Social Inference Risk Modeling

in Mobile and Social Applications

[1]Sara Motahari, [1]Sotirios Ziavras, [2]Mor Naaman, [3]Mohamed Ismail, [3]Quentin Jones
[1]Electrical and Engineering Department, [2]Department of Library and Information Science, [3]Information Systems Department
[1,3]New Jersey Institute of Technology, [2]Rutgers University
sg262@njit.edu, ziavras@njit.edu, mor@scils.rutgers.edu, mxi8616@njit.edu, qgjones@acm.org

***Abstract*— The emphasis of emerging mobile and Web 2.0 applications on collaboration and communication increases threats to user privacy. A serious, yet under-researched privacy risk results from *social inferences* about user identity, location and other personal information. In this paper, after analyzing the social inference problem theoretically, we assess the extent of the risk to users of computer-mediated communication and location based applications through 1) a laboratory experimentation, 2) a mobile phone field study, and 3) simulation. Our experimentation involved the use of 530 user-created profiles and a 292-subject laboratory chat-study between strangers. The field study explored the patterns of collocation and anonymity of 165 users using a location-aware mobile-phone survey tool. The empirical data was then utilized to populate large-scale simulations of the social inference risk. The work validates the theoretical model, highlights the seriousness of the social inference risk, and shows how the extent and nature of the risk differs for different classes of social computing applications. We conclude with a discussion of the system design implications.**

Keywords-privacy, inference, ubiquitous social computing.

I. INTRODUCTION

Changes in the technological environment are creating numerous new and unaddressed risks to user privacy. Today's social computing applications such as Facebook enable users to exchange messages, reveal aspects of their profile, and even find profile-based matches. Location-based applications such as LoveGety leverage location, mobility, or proximity information to support navigation, recommendations, match making, etc. The resulting use and sharing of such personal information raise many serious privacy concerns. Previous efforts to protect users' privacy have made considerable advances in terms of computer and network security [1], user control mechanisms [2, 3], ethical considerations, and privacy policies [4]. However, the collaborative and pervasive nature of new mobile and social computing applications can give users the ability to leverage background knowledge about the social environment/context to make unwanted inferences.

The term *inference* as used in the privacy literature is the process of deducing unrevealed information as a consequence of being presented with authorized information. A well known example of the inference problem relates to an organization's database of employees [5], where the relation <Name, Salary> is a secret, but user *u* requests the following two queries: "List the *rank* and *salary* of all employees" and "List the *name* and *rank* of all employees." None of the queries contain the secured < *name; salary* > pair; however, an individual may utilize the known information <Rank, Salary> and <Rank, Name> to infer the private <Name, Salary> information through deductive reasoning. E.g., the knowledge that Bob is a manager and all managers earn $x, can help one deduce that Bob earns $x. Inference is mostly known as a security threat to databases [5] and sometimes as a privacy risk in data mining [6]. Although the inference problem as a threat to database confidentiality is discussed in many studies, Ubiquitous Social Computing (USC) raises new classes of inferences which we call *social inference*s. Social inferences are unwanted inferences that result from the use of social computing applications by the inferrer and are about user information associated with these applications such as identity, location, activities, social relations, and profile information. Threats to user privacy in mobile social computing systems have been placed into seven categories in [7]. In this paper, we focus on the two categories that relate to social inferences:

1. *Instantaneous Social Inferences* (e.g. my cell phone shows that I have a romantic match, Bob, who is nearby and I can only see two people with a similar cell phone around me. One of them must be Bob, thus increasing my chance of identifying him).
2. *Historical Social Inferences* through persistent user observation (e.g. two nicknames are repeatedly shown on the first floor of the gym where the gym assistant normally sits. One of them must be the gym assistant).

Previous inference prevention methods are inadequate in addressing social inference risks for one or more of the following reasons:

- Users typically utilize information outside the application (background knowledge) as a premise for inferences;
- The sensitivity of user information may have a dynamic nature based on the context, such as time and location.
- The user attribute being inferred (e.g. users' identity at physical appearance granularity) may not be stored in the application database; and
- Social inferences do not necessarily result from deductive reasoning [8] as shown in the salary example above.

In this paper, we aim to expand privacy research in the domain of mobile and social computing. While numerous social computing applications deal with privacy concerns through access control [2, 9] (e.g., Facebook enables users to set privacy preferences) it is clear on its own such control models will be unable to prevent unwanted social inferences.

Currently, we do not know how significant the social-inference risk is and how it differs for users of different classes of social computing applications. Gaining such knowledge requires advances in theory and systematic empirical studies that can then be utilized to extract important privacy and design implications. Consequently, we first analyze the social inference problem theoretically in the context of ubiquitous social computing and propose methods to predict the risk of social inference. We then employ experiments and large-scale simulations to explore the dangers and prevalence of social inferences in two critical applications of social and/or mobile computing: Computer-Mediated Communication (CMC) and location-aware applications. Simulations show how social inference dangers are divergent based on the application type.

II. Present Privacy Management Solutions

We categorize research into enhancing user privacy into four categories:

1. *Ethics, principles, and rules:* Privacy concerns can be partially addressed through the application of ethical principles and rules. Langheinrich [4] defines the principles of fair information practices as openness and transparency, individual participation, collection limitation, data quality, use limitation, reasonable security, accountability and explicit consent. He then sets principles for privacy in mobile computing, that consist of notice, choice, proximity, anonymity, security, and access.

2. *Access control systems:* Access control systems provide the user with an interface to set their privacy preferences. They directly control people's access to the user's information based on their privacy settings. Access control systems with an interface to protect user privacy started with internetworking, and were later extended to context-aware and ubiquitous computing systems. The earliest work within this area is P3P [10]. P3P enables users to regulate their settings based on different factors including consequence, data-type, retention, purpose and recipient. Ackerman [11] implemented a critic-based agents system called Privacy Critics, for online interactions. These agents watch the user's actions and make appropriate privacy suggestions. Access control mechanisms for mobile and location-aware computing were introduced later [2, 12, 13].

3. *Security protection:* Security protection handles the following aspects [1]:

- Availability (services are available to authorized users).
- Integrity (free from unauthorized manipulation).
- Confidentiality (only the intended user receives the information).
- Accountability (actions of an entity must be traced uniquely).
- Assurance (assure that the security measures have been properly implemented).

The inference problem is mostly known as a security problem that targets system-based confidentiality. Confidentiality protection is the area that includes most of the previous research on the inference problem. Therefore, suggested inference solutions often deal with secure database design.

4. *Inference management:* Two different techniques have been proposed to identify and remove inference channels. One makes use of semantic data modeling methods to locate inference channels in the database design, then redesign the database to remove these channels. The other technique evaluates database queries to understand whether they lead to unauthorized inferences. Each of these database management techniques has its drawbacks, including vulnerability to false positives and negatives, denial of service attacks, high computational complexity and overly restrictive limits on user access to information. These techniques have been studied for statistical databases [14], multilevel secure databases [15, 16] and general purpose databases [5, 17]. A few researchers have also addressed the inference problem in data mining [18, 19]. Denning and Morgenstern employed classical information theory to measure the inference chance in the realm of multilevel databases [20].

Although inferences can be made about a wide range of attributes, studies and polls suggest that identity is the most sensitive piece of users' information [7] and anonymity preservation is a key aspect of application design [21, 22]. Anonymity is defined as *"not having identifying characteristics such as a name or description of physical appearance disclosed so that the participants remain unidentifiable to anyone outside the permitted people promised at the time of informed consent"* [23].

Serjantov and Danezis [24], Diaz et al [25], and Toth et al [26] suggested information theoretic measures of degree of anonymity of the transmitter node in a network of message transmission systems that use mixing and delaying in routing the messages. [24] and [25] try to measure the average anonymity of the nodes in the network and [26] measures the worst case anonymity in a local network. They make very abstract and limited assumptions about the attacker's background knowledge which do not result in a realistic estimation of probability distributions for nodes.

Recently, new measures of privacy called *k*-anonymity and L-diversity have gained popularity [27, 28]. *k*-anonymity is suggested to manage identity inference in data mining, while L-diversity is suggested to protect both identity inference and attribute inference in databases. In a *k*-anonymized dataset, each record is indistinguishable from at least $k-1$ other records with respect to certain "identifying" attributes. These techniques can be broadly classified into generalization techniques, generalization with tuple suppression techniques, and data swapping and randomization techniques. As noted in the introduction, the challenge of social inferences cannot be addressed by previous management techniques alone because the user attribute being inferred may not be stored in the application database, users typically utilize their background as a premise for inferences, and the sensitivity of user information may have a dynamic nature.

In section III, we will explain, modify, and expand Denning and Morgenstem's formulation to predict the risk of social inference in mobile and social applications.

III. Social Inference Risk Prediction Framework

In this section we frame the social inference problem and explain the relation between social inferences and information entropy. We will also provide a framework for modeling users' background knowledge so that we can calculate information entropy and predict the risk of social inferences.

The logic is as follows: as we collect more information about a user, such as his/her contextual situation, our uncertainty about other aspects such as his/her identity may be reduced, thus increasing our probability of correctly guessing these aspects. This uncertainty is measured by *information entropy* in information theory. *Information* [29] as used in telecommunications is a measure of the decrease of uncertainty of a signal at the receiver. Here we use the fact that the more uncertain or random an event (outcome) is, the higher *entropy* it will possess. If an event is very likely or very unlikely to happen, it will not be highly random and will have low entropy. Therefore, entropy is influenced by the probability of possible outcomes. It also depends on the number of possible events, because more possible outcomes make the result more uncertain. In our context, the probability of an event is the probability that an attribute (such as a user's name) takes a specific value. As the inferrer collects more information, the number of entities that match her/his collected information decreases, resulting in fewer possible values for the attribute and lower information entropy.

To explain this in more detail, we bring an example from the user experiment described in [30] and built upon in this paper; Bob engages in an online communication with Alice. At the start of communication Bob does not know anything about his experimental chat partner, so the information entropy is maximum. After Alice starts chatting, her language and chat style help Bob determine (guess correctly) her gender and home country [30]. At this point, users of the same gender and region are most likely to be his chat partner. Thus, the probability of events is no longer uniformly distributed and entropy decreases. After a while Alice reveals that she is Hispanic and she plays for women's soccer team. Bob who has seen the soccer team playing before, knows that there is only one Hispanic female member and infers Alice's identity at physical appearance granularity. At this point, while Alice thinks she kept her identity a secret [30], Bob knows who she is because there is only one possible value for her identity. Therefore, social inferences happen when collected information reduces the inferrer's uncertainty about an attribute to a level that she/he could deduce that attribute's value. Collected information includes not only the information provided to users by the system, but also the information available outside of database or background knowledge.

Classical information theory was first employed by Denning and Morgenstern to measure the inference chance in the realm of multilevel databases [20]. Given two data items x and y, let $H(y)$ denote the entropy of y and $H_x(y)$ denote the conditional entropy of y given x. They defined the reduction in uncertainty of y given x is defined as follows:

$$Infer(x \rightarrow y) = (H(y) - H_x(y)) / H(y)$$

The value of *Infer* ($x \rightarrow y$) is between 0 and 1, representing how likely it is to derive y given x. If the value is 1, then y can be definitely inferred given x. Denning and Morgenstern did not suggest using this formulation in real situations because they did not know how to calculate conditional entropies.

Cuppons and Trouessin [31] formulate inference control as follows; If ***A*** is permitted to know information Q and ***A*** can derive information Φ from information Q (Q=>Φ), then ***A*** should be permitted to know Φ. Consequently, if we want Φ to be forbidden for ***A*** and Φ can be inferred from Q, Q should be forbidden for ***A*** as well.

None of the above formulations in their current state can address the social inference risk; Denning and Morgenstern's formulation has the problem of highly associating the inference risk with the maximum entropy, which undermines the importance of conditional entropies. Furthermore, it doesn't show how to calculate the conditional entropy. Cuppon's definition is formulated for logical deductions. We must remember that considering partial inferences in a social computing system, Φ may not be logically deduced from Q as indicated by Q => Φ. Morgenstern and Cuppon don't consider users' privacy preferences as a factor, because their focus is on database confidentiality protection.

We frame the social inference problem and define Q and Φ as follows: Information Φ is defined to be inferable from information Q if knowing Q could reduce the uncertainty about Φ and bring the entropy of Φ down to a risky *threshold*. Q is safe to be completely known by user ***A*** if he is permitted to know everything that can be inferred from Q: $\forall\Phi,\ [(H(\Phi|Q) < threshold \wedge PK_A(Q)) \Rightarrow PK_A(\Phi)]$ where $H(\Phi|Q)$ is the conditional entropy of Φ given Q.

We denote significant information available to the inferrer, including the background knowledge by Q; Q includes the inferrer's background knowledge as well as answers to their queries. In the case of historical inferences, Q includes the answers to previous queries starting at the current time and going back a given amount of time equal to T. Before the inferrer knows Q, Φ maintains its maximum entropy. The maximum entropy of Φ, $H_{\max}$, is calculated as follows:

$$H_{\max} = -\sum_{1}^{X} P.\log_2 P \ , \qquad \textbf{(1)}$$

where P=1/X and X is the maximum number of entities (users) related to the application.

After estimating all the information available to the inferrer, Q, we can calculate the conditional information entropy of attribute Φ as defined in information theory:

$$H_c = H(\Phi \mid Q) = -\sum_{i=1}^{V} P1(i).\log_2 P1(i) \ , \qquad \textbf{(2)}$$

where V is the number of possible values for attribute Φ. $P1(i)$ is the probability that the i^{th} possible value is thought to

be the correct one by the inferrer. $P1(i)$ is the posterior probability of each value given Q.

We illustrate this model through the study example mentioned above; Alice is engaged in an on-line chat with Bob. After a while her chat style may enable Bob to guess her gender and home country. Then she reveals she is a Hispanic female and a member of the soccer team. In this case, Φ is Alice's identity at name or face granularity. At first, Q comprises a guess on gender and home country, which changes the probability distribution of values as below;

$$P1(i) = \begin{cases} \varsigma.\sigma/X3 + \varsigma.(1-\sigma)/(X1) + (1-\varsigma).\sigma/(X2) & \text{for users of the same gender and country} \\ \varsigma.(1-\sigma)/(X1) + (1-\varsigma).(1-\sigma)/V & \text{for users of only the same gender} \\ (1-\varsigma).\sigma/(X2) + (1-\varsigma).(1-\sigma)/V & \text{for users of only the same country} \\ (1-\varsigma).(1-\sigma)/\mathrm{V} & \text{for the rest of users} \end{cases}$$

where V is the number of possible users of the applications, ζ is the probability of correctly guessing Alice's gender, σ is the probability of correctly guessing her home country[30], $X1$ is the number of users of the same gender, and $X2$ is the number of users of the same country.

After Alice reveals her gender and team membership, Q comprises the revealed information (gender=female, ethnicity=Hispanic, and group membership=soccer team) and background knowledge. Since personal profiles were found to be part of her partner's background knowledge, background knowledge includes users that are Hispanic female soccer players. V is the number of users that satisfy Q, which is the number of Hispanic female soccer players. At this point, V=1, $P1(i)$=1, and entropy is at its minimum level.

If we assume that unlike the above example all the information available to users is deterministic (which means they are either able to know the answer or not), assume that all information available outside the database is included in Q, and focus on anonymity protection, then $P1(i)$ in equation **(2)** equals 1/V. Consequently,

$$H_c = -\Sigma(1/\mathrm{V}).\log_2(1/\mathrm{V}) = \log_2(\mathrm{V}). \quad \textbf{(3)}$$

In this simplistic case, entropy is only a function of V. Since ***A*** is indistinguishable from (V-1) other users, V is ***A***'s *degree of anonymity*. In this simplistic case, the problem simplifies into a dynamic *k*-anonymity problem [28].

Focusing on anonymity protection again, we call **U** a user's *desired degree of anonymity* if he/she wishes to be indistinguishable from (**U**-1) other users. A user is at the risk of identity inference if her/his identity entropy is less than a certain *threshold*. This entropy threshold can be obtained using the desired degree of anonymity and replacing V by **U** in equation **(2)**. Assuming a uniform distribution results in:

$$\textit{Entropy Threshold} = \log_2(\mathbf{U}) . \quad \textbf{(4)}$$

To correctly calculate the conditional information entropy in a social application, we need to model the significant information available to the inferrer, including the background knowledge. The need to model background knowledge has been recognized as an issue in database confidentiality and integrity for a number of years [32]. However, as Jajodia and Midows [15] say, "we have no way of controlling what data is learned outside of the database, and our abilities to predict it will be limited". Thus, even the best model can give us only an approximate idea of how safe a database is from illegal inferences". The purpose of modeling background knowledge in this context is to identify 1) what attribute can be inferred (Φ) even if it is outside the database; 2) what attributes, if revealed, can help the inferrer reduce the number of possible values of Φ. Background knowledge can be estimated with different levels of accuracy: 1) The simplest method is to assume that the inferrer knows what we have in the existing application database, then estimate the number of possible values of Φ and their probabilities. The weakness of this method is that some of the attributes in the database are not usually known by the inferrer and some parts of the inferrer's background knowledge may not exist in the database; 2) The second method is to extend method 1 through hypothesizing about the inferrer's likely background knowledge; 3) The third method, is to utilize the results of user studies of background knowledge; 4) finally we could extend methods 2 or 3 with application usage data that allows for continuous monitoring of inferrer's background knowledge. An analysis of the relative utility of these approaches is beyond the scope of this paper.

IV. Research Questions

In section III, we described how to predict the social inference risk using information entropy. We saw low entropy that is less than a certain threshold indicates a high inference risk and the need to take an appropriate action. In sections V and VI, we try to answer the following questions in terms of system design and the appropriate action.

1. How serious is the social inference risk in USC applications and can we ignore it?
2. How does the risk change based on the application type?

V. Method

We investigated the danger of the social inference problem, the extent of the risk and the appropriate system design in two steps; user experiments and simulations. As a first step, we conducted two user experiments in two critical domains of mobile and social computing; Computer-Mediated Communication (CMC) and location-awareness. In the second step, we used the data obtained from the studies to simulate the risk on a larger scale and for various situations. A detailed presentation of Study 1 can be found in [30] where we explore the relation between information entropy and social inference in CMC. In this paper, we present a brief overview of the method and results of Study 1 as they are utilized to enable our simulations and to extract important design implications for CMC and location based social computing applications.

A. *Study 1*: *on-line communication between unknown chat partners*

This experiment was originally designed to:

1. Investigate users' background knowledge in CMC in order to be able to calculate the information entropy.
2. Test the ability of information entropy, calculated as explained in section III, to predict the inference risk.
3. Explore the risk of social inferences in CMC.
4. Provide real world data for large-scale simulations of online chats.

Our subjects participated in a study consisting of three phases: 1) online personal profile entry; 2) an experiment involving subjects chatting with an unknown online partner; followed by 3) a post chat survey about the subject's ability to guess their chat partner's identity. Five hundred and thirty students entered a personal profile, 304 participated in the chat session of which 292 subjects completed all three study components.

B. Study 2: The inference problem in proximity-based applications

This study, first presented in this paper, was conducted as part of a larger study focused on location-aware cell phone gaming. The field study aimed to:

1. Investigate users' background knowledge in proximity-based and location aware applications.
2. Verify that our calculation of information entropy predicts the inference risk.
3. Explore the risk and frequency of social inferences in the domain of location-awareness.
4. Provide population distribution and co-location data for large-scale simulations of a proximity-based application.

1) Subjects

All subjects were students of a medium sized urban university who were offered raffle tickets for answering a pre-study survey, carrying our Window Mobile phones for three weeks, and answering questionnaires on the phone. One hundred seventy five students participated in the study of which 165 completed the questionnaire at least once. Subjects were exclusively university students representative of the various majors offered on campus and ranging from 18 to 44 years old. Twenty percent of the subjects were female, and 65% of the subjects were commuters (living off campus).

2) Procedure

Phase I: Online Pre-study Survey – Subjects entered their contact information, demographic information such as age and gender, and questions related to their physical appearance, such as height and body type.

Phase II: Installation of a location estimation system that tracks users' locations on campus. The algorithm used to locate devices on campus is the Bayesian inference algorithm explained in [33]. Prior to implementation of the algorithm, a few students were employed to divide all publicly accessible rooms and paths on the campus into smaller cells. Cells subdivide rooms into smaller polygons to provide better location estimation accuracy. Students also visited all cells to collect sample signals that were used in building signal strength probability distributions for each cell.

Phase III: Installation of the 'Nearby' application on each phone which shows the nicknames of the users in the vicinity of the phone user on campus. Two users were considered to be nearby if they were in the same room or adjacent cells.

Phase IV: Pop-up Questionnaires Using Context-Aware Experience Sampling Method (CA-ESM) [34]. The questionnaire popped up every time subjects changed their location and stayed in a new location for 5 minutes or when they had not answered a questionnaire for at least 2 hours. The questionnaire started by asking the subjects about the accuracy of their location as captured by the location estimation system. If the location was captured correctly, the study continued to ask how often the subjects visited the location, how many people they saw in their physical vicinity, and how many of them were friends or acquaintances. In the subsequent questions, subjects were asked questions about the nicknames they saw on their nearby application, what they could guess about the identity of the nickname owner, and how they could map them to people in their vicinity. They elaborated on their guess by mentioning names or physical characteristics of the nickname owner.

C. Large Scale Simulations of the Risk

As a final step we used simulations to investigate the problem and appropriate actions on a larger scale for various situations. The simulation models were populated with parameters derived from our user study providing a good approximation to real world deployments.

VI. RESULTS

A. Results of Study 1: on-line communication between unknown chat partners

The key findings of the on-line communication study show how social inferences happen and how they pose a serious threat in CMC. In particular:

- The only measure found to strongly predict the identity inference was information entropy.
- Identity inferences are frequently made in CMC.
- Different users desire different levels of anonymity.
- Even when users are in complete control of the information they reveal, they are not able to maintain their desired level of anonymity, because they do not know what can be inferred from the information they reveal.

The detailed result of this study can be found in [30].

B. Results of Study 2: Proximity-based applications

A total of 1841 ESM questions were answered, with the questionnaire completion rate ranging from zero to twenty. Some users filled out up to 20 questionnaires, and some answered one. The location was captured correctly 86% of the time, which means 1583 valid questions were obtained.

Subjects' answers and their elaborations on what they guessed were compared to the demographic and physical information that we collected in the pre-study survey. In 46% of the cases, subjects were able to either exactly identify a

nickname on their nearby application or bring her/him down to two people in their vicinity. The survey ended by asking the subjects how they could identify the owner of the nickname. Forty percent of correct guesses were followed by choosing "*I know this person and I know this is the nickname s/he picks*" as the answer to this question. Those nicknames belonged to a friend (or an acquaintance) whose nickname and real name were the same. Thus, the nearby user was not anonymous and no identity inferences happened. Twenty nine percent of correct guesses said they could guess because they saw only a few people or a few cell phone users around and 28% of them said because they repeatedly saw the nickname and the person among their nearby users. The remaining 2.8% had other reasons, out of which one subject did not know how he guessed, one subject said the nickname made him think that it belonged to an Asian girl and he could only see two Asian females, and one subject said that the nickname looked like a girl's nickname and that he found one girl around. This means in 28% percents of all cases, subjects were able to bring an anonymous nearby user down to one or two people in their vicinity. Thus, assuming a desired degree of anonymity of 3 for everyone, instantaneous and historical social inferences happened in 15% and 13% of all cases respectively. While the risk was lower than the risk in CMC, it was still a serious risk.

C. *Simulation of the Risk of Identity Inferences in CMC*

Experimental results show that social inferences are not rare and are more common in CMC. We used the experimental data from study 1 to investigate the risk of identity inference in computer-mediated communication on a larger-scale. We first simulated personal profiles for a campus similar to our downtown campus environment. Profiles consisted of 20 individual profile items, including personal information, on campus activities, education information, and contact information. Parameters, such as the diversity of profile items, their statistical distribution, etc. were derived from the 532 user profiles obtained from study 1. Additional information such as the number of courses, statistical distribution of the number of students in a class, and enrollment statistics were obtained from university admission statistics. We then simulated online interactions of the users. The probability of revealing profile items and users' desired degree of anonymity were derived from the user experiment. Information entropy was calculated for each simulated chat based on their revealed profile information.

Fig. 1 shows the probability that a user's identity entropy is lower than its threshold. The y-axis shows the percentage of users for whom entropy was less than the threshold. The x-axis was chosen to represent the population because the size of the community highly affects the inference probability. The depicted curves show this probability for desired degrees of anonymity of 2, 3, and 5 (Entropy thresholds were calculated based on **U**=2, **U**=3, and **U**=5). In user study 1, 80.8% of the users who wanted to stay anonymous desired a degree of anonymity of two: **U**=2; and 5.1% of them desired a degree of anonymity of three: **U**=3. As expected, increasing the population decreases this probability. As the figure shows,

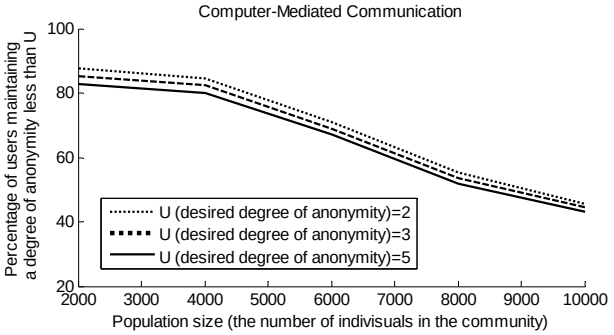

Figure 1. Risk of Identity Inference for Computer-Mediated Communication

while in a small school the risk can be very high, in a campus of 10,000 students, it is still about 50% in online chats between students. This means even in a rather big school, users reveal information that 50% of the time could lead to the invasion of their desired degree of anonymity. Therefore, identity inferences can be quite prevalent in CMC. This was also shown in our user study 1.

D. *Simulation of the Risk of Identity Inferences in Proximity-Based Applications*

We simulated a proximity-based application that shows nearby users by their nickname or real name based on nearby users' privacy preferences. Anonymity invasions (identity inferences) happen when a user's real name or nickname is mapped to the person or a few individuals using their nearby presence. Population density and distribution of nearby people has an important impact on the inference risk. In order to derive the related parameters needed for the simulations, we first analyzed our experimental data, which we explain below.

Based on the results, the mean of the number of people that subjects saw in their vicinity was 9.1 and its distribution is shown in Fig. 2. Among Poisson, Gaussian, exponential, Gamma, Lognormal, and Negative Binomial distributions, this distribution best fit the Negative Binomial distribution. We also measured the number of application users collected by the nearby application in the vicinity of each subject at each situation. The average number of nearby application users was 3.9 and probability distribution is shown in Fig. 2. These two measures are highly correlated (N=167, correlation coefficient, ρ=0.92; statistical significance, $p<0.001$) and the number of nearby people can be estimated as a linear function of the number of nearby application users with rms_{err}=1.6.

Subjects' answers show that their background knowledge mostly consists of their visual information about their vicinity and presence of nearby users. Therefore, significant information available to the inferrer includes the names shown by the application and physical appearance of current and past nearby users.

Simulations were first carried out for the nearby population distribution obtained from the user study 2, assuming mass usage of the application among the population. Equation **(2)** was used to estimate each nearby user's information entropy. When the number of possible values for a user's identity (V in

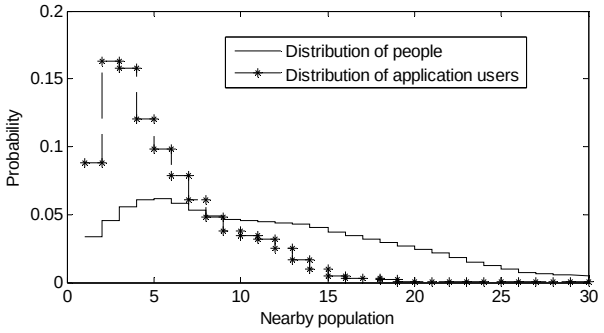

Figure 2. Probability distribution of nearby population

equation **(2)**) is set to the number of current nearby users, our measure of entropy only relates to instantaneous inferences. To measure the historical entropy for a nearby user ***A***, we should count the users whose co-presence history with the inferrer is the same as ***A***'s co-presence history. This measure depends on the history time, T. Optimization of history time and calculating the risk of historical inferences are beyond the scope of this paper.

Fig. 3 shows the probability that a user is at the risk of instantaneous identity inference in a proximity-based application. The y-axis shows the percentage of users whose identity entropy was lower than its threshold. Entropy threshold was calculated based on their desired degree of anonymity, **U** using equation **(4)**. The x-axis represents the desired degree of anonymity**.** Each curve depicts the risk for a different mean of nearby population density. The average density in the middle curve is equal to the average density of our experimental data. We see that assuming mass usage, the risk of identity inference is about 7% for a desired degree of 3, and 20% for a desired degree of anonymity of 5. As expected more crowded environments have a lower chance of being at the identity inference risk.

Fig. 4 shows the same risk for two more general nearby distributions; Gaussian distribution and a completely random spatial distribution of people (Poisson distribution). Again we see that the risk is less than 30% in the worst case, which is for a desired degree of anonymity of 5 and an environment that is 30% less populated than our campus. These results are also confirmed by the results that we obtained from the user experiment. Simulation of the risk of historical inferences and experimental results show that for a given population density,

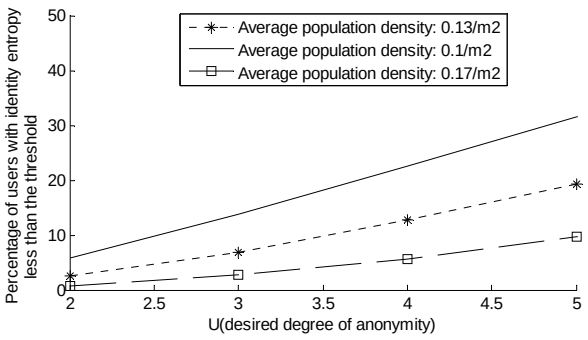

Figure 3: Risk of Identity Inference for the experimental distribution

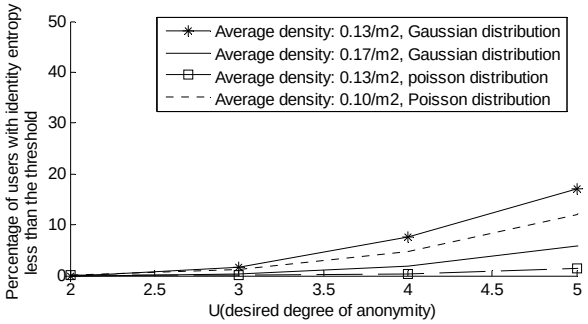

Figure 4. Risk of Identity Inference for Poisson and Gaussian distributions

historical inferences happen less frequently than instantaneous inferences.

VII. Discussion and Design Implications

The theoretical framework, user studies and simulations presented in this paper show that social inferences are not rare and pose a serious threat to user privacy. User experiments also showed that even when users are in complete control of the information they reveal, they are not able to maintain their desired degree of anonymity [30]. This is because individuals are unable to correctly judge inference risks. As a result, in order to protect individuals from unwanted social inferences, systems will need to be deployed that systematically reduce this risk. Fortunately, we can derive from our results (particularly from the simulations) a number of design implications that can aid in the development of such systems.

First, we need to provide users with the means to set their desired level of anonymity since we observed that users of USC applications can have a wide range of anonymity preferences.

Second, the calculation of information entropy, which was shown to be the best predictor of the risk, should be *automated* based on our framework. If entropy is less than its threshold, an appropriate action needs to be taken. The appropriate action can be rejecting the query, blurring the answer, sending a warning to the owner of the information, etc.

Third, for CMC systems user interfaces should be built to improve user inference-risk judgments through techniques such as risk visualizations or warning messages (perhaps applied to customized introduction tools). This implication is derived from the results that indicate that identity inference risks are quite common in CMC, which means that automatic control of information exchange in such applications can degrade system usability or be frustrating for the user. Furthermore, such applications are designed for users to consciously exchange information and users may be willing to compromise their anonymity settings to have more meaningful and productive communication. The systems can show users how uniquely they have specified themselves so far, or send a warning message when revealing a piece of information would enable their partner to invade their desired degree of anonymity. User studies will be needed to optimize such visualization so that they do not overly interrupt users.

The final design implication is derived from the finding that the prevalence of situations with identity inference risks in location-based services is lower than in CMC. In location-aware mobile applications, inference protection systems should modify information exchange, for example, by lowering the granularity of revealed information, rejecting a query, or blocking information exchange. In most cases lowering the information granularity, such as revealing the location at floor precision instead of room precision, or showing an anonymous name instead of a nickname can address the inference risk. This technique should not overly interfere with information exchange, or overly burden the user with privacy management actions.

REFERENCES

[1] W. Stallings, *Cryptography and Network Security Principles and Practices.*: Pearson Prentice Hall., 1999.

[2] D. Hong, M. Yuan, and V. Y. Shen, "Dynamic Privacy Management: a Plugin Service for the Middleware in Pervasive Computing," in *ACM 7th international conference on Human computer interaction with mobile devices & services* 2005.

[3] A. Gal and V. Atluri, " An Authorization Model for Temporal Data " in *ACM Conference on Computer and Communication Security*, 2000.

[4] M. Langheinrich, "Privacy by Design – Principles of Privacy-Aware Ubiquitous Systems," in *Third International Conference on Ubiquitous Computing (UbiComp 2001).* , 2001.

[5] A. Brodsky, C. Farkas, and S. Jajodia, "Secure databases: constraints, inference channels, and monitoring disclosures," *IEEE Transactions on Knowledge and Data Engineering,* vol. 2, pp. 900-919, 2000.

[6] A. Narayanan and V. Shmatikov, "Obfuscated Databases and Group Privacy," in *12th ACM conference on Computer and communications security* 2005, pp. 102-111.

[7] S. Motahari, C. Manikopoulos, R. Hiltz, and Q. Jones, "Seven privacy worries in ubiquitous social computing," in *ACM International Conference Proceeding Series; Proceedings of the 3rd symposium on Usable privacy and security* 2007.

[8] D. M. Gabbay and Guenthner, *Handbook of Philosophical Logic*: Kluwer Publishers, 2005.

[9] J. Cornwell, I. Fette, G. Hsieh, M. Prabaker, J. Rao, K. Tang, K. Vaniea, L. Bauer, L. Cranor, J. Hong, B. McLaren, M. Reiter, and N. Sadeh, " User-controllable security and privacy for pervasive computing.," in *Proceedings of the 8th IEEE Workshop on Mobile Computing Systems & Applications*, 2007.

[10] L. Cranor, M. Langheinrich, M. Marchiori, and J. Reagle, "The Platform for Privacy Preferences 1.0 (P3P1.0) Specification," in *W3C recommendation*, 2002.

[11] M. Ackerman and L. Cranor, "Privacy Critics: UI Components to Safeguard Users' Privacy," in *CHI 99*, 1999.

[12] S. Lederer, "Designing Disclosure: Interactive Personal Privacy at the Dawn of Ubiquitous Computing," in *Computer Science Division*: University of California at Berkeley, 2003.

[13] P. Osbakk and N. Ryan, "Context, CC/PP, and P3P," in *UbiComp 2002 Adjunct Proceedings*, Göteborg, Sweden, 2002.

[14] T. F. Lunt, "Current Issues in Statistical Database Security," *IFIP Transactions, Results of the IFIP WG 11.3 Workshop on Database Security V: Status and Prospects* vol. A-6, 1991.

[15] S. Jajodia and C. Meadows, *Inference Problems in Multilevel Secure Database Management Systems*. Los Alamitos, California, USA IEEE Computer Society Press, 1995.

[16] P. D. Stachour and B. Thuraisingham, "Design of LDV: A Multilevel Secure Relational Database Management," *IEEE Transactions on Knowledge and Data Engineering,* vol. 2, pp. 190-209, 1990.

[17] S. Dawson, S. D. Capitani, and d. V. P. Samarati, "Specification and Enforcement of Classification and Inference Constraints " *IEEE Symposium on Security and Privacy,* 1999.

[18] D. E. O'Leary, "Some Privacy Issues in Knowledge Discovery: The OECD Personal Privacy Guidelines," *IEEE Expert: Intelligent Systems and Their Applications* vol. 10, 1995.

[19] J. Zhan and S. Matwin, "A Crypto-Based Approach to Privacy-Preserving Collaborative Data Mining," in *Sixth IEEE International Conference on Data Mining Workshops*, 2006.

[20] D. E. Denning and M. Morgenstern, "Military database technology study: AI techniques for security and reliability," 1986.

[21] E. Heinrich, "Electronic Repositories of Marked Student Work and their Contributions to Formative Evaluation," *Educational Technology & Society,,* vol. 7, pp. 82-96, 2004.

[22] A. Kobsa and J. Schreck, "Privacy through pseudonymity in user-adaptive systems," *ACM Transactions on Internet Technology (TOIT),* vol. 3, pp. 149 - 183 2003.

[23] D. Doyle, "Trans-disciplinary Inquiry – Researching with Rather than Researching on," in *An Ethical Approach to Practitioner Research: Dealing with Issues and Dilemmas in Action Research*, A. Campbell and S. Groundwater-Smith, Eds.: Routledge 2007.

[24] A. Serjantov and G. Danezis, "Towards an Information Theoretic Metric for Anonymity," in *Proceedings of Privacy Enhancing Technologies Workshop (PET 2002)*, 2002.

[25] C. Diaz, S. Seys, J. Claessens, and B. Preneel, "Towards measuring anonymity," in *Proceedings of Privacy Enhancing Technologies Workshop (PET 2002)*, 2002.

[26] G. Tóth, Z. Hornák, and F. Vajda, "Measuring Anonymity Revisited," in *Proceedings of the Ninth Nordic Workshop on Secure IT Systems*, 2004, pp. 85-90.

[27] A. Machanavajjhala, J. Gehrke, and D. Kifer, "ℓ-Diversity: Privacy Beyond k-Anonymity," in *Proceedings of the 22nd IEEE International Conference on Data Engineering*, 2006.

[28] L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization And Suppression," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* vol. 10, pp. 571-588, 2002.

[29] C. E. Shannon, "Prediction and entropy of printed English," *The Bell System Technical Journal,* vol. 30, pp. 50-64, 1950.

[30] S. Motahari, S. Ziavras, R. Schular, and Q. Jones, "Identity Inference as a Privacy Risk in Computer-Mediated Communication.," in *IEEE Hawaii International Conference on System Sciences (HICSS-42)*, 2008.

[31] F. Cuppens and G. Trouessin, "Information Flow Controls vs Inference Controls: An Integrated Approach " in *Third European Symposium on Research in Computer Security* 1994.

[32] B. Thuraisingham, "Privacy Constraint Processing in a Privacy-Enhanced Database Management System," *Data and Knowledge Engineering,* vol. 55, pp. 159-188, 2005.

[33] K. E. B. Andrew M. Ladd, Algis Rudys, Guillaume Marceau, Lydia E. Kavraki, Dan S. Wallach "Robotics-Based Location Sensing using Wireless Ethernet," in *International Conference on Mobile Computing and Networking* 2002, pp. 227 - 238.

[34] J. A. S. Joel M. Hektner, Csikszentmihalyi, Mihaly., *Experience Sampling Method*: Sage Publications, 2006.