity of Social Network Services (SNS), more and more online societies such as Friendster, Blogger, Orkurt and Livejournal have emerged. Unlike traditional personal homepages, people in these societies not only publish their own personal profiles (e.g., age, gender, interests, professions and weblogs), but also reveal their social relations such as friends and family. As these networks grow rapidly, many interesting research topics arise [2]–[4]. Unfortunately, among these topics, privacy has not been fully addressed. Given the huge amount of personal data and social relations available in online social networks (for example, Friendster claims that it owns over 24 million personal profiles), it is foreseeable that privacy may be compromised if people are not careful in releasing their personal information.

This research is in part supported by NSF grant # IIS-03113283. A preliminary version of this work was presented in [1].

Information privacy has become one of the most urgent research issues in building next-generation information systems. A great deal of research effort has been devoted to protecting people's privacy. Aside from recent developments in cryptography and security protocols that provide secure data transfer capabilities, there has been effort in enforcing industry standards (e.g., P3P [5]) and government policies (e.g., the HIPAA Privacy Rule [6]) to grant individuals control over their own privacy. These techniques and policies aim to effectively block *direct* disclosure of sensitive personal information. For instance, the Platform for Privacy Preferences Project (P3P) enables Web users to explicitly specify their privacy preferences during their interaction with Web applications such as online shopping. Users will be alerted when applications fall short of their privacy preferences. However, to the best of our knowledge, none of the existing techniques handle *indirect* disclosure which can often be achieved by intelligently combining pieces of seemingly innocuous or unrelated information. Specifically, in scenarios like social networks, we realize that even if individuals do not release their personal information to the public, such data can still be disclosed by individuals' social relations.

Intuitively, friends tend to affect each other. Individuals connected in social networks often share common attributes. For instance, in a dance club, people come together because of their common interest; in an office, people get acquainted with each other due to similar professions. Therefore, it is possible that one may be able to infer someone's attribute from the attributes of his/her friends. For example, Joe does not wish to release his salary information to other people. However, a third party, such as an insurance company, can use online societies to obtain a report on Joe's social network, which lists Joe's friends and colleagues and their personal information. After looking carefully into this report, the insurance company might realize that most of Joe's colleagues are professors at a university. Thus, they can deduce that Joe is also a professor with a high probability (if not provided by Joe himself). With the knowledge of a professor's salary range, the insurance company can in turn figure out Joe's approximate salary. Therefore, in this example, privacy is indirectly disclosed from Joe's social relations rather than from himself directly, and existing techniques such as P3P cannot prevent such a disclosure.

In this paper, we study protection techniques for potential privacy disclosure in social networks. We first analyze under what conditions and to what extent privacy might be disclosed by social relations. More specifically, we propose an approach to map social networks into Bayesian networks. Using causability encoded in Bayesian networks, we successfully model social relations among people. We perform experiments on a real social network structure and study the impact of social network characteristics on the inference result. Based on these results, we then propose privacy protection schemes and evaluate their effectiveness.

The major contributions of this paper are as follows.

- Identify that privacy disclosure can take place via social relations in social networks.
- Present an approach to model social networks into Bayesian networks and use Bayesian inference to predict sensitive attribute values.
- Study the impact of such social network characteristics as influence strength, prior probability, and society openness on Bayesian inference.
- Propose methods for privacy protection by selective altering individuals' social network.
- Evaluate the effectiveness of the privacy protection methods via experimental and analytical study.

The paper is organized as follows. In Section II, we introduce the background and related work. In Section III, we explain the target scenarios. In Section IV, we propose a Bayesian network approach to infer personal attributes. In Section V, we discuss three key characteristics of social networks and conduct experiments to investigate their impact on inference result. We propose privacy protection schemes and compare their performance in Section VI, and investigate the effectiveness of these protection methods using analysis in Section VII.

II. Background and Related Work

A. Social Networks

Social network analysis has been conducted in many areas. Milgram's classic paper [7] in 1967 estimates that every person in the world is at most six hops away from one another. The recent success of the Google search engine [8], which applies social network ideas to the Internet, draws great attention on social network analysis again. For instance, Newman [9] reviews the relationship between graph structure and dynamical behavior of large networks. The ReferralWeb project mined social networks from a wide variety of public-available information [3]. An individual's decision to buy products may be influenced by his/her friends, so social network is modeled as a Markov random field to find customers' network value in [2]. In contrast, we map a social network into a Bayesian network, and a person's attribute can be reflected from his/her friends' attributes.

B. Bayesian Networks

A Bayesian network [10]–[12] is a graphic representation of the joint probability distribution over a set of variables. It consists of a network structure and a collection of *conditional probability tables* (CPT). The network structure is represented as a *Directed Acyclic Graph* (DAG) in which each node corresponds to a random variable and each edge indicates a dependent relationship between connected variables. In addition, each variable (node) in a Bayesian network is associated with a CPT, which enumerates the conditional probabilities for this variable, given all the combinations of its parents' value. Thus, for a Bayesian network, the DAG captures causal relations among random variables, and CPTs quantify these relations.

Bayesian networks have been extensively applied to fields such as medicine, image processing, and decision support systems. Since Bayesian networks include the consideration of network structure, we use them as our inference model. Individuals in a social network can be represented as nodes and the relations between individuals can be modelled as edges in Bayesian networks.

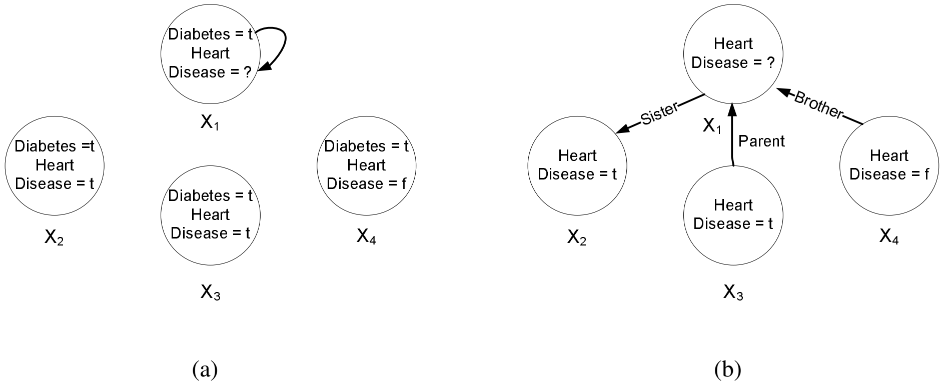


Fig. 1. Illustrator of Social Network Inference (a) Inference across attributes. (b) Inference over relations.

III. Problem Statement

Privacy is a subjective notion. Information such as social security numbers or credit card numbers is considered as privacy by most people. On the other hand, some information, e.g, age, profession and GPA may not be considered as privacy depending on people. The first type of privacy seldom appears in social network, and most privacy disclosure in social networks is related to the second type. In this paper, we focus our discussion on the second type of privacy, and refer to any personal attribute that is not released by its owner as privacy. Specifically, for a given attribute A, we define the value of an individual as t ($true$) if this individual possesses A, or f ($false$) otherwise.

Data mining may be used to predict missing attribute values by finding correlations of attributes. For example, Fig. 1(a) shows a network of four people X_1, X_2, X_3 and X_4. Given that we know person X_1 has diabetes, we want to predict whether X_1 suffers from heart disease or not. From the rest of the network, we learn that two thirds of people having diabetes also have heart disease; therefore, we can predict that there is approximately 67% chance that X_1 also has heart disease.

An alternative solution is to exploit social relations among people to predict their attribute values. Fig. 1(b) shows an example of how such information can be utilized. In this figure, four persons are connected by family relationships. From this social network, we discover that most of X_1's family members have heart disease. Knowing that heart disease is prone to inheritance, we conclude that X_1 has heart disease with a high probability. In this example, prediction is solely based on inference via social relations.

We focus on the latter approach in our study. We are mainly interested in how inference utilizes the knowledge concerning social relations and how privacy protection can be achieved to prevent such inference. Without the loss of generality, we perform single attribute analysis, that is, we study privacy inference and protection which involves only one attribute A. It is worth pointing out that we could also utilize the knowledge from data mining approaches to improve the inference accuracy.

In the real world, people are acquainted with each other via different types of relations, and personal attributes are only sensitive to a specific set of relations. For example, heart disease is sensitive to family relationships but not to officemate relationships. Therefore, to infer individuals' privacy from their social relations, one must be able to differentiate social relations between connected people and consider the characteristics of these relations in the inference.

To simplify the problem, we focus on privacy inference and protection in *homogeneous societies* where people are connected by a single type of social relations, and the attribute A that we study is sensitive to this social relation. Homogeneous networks can be viewed as small closely related groups where people are connected by a relatively pure relationship. In fact, real social networks can be considered as the combination of many homogeneous societies.

In the following sections, we will first investigate how privacy can be inferred in social networks and then study how to protect privacy from being inferred.

IV. Bayesian Inference Via Social Links

To study privacy inference, we propose to use Bayesian networks to model the causal relations among people in social networks. Specifically, if we want to infer the attribute value for a particular individual (referred to as *query node* Z), we will first construct a Bayesian network from Z's social network. We then analyze the Bayesian network and obtain the probability that Z has this attribute.

We start from a simple case in which privacy inference only involves the direct friends of the query node, and then treat the more complex case where attribute values from friends at multiple hops away are considered.

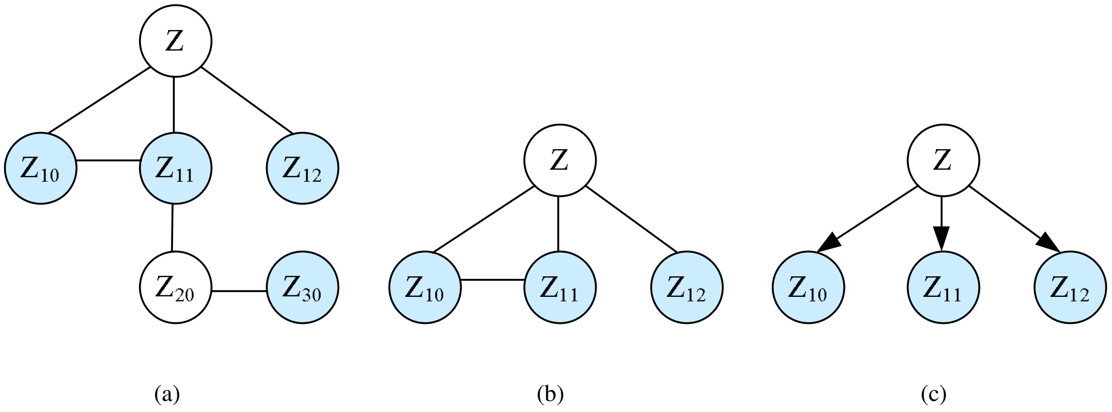


Fig. 2. Reduction of a social network (a) into a Bayesian network to infer Z from his friends via Localization assumption (b) and via Naive Bayesian Assumption (c). The shaded nodes represent friends whose attribute values are known.

A. *Single Hop Inference*

Let us consider the case in which we know the attribute values for *all* the direct friends of the query node Z. We define Z_{ij} as the jth friend of Z at i hops away. If a friend can be reached via more than one route from Z, we use the depth of the shortest path as the value of i. Let Z_i be the set of Z_{ij} ($0 \leq j < n_i$), where n_i is the number of Z's friends at i hops away. For instance, $Z_1 = \{Z_{10}, Z_{11}, ..., Z_{1(n_1-1)}\}$ is the set of Z's direct friends who are one hop away. Note that Z can also be represented as Z_{00}.

An example of a social network with six friends is shown in Fig. 2(a). In this figure, Z_{10}, Z_{11} and Z_{12} are direct friends of Z. Z_{20} and Z_{30} are the direct friends of Z_{11} and Z_{20} respectively. In this scenario, the attribute values of Z_{10}, Z_{11}, Z_{12} and Z_{30} are known (represented as shaded nodes).

1) Bayesian Network Construction: To facilitate the construction of the Bayesian network, we make two assumptions.

Intuitively, our direct friends have more influence on us than friends who are two or more hops away. We assume that it is sufficient to consider only the attribute values of direct friends Z_1 to infer Z's attribute. Once all the attribute values of Z_1 are known, knowing the attribute values of friends at multiple hops away provides no additional information for predicting Z's attribute. Formally, we state this assumption as follows.

Localization Assumption Given the attribute values of the direct friends Z_1 of the query node Z, friends at more than one hop away (i.e., Z_i for $i > 1$) are conditionally independent of Z.

Based on this assumption, Z_{20} and Z_{30} in Fig. 2(a) can be pruned, and the inference of Z only involves Z_{10}, Z_{11} and Z_{12} (Fig. 2(b)). Then the next question is how to decide a DAG linking the remaining nodes. If the resulting social network does not contain cycles, a Bayesian network can be obtained immediately. Otherwise, we must employ more sophisticated techniques to remove cycles, such as the use of auxiliary variables to capture non-causal constraints (*exact conversion*) and the deletion of edges with the weakest relations (*approximation conversion*). We adopt the latter approach and make a *Naive Bayesian* Assumption. That is, the attribute value of Z influences that of Z_{1j} ($0 \leq j < n_1$), and there is a direct link pointing from Z to each Z_{1j}. By making this assumption, we consider the inference paths from Z to Z_{1j} as the *primary* correlations, and disregard the correlations among the nodes in Z_1. Formally, we have:

Naive Bayesian Assumption Given the attribute value of the query node Z, the attribute values of direct friends Z_1 are conditionally independent of each other.

This Naive Bayesian model has been used in many classification/prediction applications including textual-document classification. Though it simplifies the correlation among variables, this model has been shown to be quite effective [13]. Thus, we adopt this assumption in our study. In Fig. 2(c), we obtain a final DAG by removing the connection between Z_{10} and Z_{11} in Fig. 2(b).

2) Bayesian Inference: After modelling Z's social network into a Bayesian network, we use the Bayes Decision Rule to predict the attribute value of Z. For a general Bayesian network with maximum depth i, let the value for Z, $\bar{z}$, be the attribute value with the maximum conditional probability given the observed attribute values of other nodes in the network (i.e., the maximum posterior probability):

$$\bar{z} = \arg\max_{z} P(Z = z \mid Z_1, Z_2, ..., Z_i) \qquad z \in \{t, f\}. \tag{1}$$

Since single hop inference involves only direct friends Z_1 which are independent of each other, the posterior probability can be further reduced using the conditional independence encoded in the Bayesian network:

$$\begin{aligned} P(Z=z \mid Z_1) &= \frac{P(Z_1 \mid Z=z)\cdot P(Z=z)}{\sum_z[P(Z_1 \mid Z=z)\cdot P(Z=z)]} \\ &= \frac{P(Z=z)\cdot \prod_{j=0}^{n_1-1} P(Z_{1j}=z_{1j} \mid Z=z)}{\sum_z[P(Z=z)\cdot \prod_{j=0}^{n_1-1} P(Z_{1j}=z_{1j} \mid Z=z)]}, \end{aligned} \tag{2}$$

where z and z_{1j} are the attribute values of Z and Z_{1j} respectively ($0 \leq j < n_1$, $z, z_{1j} \in \{t, f\}$) and the value of each z_{1j} is known.

To compute (2), we need to further learn the Conditional Probability Table (CPT) for each $P(Z_{1j} \mid Z)$. Since the attribute values for each pair of friends are fixed, so it is infeasible to learn their CPT. Therefore we apply the *parameter estimation* [12] technique on the whole sample network. For every pair of parent X and child Y, we obtain:

$$P(Y=y \mid X=x) = \frac{\text{\# of friendship links connecting people with } X=x \text{ and } Y=y}{\text{\# of friendship links connecting a person with } X=x} \tag{3}$$

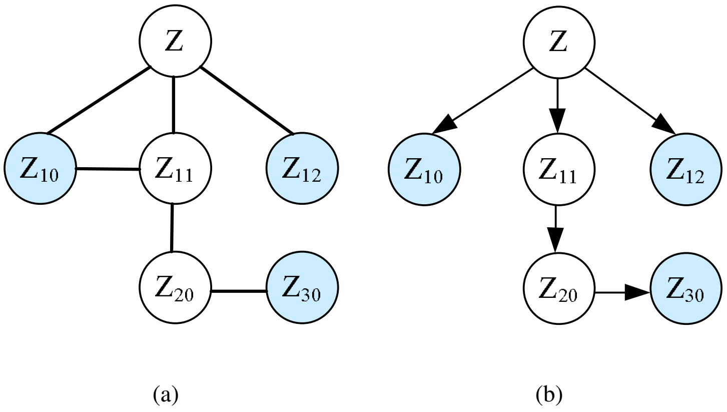
where x, $y \in \{t, f\}$. $P(Y=y \mid X=x)$ is the CPT for every pair of friends Z_{1j} and Z in the network. Since $P(Z_{1j} \mid Z)$ is the same for $0 \leq j < n_1$, Z_{1j} becomes equivalent to one another, and the posterior probability now depends on N_{1t}, the number of direct friends with attribute value t, rather than individual attribute values. We can rewrite the posterior probability $P(Z=z \mid Z_1)$ as $P(Z=z \mid N_{1t}=n_{1t})$. If $N_{1t}=n_{1t}$, we obtain:

$$P(Z=z \mid N_{1t}=n_{1t}) = \frac{P(Z=z)\cdot P(Z_{10}=t \mid Z=z)^{n_{1t}} \cdot P(Z_{10}=f \mid Z=z)^{n_1-n_{1t}}}{\sum_z[P(Z=z)\cdot P(Z_{10}=t \mid Z=z)^{n_{1t}} \cdot P(Z_{10}=f \mid Z=z^{n_1-n_{1t}}]}. \tag{4}$$

After obtaining $P(Z=t \mid N_{1t}=n_{1t})$ and $P(Z=f \mid N_{1t}=n_{1t})$, we predict Z has attribute value t if the former value is greater than the latter value, and vice versa.

B. Multiple Hops Inference

In single hop inference, we assume we know the attribute values of all the direct friends of Z. However, in reality, those attribute values may not be observed since people hide their sensitive information, and the Localization Assumption in the previous section is no longer valid. To incorporate more attribute information into our Bayesian network, we propose a *generalized localization assumption*.



(a)

(b)

Fig. 3. Reduction of a social network (a) into a Bayesian network to infer Z from his friends via Generalized Localization assumption (b). The shaded nodes represent friends whose attribute values are known.

Generalized Localization Assumption Given the attribute value of the jth friend of Z at i hops away, Z_{ij} ($0 \leq j < n_i$), the attribute of Z is conditionally independent of the descendants of Z_{ij}.

This assumption states that if the attribute value of Z's direct friend Z_{1j} is unknown, then the attribute value of Z is conditionally dependent on the attribute values of the direct friends of Z_{1j}. This process continues until we reach a descendent of Z_{1j} whose attribute value is known. For example, the network structure in Fig. 3(a) is the same as in Fig. 2(a), but the attribute value of Z_{11} is unknown. Based on the Generalized Localization Assumption, we extend the network by branching to Z_{11}'s direct child Z_{20}. Since Z_{20}'s attribute is unknown, we further branch to Z_{20}'s direct friend Z_{30}. The branch terminates here because the attribute of Z_{30} is known. Thus, the inference network for Z includes all the nodes in the graph. After applying Naive Bayesian assumption, we obtain the DAG shown in Fig. 3(b). Similar to single hop inference, the resulting DAG in multiple hops inference is also a tree rooted at the query node Z. One interpretation of this model is that when we predict the attribute value of Z, we always treat him/her as an egocentric person who have strong influences on his/her friends. Thus, the attribute value of Z can be reflected by those of his/her friends.

For multiple hops inference, we still apply the Bayes Decision Rule. Due to additional unknown attribute values such as Z_{11}, the calculation of the posterior probability becomes more complicated. One common technique to solve this equation is variable elimination [14]. In this paper, we use this technique to derive the value of $\bar{z}$ in (1).

V. Experimental Study of Bayesian Inference

In this section, we investigate to what extent privacy can be inferred by Bayesian inference. We first introduce three characteristics of social networks: influence strength, prior probability, and society openness. Then we evaluate their impacts on inference accuracy in a real social network structure. *Inference accuracy* is defined as the percentage of nodes predicted correctly by inference.

A. Characteristics of Social networks

1) Influence Strength: Analogous to the interaction between inheritance and mutation in Biology, we define two types of influence in social relations. Particularly, for relationship between every pair of parent X and child Y, we define $P(Y = t \mid X = t)$ (or $P_{t|t}$ for simplification) as *inheritance strength.* This value measures the degree that a child inherits an attribute from his/her parent. A higher value of $P_{t|t}$ implies both X and Y will possess the attribute with a higher probability. On the other hand, we define $P(Y = t \mid X = f)$ (or $P_{t|f}$) as *mutation strength.* $P_{t|f}$ measures the potential that Y develops his/her attribute by mutation rather than inheritance. An individual's attribute value is the result of both types of strength.

There are two other conditional probabilities between X and Y, i.e., $P(Y = f \mid X = t)$ (or $P_{f|t}$) and $P(Y = f \mid X = f)$ (or $P_{f|f}$). These two values can be derived from $P_{t|t}$ and $P_{t|f}$ respectively ($P_{f|t}$ = 1 - $P_{t|t}$ and $P_{f|f}$ = 1 - $P_{t|f}$). Therefore, it is sufficient to only consider inheritance and mutation strength.

2) Prior Probability: Prior probability $P(Z = t)$ (or P_t in short) is the percentage of people in the social network who possess attribute A. When no additional information is provided, we could use prior probability to naively predict attribute values for the query nodes: if $P_t \geq 0.5$, we predict that every query node has value t; otherwise, we predict that it has value f. We call this method *naive inference*. The average naive inference accuracy that can be obtained is $max(P_t, 1 - P_t)$. In our study, we use it as a reference to compare with our Bayesian inference approach.

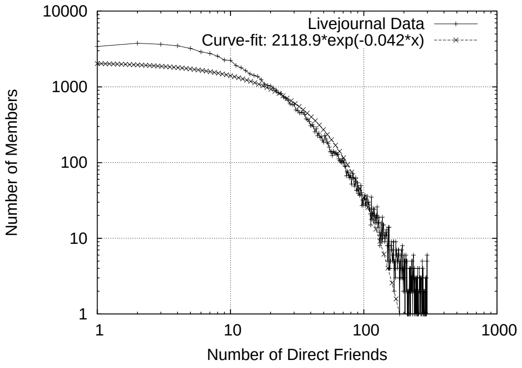


Fig. 4. Number of direct friends vs. number of members in Livejournal

It is worth pointing out that when $P_{t|t}$ is equal to P_t, people in a society are indeed independent of each other. Hence, knowing extra evidence of a friend provides no contribution to the prediction for the query node.

3) Society Openness: We define society openness O_A as the percentage of people in a society who release their values of attribute A. The more people release their values, the higher the society openness is, and the more information about attribute A could be observed. We want to examine the accuracy of Bayesian inference for various society openness.

B. Data Set

For the experiment, we collect $66,766$ personal profiles from an online weblog service provider Livejournal [15], which owns 2.6 million active members in the world. For each member, Livejournal generates a personnel profile which specifies the member's personal information as well as the URLs for the profiles of this member's friends. Among the collected profiles, there are $4,031,348$ friend relationships. The degree of the number of friends follows the power law distribution (Fig. 4). About half of the population have less than 10 direct friends.

In order to evaluate the inference behaviors for a wide range of parameters, we use a hypothetical attribute and synthesize the attribute values: for each member, we assign a CPT and determine the actual attribute value based on the parent's value and the assigned CPT. The attribute assignment starts from

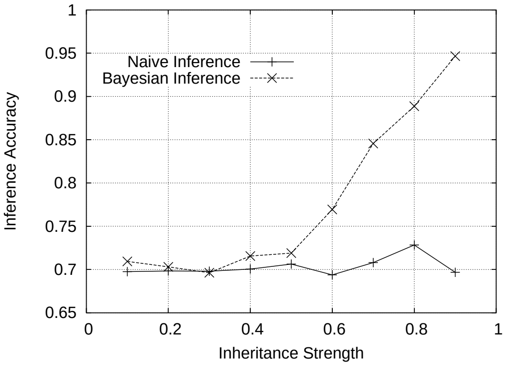


Fig. 5. Inference accuracy of Bayesian vs. naive inference when $P_t = 0.3$.

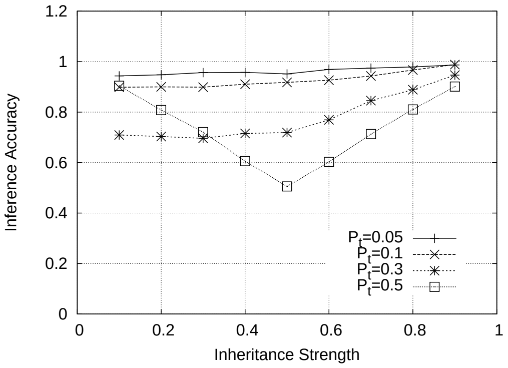


Fig. 6. Inference accuracy of Bayesian inference for different prior probabilities.

the set of nodes whose in-degree is 0 and explores the rest of the network following friendship links. We use the same CPT for each member. For all the experiments, we evaluate the inference performance by varying CPTs.

After the attribute assignment, we obtain a social network. To infer each individual, we build a corresponding Bayesian network, and then conduct Bayesian inference as described in Section IV.

C. Experimental Results

1) Comparison of Bayesian and Naive Inference: In the first set of experiments, we compare the performance of Bayesian inference with naive inference. We want to study whether privacy can be deduced from social relations. We fix the prior probability P_t to 0.3 and vary inheritance strength $P_{t|t}$ from 0.1 to 0.9. [1] We perform inference using both approaches on every node in the network, and obtain the inference accuracy by comparing predicted values with their actual values. Fig. 5 shows the inference accuracy of the two methods as $P_{t|t}$ increases. It is clear that Bayesian inference outperforms naive inference. The curve for naive inference fluctuates around 70%, because with $P_t = 0.3$, the average accuracy we can achieve is 70%. The performance of Bayesian inference varies with $P_{t|t}$. We achieve a very high accuracy, especially at high inheritance strength. The accuracy even reaches 95% when $P_{t|t} = 0.9$, which is much

[1] In an equilibrium state, the value of $P_{t|f}$ can be derived from P_t and $P_{t|t}$. When P_t is fixed, increasing $P_{t|t}$ results in a decrease in $P_{t|f}$.

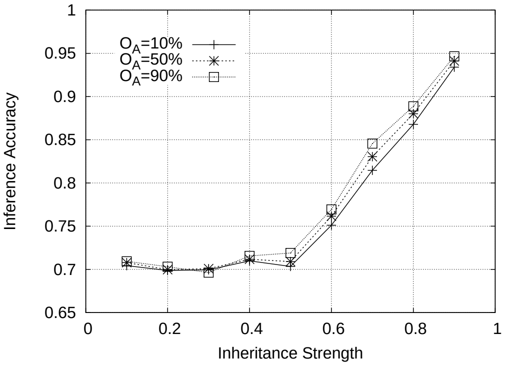


Fig. 7. Inference accuracy of Bayesian inference for different society openness.

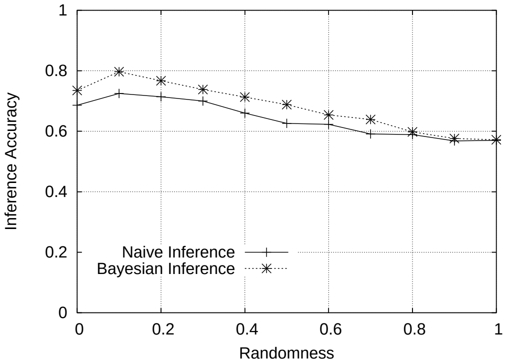


Fig. 8. Robustness study when $P_t = 0.3$ and $P_{t|t} = 0.7$.

higher than the 70% accuracy of the naive inference. We observed a similar trend between these two methods for other prior probabilities as well.

2) Effect of Influence Strength and Prior Probability: Fig. 6 shows the accuracy of Bayesian inference when the prior probability P_t is 0.05, 0.1, 0.3 and 0.5, and the inheritance strength $P_{t|t}$ varies from 0.1 to 0.9. As P_t varies, the inference accuracy yields different trends with $P_{t|t}$. The lowest inference accuracy always occurs when $P_{t|t}$ is equal to P_t. For example, the lowest inference accuracy (approximately 70%) at $P_t = 0.3$ is achieved when $P_{t|t}$ is 0.3. At this point, people in the network are actually independent of one another. Furthermore, the higher the difference between $P_{t|t}$ and P_t, the stronger the influence of parent on children, and the better Bayesian inference performs.

3) Society Openness: In the previous experiments, we assume that society openness is 100%. That is, the attribute values of all the friends of the query node are known. In this set of experiments, we study the inference behavior at different levels of society openness. We randomly hide the attributes of a certain percentage of members, ranging from 10% to 90%, and then perform Bayesian inference on those nodes.

Fig. 7 shows the experimental results for the prior probability $P_t = 0.3$ and the society openness $O_A = 10\%$, 50% and 90%. The inference accuracy decreases as more members hide their attributes. For instance, at inheritance strength 0.7, when the openness is decreased from 90% to 10%, the accuracy reduces from 84.6% to 81.5%. However, the reduction in inference accuracy is relatively small (on average less than 5%). We also observe similar trends for other prior probabilities. This phenomenon reveals that randomly hiding friends' attributes only has a small effect on the result of Bayesian inference. Therefore, an effective privacy protection method should consider selectively altering social networks.

4) Robustness: In this study, we evaluate the robustness of the performance of Bayesian inference when people provide false information. In this set of experiments, we control the percentage of members who randomly select their attribute values (which is referred to as "randomness") from 0% to 100%. Fig. 8 shows the impact of randomness on the inference accuracy at prior probability $P_t = 0.3$ and inheritance strength $P_{t|t} = 0.7$. At low randomness, Bayesian inference has higher accuracy than naive inference. For example, when the randomness is 0.1, the inference accuracy of Bayesian and naive inference is 79.7% and 72.9% respectively. As the randomness increases, the advantage of Bayesian inference gradually diminishes. Especially, when the randomness is 1.0, there is almost no difference in the inference accuracy between these two methods. This is because when people randomly negate their attribute values, their attribute values no longer follow the causal relations between friends. Thus, Bayesian inference behaves similarly to naive inference. This phenomenon provides an insight that falsifying personal attribute might be an effective method for privacy protection. In Section VI, we will propose several schemes for privacy protection and evaluate their effectiveness.

VI. Privacy Protection

The above study shows that privacy can be inferred from social relations. To prevent such inference, we consider altering an individual's social network. Basically, we could alter two things: the individual's social relations (i.e., links) and personal attributes of the individual's friends (i.e., nodes). For social relations, we can either hide existing social relations or add fraudulent ones. For friends' personal attributes, we can either hide or falsify their values. Our study on society openness suggests that random changes on a social network only have a small effect on the result of Bayesian inference. Therefore, an effective protection method requires choosing appropriate candidates for alteration.

In this section, we study privacy protection schemes. We first present a theorem which captures the causal effect between friends' attribute values in a chain topology. We then apply this theorem to our protection schemes. We conduct experiments on the Livejournal network structure and evaluate the performance of the proposed protection schemes.

A. Causal Effect between Friends' Attribute Values

As mentioned earlier, children's attribute values are the result of the interaction between the inheritance strength $P_{t|t}$ and the mutation strength $P_{t|f}$ with their parents. Intuitively, in a family, when the inheritance strength is stronger than the mutation strength, children tend to inherit the same attribute from their parents; thus, the evidence of a child having this attribute increases our belief that his/her parent has the same attribute. On the contrary, when the inheritance strength is weaker than the mutation strength, parents and children are more likely to have opposite attribute values, and the evidence of a child having an attribute reduces our belief of his/her parent having the same attribute. Inspired by this observation, we derive a theorem to quantify the causal effects between friends' attribute values.

Theorem: Casual Effect between Friends' Attribute Values in a Chain Network

Given a chain topology social network, let Z be the query node, Z_{n0} be Z's descendant at n hops away. Assuming the attribute value of Z_{n0} is the only evidence observed in this chain, and the prior probability P_t satisfies $0 < P_t < 1$, we have $P(Z = t \mid Z_{n0} = t) > P(Z = t)$ *iff* $(P_{t|t} - P_{t|f})^n > 0$, and $P(Z = t \mid Z_{n0} = f) > P(Z = t)$ *iff* $(P_{t|t} - P_{t|f})^n < 0$, where $P_{t|t}$ and $P_{t|f}$ are the inheritance strength and mutation strength of the network respectively.

Proof: see Appendix.

This theorem indicates that, when $P_{t|t} > P_{t|f}$, the posterior probability $P(Z = t \mid Z_{n0} = t)$ is greater than the prior probability $P(Z = t)$. Under this condition, the evidence of $Z_{n0} = t$ always increases our prediction for $Z = t$. We can use an example to explain this intuitively. If the inheritance strength is stronger than the mutation strength for having cancer in a family, it is very likely that cancer will be carried out to the next generation. If we observe a descendant in a family having cancer, we can predict that his ancestors have cancer with a higher probability . However, if this descendant hides his record of cancer, our prediction for his ancestors will decrease.

When $P_{t|t} < P_{t|f}$, i.e., the mutation strength is stronger than the inheritance strength, $(P_{t|t} - P_{t|f})^n < 0$ and $P(Z = t \mid Z_{n1} = t) > P(Z = t)$ only happens when n, the depth of current node, is odd. When n is even, $P(Z = t \mid Z_{n1} = t) < P(Z = t)$. This means when $P_{t|t} < P_{t|f}$, whether the evidence of $Z_{n1} = t$ increases our prediction for $Z = t$ also depends on the value of n. We use the same example as above, assuming now the mutation strength is stronger than the inheritance strength. In this case, the next generation may develop cancer by mutation instead of inheritance. That is, if a parent has no cancer, his child is more likely to have it because of the strong mutation strength. If a parent has cancer, there is a high tendency that his child will not have it because of the weak inheritance strength. Therefore, we can observe that the cancer records alternate across generations in this family. Correspondingly, the observation that a child has cancer (n=1) will decrease our prediction for his/her parent having cancer, whereas, the evidence of cancer in a grandchild (n=2) will increase our prediction for his/her grandfather.

B. A Privacy Protection Rule

The theorem states that the evidence of a friend having attribute value t when $(P_{t|t} - P_{t|f})^n > 0$ or having attribute value f when $(P_{t|t} - P_{t|f})^n < 0$ will increase the posterior probability of the query node. Based on this theorem, we propose the following protection rule. *For reducing others' belief that the query node has the attribute value t, we should: 1) hide or falsify the attribute values of friends who satisfy the above conditions, 2) hide relationships to friends who satisfy the above conditions, or add fraudulent relationships to friends who do not satisfy. On the contrary, for misleading people to believe the query node possesses the attribute, we can apply these techniques in an opposite way.*

Based on the protection rule, we propose the following protection schemes:

1) Selectively Hiding Attribute (SHA). SHA hides appropriate friends' attribute value base on the protection rule.

2) Selectively Falsifying Attribute (SFA). In this scheme, we falsify friends' attribute values according to the protection rule. Because of the Localization Assumption in Section IV-A, we do not need to make further changes to friends at multiple hops away even if the protection fails.
3) Selectively Hiding Relationships (SHR). Instead of hiding people's attribute as in previous schemes, we hide the relationship between the query node and selected direct friends. When all the friend relationships of this individual are hidden, the individual becomes a singleton, and the prediction will be made based on the prior probability.
4) Selectively Adding Relationships (SAR). Opposite to hiding relationships in SHR, SAR adds fraudulent relationships with people whose attribute values could mislead the inference based on the protection rule.

C. Experimental Study of Privacy Protection Performance

In this section, we conduct a set of controlled experiments to evaluate different schemes for privacy protection. To protect each individual's attribute, we continuously alter this individual's social network until protection succeeds when the predicted attribute value becomes opposite to its original prediction, or until protection fails if no more alteration can be made.

In the experiments, we compare the proposed protection schemes with Randomly Hiding Attribute (RHA) scheme. In RHA, we do not follow the protection rule. We randomly select a friend and hide his/her attribute value in each step. We start with the direct friends. If an attribute value can still be correctly inferred after we hide all the direct friends' attribute values, we continue to hide attribute values of friends at two hops away and so on, until the protection either succeeds or fails.

We conduct experiments on 3000 nodes in the Livejournal data set and compare their performance. For each node, we apply the above five schemes. The performance metrics are the percentage of successful protection and the average number of alterations needed to protect privacy.

Fig. 9 illustrates the percentage of successful protection for different $P_{t|t}$ values when $P_t = 0.3$. From this figure, it is clear that the effectiveness of these schemes is in the order of SAR $>$ SFA $>$ SHR $>$

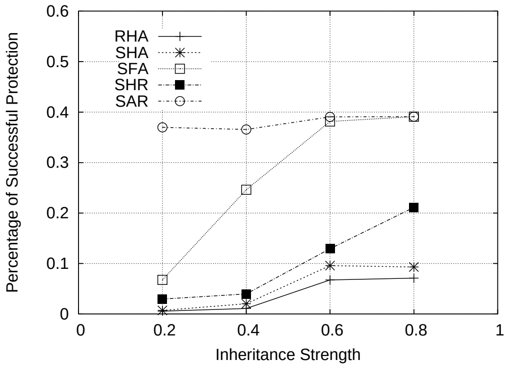


Fig. 9. Percentage of successful privacy protections when $P_t = 0.3$.

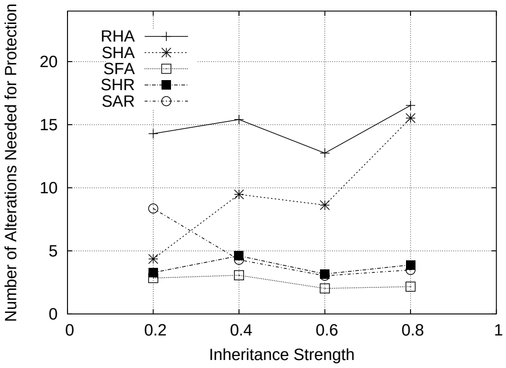


Fig. 10. Number of changes needed for privacy protection when $P_t = 0.3$.

SHA $>$ RHA, from the most effective to the least effective. Since SAR could add non-existing friend relationships, the upper bound of the number of alterations is the size of the entire network. In other schemes, the maximum numbers of alterations are the number of descendants (e.g., RHA, SHA) and the number of direct friends (e.g., SFA, SHR). SAR has the highest upper bound in terms of the number of possible alterations and thus provides more room for protection. The performance of SFA and SHR follows SAR. In fact, we can think of SFA as a special combination of SHR and SAR, i.e., hiding a friend relationship followed by adding a fraudulent relationship. Therefore, the performance of SFA is better than SHR. SHA is worse than the above schemes because friends at multiple hops away still leave clue for Bayesian inference. RHA does not follow the protection rule, thus yielding the worst performance.

The average number of alterations needed for each protection scheme is shown in Fig. 10. Schemes which require fewer alterations to achieve protection is regarded as more effective. Based on this metric, RHA is again the worst. SFA is the best among the five schemes. The average number of alterations of SHR and SAR is close to each other in most cases. Note that the average number of changes of SAR is high at $P_{t|t} = 0.2$ (which is 8.36). At this point, SAR protects many cases that other schemes cannot protect by adding a relatively large number of fraudulent friend relationships. Finally, SHA is better than RHA, but not as good as the other schemes.

Both figures demonstrate the effectiveness of friend selection based on the protection rule. Furthermore, among the four schemes which consider friend selection, SFA is the best one. It can protect many cases while reducing the number of alterations to the original social network.

VII. Discussions on Friend Selection

In the previous section, we demonstrate that selective social network alterations based on the protection rule are more effective than such an alteration as RHA which does not follow the protection rule. In this section, we shall analytically compare the difference between randomly hiding friend relationships (denoted as RHR) and selectively hiding friend relationships (abbreviated as SHR) through analysis. Specifically, we study the frequency of posterior probability variation after hiding friend relationships. The hiding scheme is considered to be more effective in preventing privacy from being inferred if the posterior probability varies significantly towards the desired value.

A. Randomly Hiding Friend Relationships

Hiding friend relationships means removing direct friends of the query node. The social network can be represented as a two-level tree with the query node Z as the root and n_1 direct friends Z_{10}, ..., $Z_{1(n_1-1)}$ as leaves. We want to derive the probability distribution of the posterior probability variation due to randomly hiding friend relationships, i.e., the difference of the posterior probability and the probability that this difference occurs.

Let random variables N_{1t} and N'_{1t} be the number of friends having attribute value t before and after hiding h friend relationships, where $0 \leq h \leq n_1$ and $max(0, N_{1t} - h) \leq N'_{1t} \leq min(N_{1t}, n_1 - h)$. If $N_{1t} = n_{1t}$ and $N'_{1t} = n'_{1t}$, we can compute the posterior probabilities $P(Z = t \mid N_{1t} = n_{1t})$ and $P(Z = t \mid N'_{1t} = n'_{1t})$ from (4) respectively. Therefore, the posterior probability variation caused by hiding h friend relationships is:

$$\Delta P(Z = t \mid N_{1t} = n_{1t}, N'_{1t} = n'_{1t}) = |P(Z = t \mid N_{1t} = n_{1t}) - P(Z = t \mid N'_{1t} = n'_{1t})|. \tag{5}$$

Now we want to derive the probability that each possible value of $\Delta P(Z = t \mid N_{1t} = n_{1t}, N'_{1t} = n'_{1t})$ occurs. In other words, we want to compute the probability of the joint event $N_{1t} = n_{1t}$ and $N'_{1t} = n'_{1t}$

(before and after hiding friend relationships), which is equal to:

$$P(N_{1t} = n_{1t}, N'_{1t} = n'_{1t}) = P(N_{1t} = n_{1t}) \cdot P(N'_{1t} = n'_{1t} \mid N_{1t} = n_{1t}). \tag{6}$$

Initially, if we know Z's attribute value is z ($z \in \{t, f\}$), the probability that $N_{1t} = n_{1t}$ follows the Binomial distribution:

$$\begin{aligned} P(N_{1t} = n_{1t} \mid Z = t) &= \binom{n_1}{n_{1t}} \cdot P_{t|t}^{n_{1t}} \cdot P_{f|t}^{n_1 - n_{1t}} \\ P(N_{1t} = n_{1t} \mid Z = f) &= \binom{n_1}{n_{1t}} \cdot P_{t|f}^{n_{1t}} \cdot P_{f|f}^{n_1 - n_{1t}}. \end{aligned} \tag{7}$$

By unconditioning on Z, we obtain:

$$P(N_{1t} = n_{1t}) = P(Z = t) \cdot P(N_{1t} = n_{1t} \mid Z = t) + P(Z = f) \cdot P(N_{1t} = n_{1t} \mid Z = f). \tag{8}$$

We define h_t and h_f as the numbers of removed nodes (i.e., hidden friend relationships) with attribute t and f, respectively ($h_t = n_{1t} - n'_{1t}$ and $h_f = h - h_t$). Then we can compute the conditional probability that $N'_{1t} = n'_{1t}$ given $N_{1t} = n_{1t}$ as:

$$P(N'_{1t} = n'_{1t} \mid N_{1t} = n_{1t}) = \frac{\binom{n_{1t}}{h_t} \cdot \binom{n_1 - n_{1t}}{h_f}}{\binom{n_1}{h}}. \tag{9}$$

In this equation, the numerator represents the number of ways to remove h_t nodes with value t and h_f nodes with value f, and the denominator represents the number of combinations to choose any h nodes from a total of n_1 nodes.

Substituting (8) and (9) into (6), we obtain $P(N_{1t} = n_{1t}, N'_{1t} = n'_{1t})$.

B. Selectively Hiding Friend Relationships

We perform a similar analysis for selectively hiding friend relationships in a two-level tree. Unlike random selection which randomly selects nodes with attribute values t or f, this method follows the protection rules and selects all the nodes with the same attribute values to hide. Thus, we can compute $\Delta P(Z = t \mid N_{1t} = n_{1t}, N'_{1t} = n'_{1t})$ in the same way as above. However, the distribution of posterior probability variation needs to computed differently.

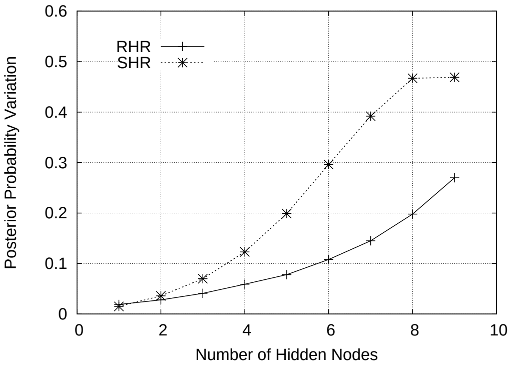


Fig. 11. Average posterior probability variation for selectively and randomly hiding friend relationships.

Given h, the number of nodes to remove, n'_{1t} is deterministic. For example, if we remove nodes with attribute t, then $n'_{1t} = m - h$; otherwise $n'_{1t} = m$. Consequently, in the former case,

$$P(N_{1t} = n_{1t}, N'_{1t} = n'_{1t}) = \begin{cases} P(N_{1t} = n_{1t}), & \text{if } n'_{1t} = m - h \\ 0, & \text{otherwise} \end{cases}, \tag{10}$$

whereas in the latter case,

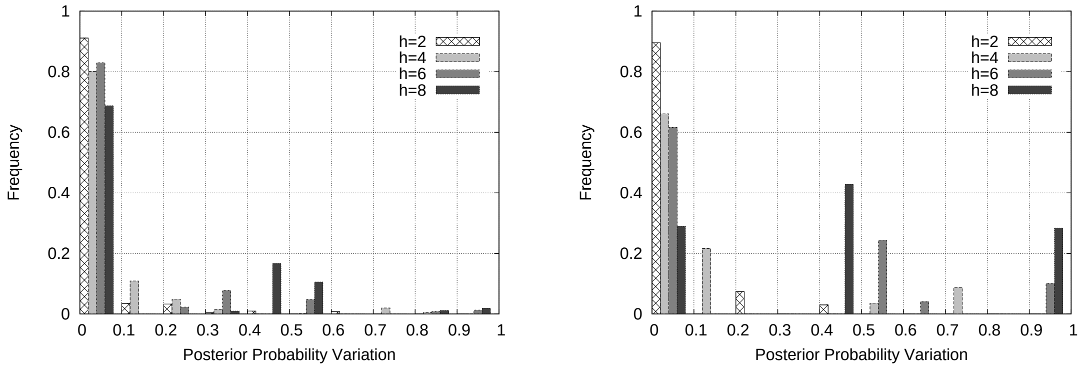
$$P(N_{1t} = n_{1t}, N'_{1t} = n'_{1t}) = \begin{cases} P(N_{1t} = n_{1t}), & \text{if } n'_{1t} = m \\ 0, & \text{otherwise} \end{cases}. \tag{11}$$

where $P(N_{1t} = n_{1t})$ can be obtained from (8).

C. Comparison of Randomly vs. Selectively Hiding Friend Relationships

We first compute the average variation in posterior probability for both RHR and SHR as shown in Fig. 11. We fix n_1 to be 10 and vary h from 1 to 9. The x-axis is the number of friends that we hide, and the y-axis is the posterior probability variation calculated based on (5). Clearly, SHR has higher posterior probability variation than RHR, especially for the case of a large number of hidden friends.

We also plot the histogram of the posterior probability variation $\Delta P(Z = t \mid N_{1t} = n_{1t}, N'_{1t} = n'_{1t})$. We divide the range of posterior probability variation into 10 equal width intervals. Then we compute the probability that the posterior probability variation falls in each interval.



(a)

(b)

Fig. 12. Frequency of posterior probability variation for (a) randomly hiding friend relationships (b) selectively hiding friend relationships.

Fig. 12 shows the histogram of the posterior probability variation for RHR and SHR respectively, when the prior probability is 0.3 and the influence strength is 0.7. The x axis represents the intervals and the y axis represents the frequency of the posterior probability variation within the interval. The frequency is derived from (6) for RHR and from (10) and (11) for SHR. For SHR, we remove friends with attribute value t. The maximum number of removed friends k cannot exceed n_{1t}. As a result, we do not consider the cases when $n_{1t} < k$, and we normalize the frequency results for selectively hiding friends based on the overall probability that $n_{1t} \geq k$.

For RHR, we observe that the variation is less than 0.1 for 70% to 90% of the cases in Fig. 12(a). Thus, the posterior probability is unlikely to be varied greatly. In contrast, the posterior probability variation in Fig. 12(b) is widely distributed, which means there are noticeable changes in the posterior probability after hiding nodes selectively. This trend is more pronounced with increasing the number of removed friends. For example, when $h = 8$, the frequency that the variation lies between 0.9 and 1.0 is about 28.8% as compared to 1.9% in Fig. 12(a). These results show the effectiveness of using the protection rule for privacy protection.

VIII. Conclusion

In this paper, we study privacy inference and protection in social networks. Using the Bayesian networks to model the causal relations among friends in social networks, we showed that privacy may be indirectly

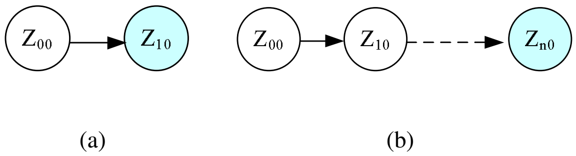


Fig. 13. The chain network structure (a) The query node with one descendant; (b) The query node with n descendants.

released via their social relations, and the inference accuracy of privacy information is closely related to the influence strength between friends. We found that Bayesian network is effective to capture the causal relations in a social network. Based on both the interaction between inheritance strength and mutation strength and the network structure, a protection rule is developed to perform selective social network alterations. Our experimental and analytical study reveals that using the protection rule is far more effective in privacy protection than random social network alterations. Further, falsifying friend attributes is the most effective among all the proposed methods.

We have focused our study on the impact of social relations on privacy disclosure and protection. We also need to consider the case of social relations with multiple attributes. We plan to investigate such complex social networks in the future.

APPENDIX I

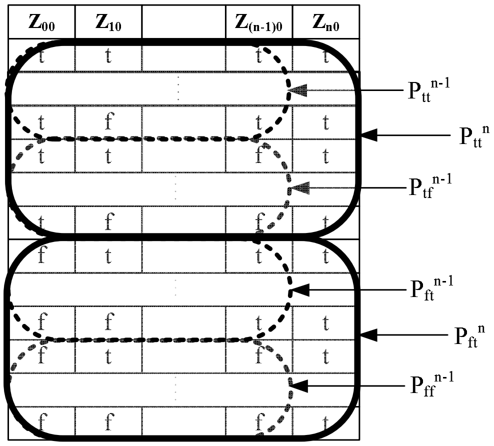
THEOREM PROOF

Theorem: Casual Effect between Friends' Attribute Values in a Chain Network

Given a chain topology, let Z be the query node, Z_{n0} be Z's descendant at n hops away. Assuming the attribute value of Z_{n0} is the only evidence observed in this chain, and the prior probability P_t satisfies $0 < P_t < 1$, we have $P(Z = t \mid Z_{n0} = t) > P(Z = t)$ *iff* $(P_{t|t} - P_{t|f})^n > 0$, and $P(Z = t \mid Z_{n0} = f) > P(Z = t)$ *iff* $(P_{t|t} - P_{t|f})^n < 0$, where $P_{t|t}$ and $P_{t|f}$ are the inheritance strength and mutation strength of the network respectively.

Proof:

Let us consider a chain topology shown in Fig. 13. The query node Z_{00} is the root node and each descendant (except the last one) has exactly one child. Consider the simplest example when $n = 1$ (i.e.,



Z_{00}	Z_{10}		$Z_{(n-1)0}$	Z_{n0}
t	t		t	t
		⋮		
t	f		t	t
t	t		f	t
		⋮		
t	f		f	t
f	t		t	t
		⋮		
f	f		t	t
f	t		f	t
		⋮		
f	f		f	t

Fig. 14. CPT for nodes in Fig. 13(b)

the query node Z has only one direct child Z_{10}) as shown in Fig. 13(a). In this example, the attribute value of Z_{10} is known.

Assuming $Z_{10} = t$, from (1), the posterior probability $P(Z_{00} = t \mid Z_{10} = t)$ is:

$$\begin{aligned} P(Z_{00} = t \mid Z_{10} = t) &= \frac{P(Z_{00} = t) \cdot P(Z_{10} = t \mid Z_{00} = t)}{P(Z_{00} = t) \cdot P(Z_{10} = t \mid Z_{00} = t) + P(Z_{00} = f) \cdot P(Z_{10} = t \mid Z_{00} = f)} \\ &= \frac{P_t \cdot P_{t|t}}{P_t \cdot P_{t|t} + (1 - P_t) \cdot P_{t|f}}. \end{aligned} \tag{12}$$

Thus,

$$\begin{aligned} P(Z_{00} = t \mid Z_{10} = t) > P(Z_{00} = t) &\Leftrightarrow \frac{P_t \cdot P_{t|t}}{P_t \cdot P_{t|t} + (1 - P_t) \cdot P_{t|f}} > P_t \\ &\Leftrightarrow P_{t|t} - P_{t|f} > 0 \text{ for } P_t \neq 1. \end{aligned} \tag{13}$$

Similarly, when $Z_{10} = f$, we can prove $P(Z_{00} = t \mid Z_{10} = f) > P(Z_{00} = t)$ iff $P_{t|t} - P_{t|f} < 0$ for $P_t \neq 1$.

Now we extend this example to show how attribute value of a node at depth n affects the prediction for Z. In Fig. 13(b), we show a network of $n + 1$ nodes. In this figure, only Z_{n0}, Z's descendent at depth n, has a known value. Fig. 14 shows the corresponding conditional probability table for these $n + 1$ nodes.

Let P^n_{tt}, P^n_{ft}, P^n_{tf} and P^n_{ff} be the joint distributions of Z and Z_{n0} respectively:

$$\begin{aligned}
P^n_{tt} &= P(Z_{00} = t, Z_{n0} = t), \\
P^n_{ft} &= P(Z_{00} = f, Z_{n0} = t), \\
P^n_{tf} &= P(Z_{00} = t, Z_{n0} = f), \\
P^n_{ff} &= P(Z_{00} = f, Z_{n0} = f).
\end{aligned} \tag{14}$$

For example, $P^1_{tt} = P(Z_{00} = t, Z_{10} = t) = P(Z_{00} = t) \cdot P(Z_{10} = t \mid Z_{00} = t) = P_t \cdot P_{t|t}$ and so on.

We know,

$$\begin{aligned}
P^n_{tt} + P^n_{tf} &= P(Z_{00} = t) = P_t \\
P^n_{tf} + P^n_{ff} &= P(Z_{00} = f) = 1 - P_t.
\end{aligned} \tag{15}$$

Further, from Fig. 14, we have the following relations:

$$\begin{aligned}
P^n_{tt} &= P^{n-1}_{tt} \cdot P(Z_{n0} = t \mid Z_{n-1} = t) + P^{n-1}_{tf} \cdot P(Z_{n0} = t \mid Z_{n-1} = f) = P^{n-1}_{tt} \cdot P_{t|t} + P^{n-1}_{tf} \cdot P_{t|f}, \\
P^n_{tf} &= P^{n-1}_{tf} \cdot P(Z_{n0} = t \mid Z_{n-1} = t) + P^n_{ff} \cdot P(Z_{n0} = t \mid Z_{n-1} = f) = P^{n-1}_{tf} \cdot P_{t|t} + P^{n-1}_{ff} \cdot P_{t|f}.
\end{aligned} \tag{16}$$

When $Z_{n0} = t$, the posterior probability is:

$$P(Z_{00} = t \mid Z_{n0} = t) = \frac{P(Z_{00} = t, Z_{n0} = t)}{P(Z_{00} = t, Z_{n0} = t) + P(Z_{00} = f, Z_{n0} = t)} = \frac{P^n_{tt}}{P^n_{tt} + P^n_{tf}}. \tag{17}$$

Therefore,

$$\begin{aligned}
P(Z_{00} = t \mid Z_{n0} = t) > P(Z_{00} = t) &\Leftrightarrow \frac{P^n_{tt}}{P^n_{tt} + P^n_{tf}} > P_t \\
&\Leftrightarrow (1 - P_t) \cdot P^n_{tt} - P_t \cdot P^n_{tf} > 0.
\end{aligned} \tag{18}$$

Based on (16),

$$\begin{aligned}
(1 - P_t) \cdot P^n_{tt} - P_t \cdot P^n_{tf} &= (1 - P_t) \cdot (P^{n-1}_{tt} \cdot P_{t|t} + P^{n-1}_{tf} \cdot P_{t|f}) - P_t \cdot P^{n-1}_{tf} \cdot P_{t|t} + P^{n-1}_{ff} \cdot P_{t|f}) \\
&= \{(1 - P_t) \cdot P^{n-1}_{tt} - P_t \cdot P^{n-1}_{tf}\} \cdot P_{t|t} + \{(1 - P_t) \cdot P^{n-1}_{tf} - P_t \cdot P^{n-1}_{ff}\} \cdot P_{t|f}.
\end{aligned} \tag{19}$$

Substituting (15) into (19), we have

$$(1 - P_t) \cdot P^n_{tt} - P_t \cdot P^n_{tf} = \{(1 - P_t) \cdot P^{n-1}_{tt} - P_t \cdot P^{n-1}_{tf}\} \cdot (P_{t|t} - P_{t|f}). \tag{20}$$

Recursively, we have

$$(1 - P_t) \cdot P_{tt}^n - P_t \cdot P_{tf}^n = \{(1 - P_t) \cdot P_{tt}^1 - P_t \cdot P_{ft}^1\} \cdot (P_{t|t} - P_{t|f})^{n-1}. \tag{21}$$

Since $P_{tt}^1 = P_t \cdot P_{t|t}$, and $P_{ft}^1 = (1 - P_t) \cdot P_{t|f}$, we obtain

$$\begin{aligned}(1 - P_t) \cdot P_{tt}^n - P_t \cdot P_{tf}^n &= \{(1 - P_t) \cdot P_t \cdot P_{t|t} - P_t \cdot (1 - P_t) \cdot P_{t|f}\} \cdot (P_{t|t} - P_{t|f})^{n-1} \\ &= P_t \cdot (1 - P_t) \cdot (P_{t|t} - P_{t|f})^n.\end{aligned} \tag{22}$$

Combining (18) and (22), $P(Z_{00} = t \mid Z_{n0} = t) > P_t$ is equivalent to $(P_{t|t} - P_{t|f})^n > 0$ (when $0 < P_t < 1$). Similarly, we can show that $P(Z_{00} = t \mid Z_{n0} = f) > P_t$ is equivalent to $(P_{t|t} - P_{t|f})^n < 0$.

References

[1] J. He, W. Chu, and Z. Liu, "Inferring privacy information from social networks," in *Proceedings of IEEE International Conference on Intelligence and Security Informatics*, 2006.

[2] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining*, 2001. [Online]. Available: citeseer.ist.psu.edu/article/domingos02mining.html

[3] H. Kautz, B. Selman, and M. Shah, "Referral Web: Combining social networks and collaborative filtering," *Communications of the ACM*, vol. 40, no. 3, pp. 63–65, 1997. [Online]. Available: citeseer.ist.psu.edu/kautz97referralweb.html

[4] D. J. Watts and S. H. Strogatz, "Collective dynamics of "small-world" networks," *Nature*, 1998.

[5] W. W. W. C. (W3C), *The platform for privacy preferences 1.1 (P3P1.1)*, 2004, http://www.w3.org/TR/P3P11/.

[6] U. D. of Health and O. for Civil Rights Human Services, *Standards for Privacy of Individually Identifiable Health Information*, 2003, http://www.hhs.gov/ocr/combinedregtext.pdf.

[7] S. Milgram, "The small world problem," *Psychology Today*, 1967.

[8] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," in *Proceedings of the Seventh International World Wide Web Conference*, 1998. [Online]. Available: citeseer.ist.psu.edu/brin98anatomy.html

[9] M. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003. [Online]. Available: citeseer.ist.psu.edu/newman03structure.html

[10] D. Heckerman, "A tutorial on learning Bayesian networks, Tech. Rep. MSR-TR-95-06, March 1995. [Online]. Available: citeseer.ist.psu.edu/article/heckerman95tutorial.html

[11] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," in *Proceedings of KDD Workshop*, 1994, pp. 85–96. [Online]. Available: citeseer.ist.psu.edu/heckerman94learning.html

[12] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer, "Learning probabilistic relational models," in *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI)*, Stockholm, Sweden, August 1999.

[13] D. Lowd and P. Domingos, “Naive bayes models for probability estimation,” in *Proceedings of the Twenty-Second International Conference on Machine Learning (ICML)*. Bonn, Germany: ACM Press, 2005.

[14] N. L. Zhang and D. Poole, “Exploiting causal independence in Bayesian network inference,” *Journal of Artificial Intelligence Research*, vol. 5, pp. 301–328, 1996. [Online]. Available: citeseer.ist.psu.edu/zhang96exploiting.html

[15] *Livejournal*, http://www.livejournal.com.

PLACE
PHOTO
HERE

Wesley W. Chu is a distinguished professor of Computer Science and was the past chairman (1988-1991) for the Computer Science Department at the University of California, Los Angeles. He researched computer communications and distributed databases at Bell Laboratories, Holmdel, New Jersey (1966-1969). He joined the University of California, Los Angeles in 1969. His current research interest is in the areas of knowledge-based medical information systems, intelligent information systems and security and privacy in information systems.

He was the recipient of the 2003 IEEE Computer Society Technical Achievement Award for contributions to Intelligent Information Systems. He is also a member of the Editorial Board for the Journal on Applied Intelligence and an Associate Editor for the Journal of Data and Knowledge Engineering. Dr. Chu is a Fellow of IEEE.

PLACE
PHOTO
HERE

Jianming He is a third year Ph.D. student in Computer Science Department at the University of California, Los Angeles. Prior to UCLA, he received his B.S. degree from Fudan University, China in 1998, and his M.S. degree in Computer Engineering and Computer Science Department from California State University at Long Beach in 2003. His research focuses on social networks and collaborative filtering.