Towards an Understanding of Social Inference Opportunities in Social Computing

Julia M. Mayer, Richard P. Schuler, Quentin Jones
New Jersey Institute of Technology, USA
jam45, rps22 @njit.edu and qgjones@acm.org

ABSTRACT
Social computing applications are transforming the way we make new social ties, work, learn and play, thus becoming an essential part our social fabric. As a result, people and systems routinely make inferences about people's personal information based on their disclosed personal information. Despite the significance of this phenomenon the opportunity to make social inferences about users and how this process can be managed is poorly understood. In this paper we 1) outline why social inferences are important to study in the context of social computing applications, 2) how we can model, understand and predict social inference opportunities 3) highlight the need for social inference management systems, and 4) discuss the design space and associated research challenges. Collectively, this paper provides the first systematic overview for social inference research in the area of social computing.

Categories and Subject Descriptors
H.1.2 [Information Systems]: User/Machine Systems

General Terms
Human Factors, Design, Theory

Keywords
Social inference, social computing, impression management, privacy, personalization

1. INTRODUCTION
Social computing applications connect users to each other to support interpersonal communication (e.g. Instant Messaging), social networking (e.g. Facebook) and the sharing of user-generated content (e.g. YouTube). These applications are transforming the way we make new social ties, work, learn and play, thus becoming an essential part our social fabric. Everyday use of social computing applications, such as posting on a Facebook wall, checking-in to a location, uploading a picture, or liking or sharing a website, generates an enormous amount of personal data. The richness and widespread availability of this data enables *social inferences* about user preferences, identity, location, and private user information. Such *social inferences* are an emergent form of social computing that can occur when unrevealed personal user information, e.g., identity, location, user preferences, or profile information, is correctly inferred from revealed information in combination with *background knowledge*. We define *background knowledge* as information that has not been provided by the user or revealed by a social computing system and is available from other sources.

The following scenarios illustrate the ubiquity, versatility and the basic dynamics of social inferences in social computing:

Scenario 1: *Susan met this cute guy on a party last weekend, Josh. When Josh adds her on Facebook the next day, she's excited and studies his profile. He has tons of friends, lots of cool pictures and she finds out that he likes the same music as her. She doesn't have a lot of private info on her Facebook profile because she has some privacy concerns. However, she now quickly adds some of her favorite bands to her profile, hoping he will infer that they have a similar music taste. Later that day, Josh, who also enjoyed meeting Susan, curiously looks over Susan's profile. Seeing that they have several favorite bands in common, confirms his plan of meeting her again because he infers she must have a good music taste which is important to him. He knows that one of the bands they both like will play in town soon so he invites her to come with him.*

Scenario 2: *Greg started using a mobile dating app. One day, his phone vibrates and says 'somebody nearby wants to meet you'. He is just having coffee at Starbucks so he looks around to see what girls are nearby with a phone in their hand. There are several, however, only one is staring right back at him. She must have observed him pulling out his phone and then scanning his surroundings. He feels uncomfortable because the girl is not really his type. He looks away, gets up and quickly leaves. He tells himself to uninstall the app later.*

Scenario 3: *Susan told her friend Lily that she can't help her with this class project because she needs to visit her grandma in Delaware over the weekend and she won't have internet there. However, when Lily checks Facebook she sees in the Ticker that Susan just liked a recipe at the food network's website. Since Lilly knows that Susan does not have a smartphone, she infers that Susan must be online on her desktop having lied about the grandma visit or the internet problem there because she didn't want to help her.*

Scenario 4: *Maggie's boyfriends just broke up with her and she is indignant. She assumes that he cheated on her but she wants to know for sure. Since he won't admit it to her she decides to try to get into his email account. She does not know his password but the system prompts her to answer two security questions that will reset the password. She can easily answer the question about her boyfriend's first pet since he posted a photo of his beloved dog on Facebook labeling it with "Charley" and she knows his mother's maiden name. After resetting his*

GROUP'12, October 27-31, 2012, Sanibel Island, Florida, USA.

password she reads his emails finding out that he has been cheating on her for weeks.

Scenario 5: *Joe the robber is specialized in what he calls 'social media robberies'. His newest line of attack is looking at public location logs of Twitter and Foursquare users in his city and searching for people who go on vacation and inferring their home address from their check-ins. This works pretty well. Just last week he captured a lot of booty when he figured that the Smiths are gone for two weeks because they tweeted that they are excited for their vacation to Mexico and regularly checked-in to a location called "home" before.*

Scenario 6: *Greg travels a lot for work. To meet interesting people during his travels, he uses a location-based social matching system. One day he's walking down the street in Tokyo when his phone vibrates and tells him there is somebody nearby who went to the same college as him. He never entered in his matching preferences that he'd be interested in people from his college. However, the phone infers from his unusual context that now this might be interesting to Greg. Since Greg is alone in Japan, he is up for meeting another student from his US College so he starts chatting with him.*

These scenarios illustrate how the increasing use and sophistication of social computing applications and their related technologies has significantly increased the opportunities for social inferences. Motahari et al. [19] found that people often make both incorrect guesses and social inferences about other users based on their social networking profiles. However, at the same time people are very poor judges of the ability of other users and software systems to make inferences about them [22].

Unfortunately, current social computing system designs do not adequately take into account the potential for social inferences and do not provide users the tools necessary to manage them. When users of social computing applications provide or share personal information through the system (e.g. their interests, demographics, a picture, or their location), they are not informed about how their decision to reveal this information may impact potential social inferences. As a result, users of social computing applications have a limited understanding of the implications of their personal information sharing decisions.

In order to gain a solid understanding of social inferences and their impact on social computing system design we believe that extensive research into the different components of social inferences and the dynamics of social inference opportunities is necessary. Unfortunately, the underlying logic, namely that *background knowledge* when combined with *revealed user information* can lead to a *social inference opportunity* (Figure 1), has not been examined and researched systematically.

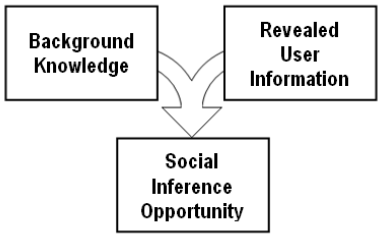

Figure 1. Underlying Logic of Social Inference Opportunities

To support the exploration of this problem we introduce a model outlining and linking the various components of social inferences. The understanding of these components and how they relate to each other will aid in effective social inference management system design and building. We pose the following questions in order to structure the discussion of our social inference model:

1. Why are social inferences important to study in the context of social computing systems?
2. How can we model, understand and predict social inference opportunities?
3. How can we design and develop social inference management systems that inform users about potential social inferences and allow them to regulate them?

Collectively, this paper provides the first holistic overview of the social inference phenomenon in social computing systems and lays out future direction for researchers.

The remainder of this paper is organized as follows. In the next section we provide an overview of the diverse aspects of social computing in which social inferences occur and explain the basic dynamics of social inferences. In section 3 we present four different social inference modeling approaches that can be used to identify social inference opportunities. Section 4 highlights the need for tools that raise awareness and allow users to control social inference opportunities. In section 5 we discuss open research challenges related with the principled design of social inference management systems. Section 6 concludes.

2. SOCIAL INFERENCES IN SOCIAL COMPUTING

Although there has been relatively little research into the social inference phenomenon in social computing systems, some other fields have explored the inference problem. The problem of inferences is well known to database researchers [8, 23, 24, 40]. A typical example comes from the database privacy literature [5]: the relation <Name, Salary> is a secret, but user *u* may request 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. For example, the knowledge that Bob is a manager and all managers earn $x, can help one deduce that Bob earns $x. This problem is known as *data re-identification* [31], the process of linking datasets without explicit identifiers such as name and address to datasets with explicit identifiers through common attributes. Another example is the linkage of hospital discharge data to voter registration lists that allows sensitive medical information to be inferred [32]. Considerable work has been undertaken exploring the general inference problem as a security threat to databases [8] and as a privacy risk in data mining [23, 24, 40]. *K-anonymity* [31], a method widely used in the database and privacy domain to detect and prevent such inferences about private user information works on the principle that "data is safe to release if at least k entities share the same attributes".

However, social inferences in social computing systems are very different in nature than inference problems that typically arise in the database and data mining domains (as seen from the example scenarios above). This is because both the inferred personal user information, as well as background knowledge used to make an

inference may not be stored in the application database. Some examples of inferable information are users' identity at physical appearance granularity (seeing the only person with a phone nearby - scenario 2) and users' activity and location (tweets about an upcoming vacation in Mexico - scenario 5). Furthermore, the availability of revealed user information and background knowledge is extremely dynamic in nature, especially in mobile social application, since it can be based on the user's context, such as time and location. Social inferences are also not only a privacy-risk but also have the potential to be beneficial and actively desired by the user (to impress or to receive valuable recommendations - scenario 1 and 6).

Social inference can affect very diverse aspects of social computing. Until now, inferences were only studied for certain and very limited contexts (e.g., as threat to database security). However, it is important to understand that social inferences occur across various areas of social computing and may have very different implications. Social inferences in social computing 1) can help users to manage their online impressions (*impression management)*, 2) may lead to privacy invasions and loss of anonymity (*user-privacy and security),* and 3) can help to receive personalized social recommendations (*system personalization)*. These three aspects of social computing affected by social inferences are illustrated in the sub-sections below.

2.1 Impression Management

According to Goffman [12], people routinely manage impressions through varying self-presentations depending on the social setting and audience. Social inferences usually occur in this context because impressions are formed both based on explicitly revealed personal information as well as unrevealed but inferred information. Similarly, users of social computing systems, particularly social networking sites (SNS), consciously or unconsciously attempt to influence the perceptions other people have about them by regulating and controlling the personal information they share [12] (see scenario 1). For example, an important part of the value proposition of business-oriented Social Networking Sites (SNS) such as LinkedIn comes from the social inferences they enable about members. Users often make inferences about other members' skills and business networks based on fairly limited user-profile information and weak-tie relationships. Social inferences about the extent of an individual's expertise made from information provided through LinkedIn could be used to find worthwhile employees.

However, while users routinely rely on social inferences made about them when presenting themselves on SNS these systems do not provide users a method to understand what others might be able to infer from their profile information. Thus, to effectively aid impression management users require methods to go beyond traditional profile management and enable them to perform what we call 'social inference management'.

2.2 User-Privacy and Security

Due to the growing ubiquity of social computing and the manner in which people use them and in turn how those systems collect and share information raise pertinent privacy and security concerns.

Social inferences can potentially lead to unwanted disclosure of personal user information, for example, a user's identity or location. This is illustrated by the following actual incident [29]. During the deployment of CampusWiki, a location-aware application that allows users to create and edit location-linked content which can be anonymous or identified, a student anonymously added unpleasant comments about a course professor. The professor was able to utilize the time-location stamps of page edits to determine that the comments were made during his class period and near the classroom. He then proceeded to monitor laptop use by students during the class in question, and was then able to determine the student responsible for the comments. The result was a confrontation, which lead to the student dropping the course. In this case a user's expected privacy was violated due to a social inference.

A simple, yet to-the-point definition of privacy is "a person's right to control access to his or her personal information" [3], i.e., the expectation that others will not know and cannot find out what they wish to keep secret. In other words, a person should have the ability to exclude others from accessing individuals' personal information and to determine when, how, and to what extent he or she will release personal information. Anonymity preservation, to be unidentifiable among a certain number of other people [22], also is an important aspect of privacy.

While privacy is well researched in the literature, there has been scant research on the social inference problem as threat to user privacy in social computing communities. Motahari et al. [20, 21, 22] investigated identity inferences in open partially anonymous computer-mediated communication (CMC) used to support interaction between partially or fully anonymous individuals and found that users were able to make social inferences about their chat partners' identity while they assumed to be anonymous. Users of social computing applications often engage in conversations (chats) where they want to stay anonymous, e.g., initial conversations on social matching / dating sites, private messages with an unknown member on Twitter, online forums, Chat roulette, or in massively multi-player online games (MMOG). Here, information disclosed during the conversation may allow for identity inferences while users still assume to be anonymous.

User-location is increasingly popular information item collected and shared via social computing applications (e.g., location-based social networks like Foursquare). Motahari et al. [22] also investigated anonymity and identity inferences in proximity-based social applications and found that disclosure of location and patterns of co-location often lead to social inferences about users' identity (see scenario 2). A study by Krumm [16] showed that the location of a user's home and his/her identity can be computed only using pseudonymous GPS data, simple algorithms and a free Web service. An official warning by the United States Military was given due to the risks posed by geotagging, adding a geographical metadata to a picture and uploading it to a social computing application [26]. This could potentially allow inferences about the exact location of a soldier by an enemy allowing for an attack. This risk can basically be extended to anyone who uploads geotagged pictures to social computing applications. A factor mediating this risk is having hundreds of 'friends' that may have never been met in person. From tagged locations like the frequently visited restaurants, the gym visited everyday and the street living in, these 'friends' (that are actually strangers) can infer routines and habits. This way, social inferences may also enable stalking. For example, a college student's SNS profile including information about residence location, class schedule, and location of last login may help a potential stalker to determine the user's whereabouts [13]. PleaseRobMe [11] is a website that aims at raising awareness

about oversharing, particularly how users of location-based services make themselves vulnerable to housebreaking by checking-in to locations online (see scenario 5).

A possible consequence of identity inferences is identity theft. Making birth date, hometown, current residence, and current phone number publicly available at the same time can be used to estimate a person's social security number and exposes her to identity theft. Since a vast majority of Facebook profiles not only include birthday and hometown information, but also current phone number and residence (often used for verification purposes by financial institutions and other credit agencies), users are exposing themselves to substantial risks of identity theft. Passwords and answers to security questions may also be jeopardized as a result of a social inference. Users often select passwords that have a personal meaning for them as they are easy to remember [28, 17, 37] (e.g. name of a favorite pet, birth date, social security number) that may be able to be determined via a social inference. Furthermore, many systems insist on the use of security questions for lost or forgotten passwords intentionally based on potentially inferable personal information [28]. For example, typical security questions ask users about their home town, their mother's maiden name or their pet's name; information that is often publicly visible on social networks or through search engine results (see scenario 4). A very public and well known example of this problem is known as the "Palin Hack" [39]. During the 2008 United States presidential election campaign someone had obtained access to Sarah Palin's private email account and posted several screenshots of her emails in an online forum. The "Palin Hack" did not require any real skill, instead, the hacker simply reset Palin's password by answering the security questions on the account using her birth date, ZIP code and information about where she met her spouse - information that was easily obtained by a simple web search. This shows how answers to security questions are prone to be inferable from revealed personal information and by searching various sources for user attributes which may lead to security holes.

These diverse examples illustrate how social inferences can have significant implications on user-privacy. In order to effectively protect users' privacy in social computing, we need to understand and model social inference opportunities which in turn will allow us to build and develop social inference management systems.

2.3 System Personalization

Social inferences can also be valuable to the user for system personalization, specifically for recommendations. System personalization is usually based on explicit data such as interests or implicit data such as context of each individual user or user group [35]. With the exponential growth of available information on the Web, social inferences about user preferences can help to tackle the information overload problem by personalizing the systems appearance and behavior and recommending information that is relevant to the user.

Social computing systems often provide personalized services based on user profile information. Social inferences about user preferences that are not explicitly stated provide a sophisticated way to personalize recommendations. For example, personalized social recommendations can help users form new relationships by suggesting other individuals of interest (e.g., Facebook's and LinkedIn's "People You May Know") [35]. Social matching systems (e.g., dating sites like Match.com) typically base match recommendations on users' explicit matching-preferences and/or measures of the overall affinity between individuals.

Current personalized services often inadequately inform the user about the reasons for the personalized recommendations and do not provide adequate control over how recommendations are made. This may lead to frustrations and undesired recommendations when users do not agree with or understand the reasons for the system's personalization decisions (e.g., repeatedly being shown the hated ex-girlfriend as "people you may know"). Furthermore, this can potentially reveal information that invades users' privacy or leads to compromised security, for example when user-location is used for people recommendations and the system reveals the proximity between two users leading to an identity inference (see scenario 2).

Social inference about users' match preferences (e.g., interest in another student from the same U.S. college while traveling in Japan – scenario 6) can potentially yield to more desirable social match recommendations. Context (user location, activity, resources, etc.) is a helpful piece of information to allow for valuable social inferences in the area of mobile social matching. Mayer et al. [18] found that recommendations made based on user-context and contextual rarity of shared user attribute are more interesting to users. We need to gain a deeper understanding as to how social inferences can be used to personalize services in socially intelligent ways based on information such as user-context, as well as how users want them to be used for system personalization. Together with impression management and user-privacy, system personalization needs to be effectively managed by users by providing them with social inference management tools.

3. DYNAMICS OF SOCIAL INFERENCES

We define social inferences as inferences about personal user information (e.g. user preferences, identity, location, and profile information) that has not been explicitly revealed but can be inferred from revealed information in combination with background knowledge. The act of inferring always involves two entities: 1) an individual about whom something is inferred, i.e. the inferee, using his/her revealed personal information and 2) an inferring entity (an individual or system), i.e. the inferer, who combines this information with background knowledge to perform a social inference (see Figure 2). Figure 3 shows a more detailed version of the simple underlying logic of social inferences in Figure 1. It illustrates what different types of *background knowledge* can, when combined with *revealed user information*, lead to a *social inference opportunity*.

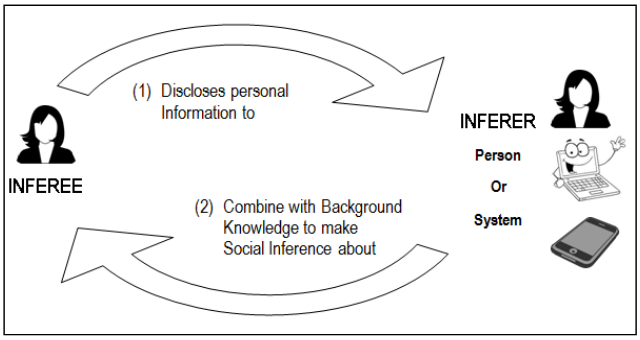

Figure 2. Dynamics of Social Inferences

Since the inferring entity can be a system or a user we differentiate between two different types of social inferences; those made by systems (*system-based social inferences*) and those made by people (*human social inferences*).

System-based social inferences can occur when a system combines its background information (implicitly collected, e.g., from internal user-profiling, web crawling, sensor data) with information explicitly disclosed by the user by applying heuristic rules. This is similar to the idea of emulating human reasoning to extend a system's knowledge base that is explored in the field of artificial intelligence and semantic web using methods like statistical inference (e.g. Bayesian inference [33]), or automated ontology construction [4]. In social computing, system-based social inferences about users' social matching preferences can be valuable for personalized social recommendations [18].

On the other hand, *human social inference* are inferences users of social computing systems can often perform when interpreting other users' revealed information. By combining personal background knowledge with what the system reveals about a user, an inferee might be able to infer unrevealed information, such as a user's nationality based on the user's name or hometown, a user's approximate age based on the graduation year, or the impression made based on personal pictures and status updates. Users can build up background knowledge in various ways over time, for example, through experience and learning, through observation, or a web search.

3.1 Revealed User Information

Revealed user information is the information a user explicitly discloses through a social computing system. This information can be static (e.g. user's demographics) or dynamic (e.g., user's location). For example, users disclose personal information, such as interests, hobbies or educational and professional information on SNS profiles like Facebook or LinkedIn. Behavioral information, such as user's activity and interactions could be revealed through status updates on Facebook or Twitter. Geotemporal information, such as a user's current location, is typically disclosed though check-ins (e.g. Foursquare, Facebook Places).

Today's social computing applications tend to promote and push users to share more and more personal information. The value of social computing systems comes from sharing. Systems are designed in a way to get users create rich personal profiles disclosing increasing amounts of personal information and sharing it with the world.

3.2 Modeling the Social Inference Opportunity

A social inference opportunity exists if it is *logically possible* to make a social inference. We consider a social inference logically possible if unrevealed information can be deduced by analysis of a user's or system's background knowledge using the additional revealed information provided by the user.

The fact that a social inference is logically possible (a *social inference opportunity*) does not mean that a person or system will make a social inference. To calculate the probability that a social inference will actually occur, requires 1) knowing that an inference is logically possible, then 2) having a reliable estimate of (a) the number of potential inferers exposed to the revealed information, (b) the range of contexts in which the potential inferers are exposed to the revealed information and (c) the likelihood that for each '*potential inferrer-inference context*' an inference will actually be made. Clearly, it is considerably more difficult to determine the likelihood of a social inference than if a social inference opportunity exists. However, for system designers and CSCW researchers we do not believe that the initial focus of our efforts should be on assessing the social inference probability. The following analogy helps illustrate why this is the case: People routinely lock their cars in public parking even though on most occasions no individual will attempt to open the door to steal items inside. The decision to lock the doors can be based on the logical possibility of easy item theft, rather than the overall probability that somebody will walk past and try opening the car door in case it is unlocked and steal an item. Car designers do not worry about such probabilities; instead they provide drivers with the ability to know if they have controlled easy access – i.e. if they have locked the door. Similarly, we first need to provide users with an understanding of the social inference opportunities of various information sharing decisions. To achieve this end we need to be able to model the logical possibilities of social inferences.

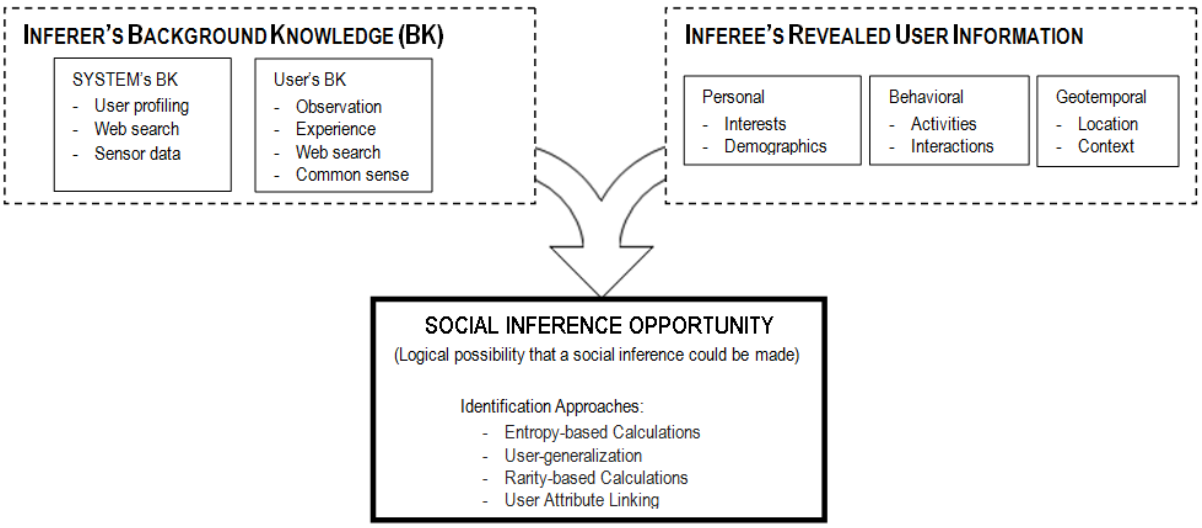

Figure 3. Modeling the Social Inference Opportunity

In following four subsections, we provide a brief overview of the approaches we are aware of for identifying social inference possibilities: 1) entropy calculations, 2) user-generalizations, 3) rarity calculations, and 4) user attribute-linking.

3.2.1 Entropy-based Calculations

The entropy approach is based on the following fact: as individuals or applications collect more information about a user, our uncertainty about other attributes, such as his/her identity, may be reduced, thus increasing the opportunity of a social inference. The following scenario from a user experiment of online chat between anonymous partners illustrates the social inference opportunity assessment framework [22].

> *Bob engages in an online chat with Alice. At the start of communication, Bob does not know anything about his chat partner. He is not told the name of the chat partner or anything else about her, so all users are equally likely to be his partner. After they start chatting, Alice's language and chat style help Bob guess her gender and that she is Hispanic. After a while, Alice reveals that she plays for the university's women's soccer team. Bob, who has prior knowledge of this soccer team, knows that it has only one Hispanic member. This allows Bob to then infer Alice's identity.*

Here, as Bob combines his background knowledge of the female Hispanic soccer players on campus with what Alice reveals, his uncertainty about his chat partner's identity decreases, thus increasing the opportunity of a social inference. This uncertainty can be measured by an *information entropy* calculation [30]. Information entropy, as used in information theory for telecommunications, is a measure of the decrease of uncertainty in a signal value at the receiver site. Here we use the fact that the more uncertain or random an event (outcome) is, the higher the *entropy* it possesses. 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 a greater number of 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.

The information entropy modeling approach was tested in two different areas of social computing: 1) for anonymous computer-mediated communication, through a laboratory chat experiment between unknown chat partners and 2) for proximity-based applications, through a mobile phone field study that explored patterns of co-location and anonymity of the subjects. In both studies, entropy-based inference modeling was the strongest predictor of social inference opportunities [20, 21, 22].

3.2.2 User-generalization

This approach is based on the following fact: Individuals or applications can derive general concepts from repeated experiences or observations and then apply these concepts to a certain user, allowing for a potential social inference. This is typically done by recommender systems that attempt to infer user preferences based on buying behavior and preferences of other users (e.g., someone buying several science fiction books must like other science fiction books because other users showed that same interest).

Furthermore, matchmaking systems are usually based on the concept of homophily, i.e., that people like people who are like them (similar demographics, interests, etc.). Currently, social matching systems calculate user affinities by weighting the similarities between users over a set of user attributes stored in the user profile. Standard user attributes include interests (hobbies, favorite music), social ties, demographics and personality. In order to find valuable social matches, rich user profiles are needed. The more information a system has about a user, the better it can compute social matches. Ideally, mobile social matching systems could improve matching by leveraging geo-temporal social data [18].

Generalizations about geo-temporal data can be used to make inferences about user's local context, for example, involvement in an activity based on location: Tony is at the gym, so he is probably working out [10, 15, 34]. Short-term and long-term trajectories, location logs, and geo-temporal patterns may also be used to infer user's interests from a user's frequently visited places (Susan regularly goes to the church; then she is Christian). Proximity patterns can be used to identify people who are nearby each other often but do not know each other yet [15].

3.2.3 Rarity-based Calculations

Potential social inferences can also be identified based on user attribute rarity. This is based on the following fact: the rarer an information item in the system is the more meaningful it is to the system, thus increasing the social inference opportunity. This is related to the concept of TF*IDF (term frequency-inverse document frequency) in data mining [25] which assumes that the frequency of a word in a document compared to the frequency of that word in the set of all documents is an indicator of the importance of that word.

The rarity-based approach was explored in the context of mobile social matching. Social matching systems are social recommender systems that help users to find other individuals of interest [35]. Traditional social matching systems calculate user affinities by weighting the similarities between users over a set of user attributes. Mobile social matching extends social matching to mobile devices by recommending people to people based on their current local context. It was found that in mobile social matching not only the similarity between two users is useful to determine the value of a match but also the rarity of this similarity in the user's current context [18]. C*ontextual rarity* is a measure of how many other individuals have the same attribute in the user's social context. A generally common attribute can become 'contextually rare' in certain contexts. The scenario that follows illustrates this point:

> *Daniel is an undergraduate student in Information Systems at an American college. He uses a mobile social matching tool with social inference management technology. When on his campus, the system knows he is not interested in receiving recommendations to meet other students. On the other hand, during the summer Daniel goes as an exchange student to Italy. As the matching system recognizes the contextual rarity of being from his American college in Italy, the system is able to infer that a recommendation to meet another unknown exchange-*

student from his college who is in his vicinity when shopping in down-town Rome would be of value.

In the above example, the match relevance was inferred from the rarity of the particular context (it is really unusual for two students who do not know each other and are from the same American college to be in each other's vicinity in Italy). This scenario highlights how the rarity of a user attribute can lead to beneficial social inferences about users' match preferences. The prevalence of an attribute in a certain population, which might be the user's greater social environment or a particular current context, is calculated the following way: $P(a) = \frac{n_a}{N}$. The prevalence *P* of an attribute *a* equals the number of occurrences of attribute a divided by the size of the population *N*. Rarity is the inverse of the prevalence, i.e. attributes with a high prevalence are not rare (i.e., are common) but attributes with low prevalence are very rare. For example, rarity of being German at an American college equals the number of people at the American college divided by the number of Germans at this college.

The contextual rarity approach was tested using a personalized self-reported web survey exploring seven different affinity types (interests, needs, geographical background, educational background, distinct characteristics, places and friends) [18]. For each section respondents were instructed to enter three of their own user attributes. Then, for each attribute and the combination of all three in this section, they were asked two sets of questions: First, they had to rate the commonness of the attribute in different contexts (home, work, social circle, etc.) and second, they were asked about their level of interest in a potential match with same user attribute in different stationary, mobile, common and rare situations. The results of this survey study confirmed that people are generally more interested in matches based on rare user attributes. They also found that the relative contextual rarity can influence the desire for a social match [18].

We consider this relationship between contextual rarity and users' match preference as a social inference opportunity. Systems can potentially infer users' context-dependent match preferences using rarity calculations. However, this approach has not yet been implemented and we only have a limited understanding of the applicability and generalizability of this method to detect logical possibilities of a social inference.

3.2.4 User Attribute Linking

This approach is based on the following fact: individuals or applications can link together a number of isolated facts from various sources, thus leading to a potential social inference.

It has been shown previously that a large portion of the US population can be re-identified using a combination of 5-digit ZIP code, gender, and date of birth [32]. For example, information can be searched from several online sources. People routinely search what personal information about them or others are publicly available online, e.g. using search engines, to make potential social inferences. In addition to these easily accessible search results, there are other personal data publicly available that are not directly accessible through common search engines, e.g., intelius.com [14] to find somebody's age, zabasearch.com [38] to find an address or home phone number, or portals to public government data to find a state employee's salary [1]. Often times, inferees are not aware that this kind of information is publicly available, which could lead to potential social inference, for example about a user's password (see "Palin Hack" [39]).

User attribute linking can also be used for face re-identification by linking facial images from social networking profiles to other available facial images, e.g., from public websites. Gross and Acquisti [13] were able to correctly link facial images from Friendster profiles without explicit identifiers with images obtained from fully identified university web pages using a commercial face recognizer.

4. DESIGNING SOCIAL INFERENCE MANAGEMENT SYSTEMS

Current social computing systems do not consider social inference opportunities. They implement basic access control systems that provide users with an interface to directly control people's access to their information [7]. Social networking sides (SNS) usually provide users with profile management and privacy setting user interfaces (UIs) that allow them to manage revealed profile information. These are usually structured for the kind of data a user can enter (e.g. name, interests, contact information, demographics and a profile picture). Some social network systems (e.g., LinkedIn, Facebook) differentiate between a public profile (visible to everybody) and a private profile (visible only certain people in the social network systems, e.g. friends, contacts). For example, Facebook's privacy settings allow users to customize the visibility of each item on their profile for different friend lists, as does Google+ with the 'circles' feature. Similarly, various mobile location-aware applications such as Loopt allow users to set rules about their locatability in particular locations for particular people. Users are generally able to view how their profile is seen by other users based on the information explicitly disclosed.

However, users are often neither aware nor informed about the issue that personal information about them might be inferable. Simple access control systems and profile management UIs only allow users to see and control the information explicitly revealed about them but none of these systems consider and inform users about social inference opportunities. This shows that current simple rule-based approaches to privacy and security do not cope well with the dynamic and context-dependent nature of social inference opportunities.

Based on the highlighted issues, we propose social inference management systems that support users' awareness of social inference opportunities through UIs and visualizations and allow users to control them through management tools. These systems would ideally provide users with visualizations of social inferences opportunities in relation to:

1) Their digital self-presentation for *impression management*;
2) Resulting *privacy and security* implications of data sharing (e.g., level of anonymity in the current context);
3) Opportunities for *personalized recommendations* (e.g., inferred contextually rare match criterion used for personalized recommendation).

Unfortunately, there is a large gap in social computing research investigating system interfaces and mechanisms to manage social inferences. In order to build effective social inference management systems we believe that broad and extensive research efforts are necessary to study the social inference phenomenon across the different presented areas.

5. RESEARCH CHALLENGES

There are several research challenges associated with the principled design of social inference management systems. Based on the previous discussion we argue that current social computing system designs do not provide users with the proper methods to understand and manage social inferences. In order to transform our understanding of social inference opportunities in social computing and be able to design and build effective social inference management systems, we need to overcome research challenges in terms of:

1) Users' current understanding of inference opportunities, their risks and benefits;
2) How to model and reliably predict social inference opportunities; and
3) User-interface design alternatives and their likely effectiveness in various contexts.

We will discuss each of these challenges in more detail in the following three sub-sections.

5.1 User's Understanding of Inference Opportunities (Risks and Benefits)

In order to design social computing systems that incorporate social inference management, we first need to explore people's current understanding and perceptions of the social inference opportunities associated with the social computing applications they presently use. Research into users' awareness of social inference opportunities (to what extent people guess that systems or other users can infer private information about them), and what users perceive to be the implications of such opportunities (to what extent people perceive social inferences as negative, e.g., privacy invasion, or as beneficial, e.g., personalization), will inform us about how users currently experience social inferences in social computing systems. In particular we need to explore:

1) People's beliefs about other users' abilities to infer information about them via various apps they use as well as the various contexts of use;
2) What they want (and do not want) to be inferable about themselves from apps;
3) Their satisfaction and frustrations with personalization, recommendations and privacy settings of applications they use (e.g. adjusting the information shared based on location). How well do the applications infer needs, preferences and desires and give users control over these settings?
4) People's inferences about other users in order to identify possibly important components in our inference opportunity modeling;
5) Their preferences, from a privacy-point of view as well as from a system personalization-point of view.

Collectively, this will provide us with a better understanding of how social inferences are perceived by the broad user-community. This in turn will provide us with important insights into social inference management system design requirements.

5.2 Social Inference Opportunity Modeling

Social computing systems that incorporate social inference management must be able to effectively and efficiently model the logical possibility of social inferences. In order for systems to provide social inference based impression management, privacy and personalization we need to learn how to evaluate and compute potential social inference opportunities associated with user-information sharing in various contexts. We must also generate methods for understanding and evaluating the quality of social inference detection. To achieve this goal we need to thoroughly examine how to:

1) Quantify and model the background knowledge used by people to make social inferences in the domains of impression management, privacy and personalization;
2) Create and populate a world model that can be used to support the social inference management UIs; and
3) Assess and refine the entropy, rarity, user-generalizations, and user attribute-linking calculation techniques used to make social inferences.

To achieve this goal, profound studies examining users' ability to make inferences about others as well as to assess other users' ability to make inferences about them are necessary. Many of the social inference calculations require the accurate assessment of the probability/prevalence of an attribute occurring in a particular context; e.g., the chance of a student knowing another random student who is on campus and the chance that a student is coming from Germany and studying HCI. As a result, the world model needed will need to contain extremely detailed information. This could be achieved using commercial entities that have much of this data on hand (e.g. over 90% of current students have Facebook accounts) or using a software tool that enables the collection of rich social network data and in-depth information from a large number of individuals within short timeframes (e.g. [22]). It is important to gain knowledge of subject's social ties, to profile individual users and collect background knowledge of the social activities subjects engage in.

Based on that, research efforts need to further examine how social inferences can logically be predicted by combining revealed information with the data contained in the world model. The aim here should not be to create an ideal world model, but rather to learn about the requirements for effective modeling.

Data management and system building is another challenge that requires our attention. To model social inference opportunities, data could be pulled from the outside sources or it could be stored in massive internal databases. Furthermore, the computational models could be constructed internally or externally.

We have to explore advantages and disadvantages related to various options and create and evaluate IT artifacts that can help us gain deeper understanding about social inference opportunity modeling.

5.3 Social Inference Management User Interfaces

A third research challenge is to learn how visualizing social inferences opportunities can help users to: 1) learn about digital self-presentation for impression management; 2) estimate more accurately the privacy and security implications of data sharing; and 3) control personalization of various social software services. We need to develop theories and tools that allow for the principled design and development of social inference management systems by providing the user with the opportunity to learn about and

control the social inference possibilities associated with their system use and personal data-sharing.

Basic design alternatives can include visualizations showing highlighted profile sections that could be inferred in addition to the user profile as seen by others, awareness displays [2, 6] that visualize the logical possibility of a social inference and visualizations that explain social inference opportunities to users in form of words, statistics, probabilities, etc., so users can learn and understand the process. Furthermore, different intervention mechanisms could be incorporated: Passive mechanisms could warn or alert the user about social inference opportunities when disclosing certain information without inhibiting the disclosure, while active mechanisms could blur the information, e.g. revealing the age bracket instead of date of birth or even blocking the information sharing. Automated mechanisms could allow users to set a certain required level of privacy and instruct the system to keep this level by prohibiting social inferences.

6. CONCLUSION

Social inferences are a ubiquitous yet under researched phenomenon in social computing systems. This current situation poses a serious challenge to the Group/CSCW research community since social inferences happen regularly, users and researchers lack awareness and understanding, current access control systems are far from sufficient in addressing this issue, and social inferences impact wide aspects of system use (impression management, user privacy and security, system personalization).

Therefore, the aim of this paper was to make the case for a focused effort to be made by the Group/CSCW community to understand and address these challenges. Much more remains to be done, however, our work shows that social inference opportunities can successfully be modeled and predicted, and shows that we can establish guidelines that will help designers make informed decisions when designing social computing systems that take social inference opportunities into account.

7. REFERENCES

[1] Asbury Park Press DataUniverse Portal to Public Government Data. Retrieved Nov 10, 2011. http://www.app.com/section/DATA

[2] Biskup, J. and Bonatti, P. 2004. Controlled query evaluation for enforcing confidentiality in complete information systems. *International Journal of Information Security*, *3* (1), 14-27.

[3] Black, G., 2011. Publicity Rights and Image. *Oxford: Hart Publishing*, page 61-62

[4] Blaschke, C. and Valencia, A. 2002. Automatic ontology construction from the literature. *Genome Informatics Series*. 13201–213

[5] Brodsky, A., Farkas, C. and Jajodia, S. 2000. Secure databases: constraints, inference channels, and monitoring disclosures. *IEEE Transactions on Knowledge and Data Engineering*, *2* (6), 900-919.

[6] Cadiz, J. J., Venolia G.D., Jancke G., Gupta A., 2002. Designing and deploying an information awareness interface. CSCW 2002: 314-323

[7] Crampton, J. and Khambhammettu, H. 2008. Delegation in role-based access control *International Journal of Information Security*, *7* (2), 123-136.

[8] Cuppens, F. and Trouessin, G. 1994. Information Flow Controls vs Inference Controls: An Integrated Approach *Third European Symposium on Research in Computer Security* Springer Berlin / Heidelberg, 447-468.

[9] Dey A., Lederer S., Beckmann, C., Mankoff, J. 2003. Managing Personal Information Disclosure in Ubiquitous Computing Environments. *Technical Report CSD-03-1257*. UC Berkeley, Berkeley, CA, USA.

[10] Eagle, N., Pentland, A. 2006. Reality mining: sensing complex social systems, Personal and Ubiquitous Computing, v.10 n.4, p.255-268

[11] Fletcher, D. February 18, 2010. Please Rob Me: The Risks of Online Oversharing. Retrieved March 5, 2012. http://www.time.com/time/business/article/0,8599,1964873,00.html

[12] Goffmann E. 1059. The presentation of self in everyday life. New York, Anchor Books.

[13] Gross, R. and Acquisti, A. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society* (2005). ACM, New York, NY, USA, 71-80.

[14] Intelius People Search. Retrieved Nov 10, 2011. http://www.intelius.com/

[15] Jones, Q. and Grandhi, S.A. P3 Systems: Putting the Place Back into Social Networks *IEEE Internet Computing*, 2005, 38-47.

[16] Krumm, J. 2007. Inference Attacks on Location Tracks. *Fifth International Conference on Pervasive Computing* (Toronto, Ontario, Canada, May 13- 16, 2007)

[17] Lennon, M. Oct 12, 2010. Survey Reveals How Stupid People are With Their Passwords. Retrieved Nov 10, 2011. http://www.securityweek.com/survey-reveals-how-stupid-people-are-their-passwords

[18] Mayer, J.M., Motahari, S., Schuler, R.P. and Jones, Q. 2010. Common attributes in an unusual context: predicting the desirability of a social match. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, New York, NY, USA, 337-340.

[19] Motahari, S., Manikopoulos, C., Hiltz, R. and Jones, Q. 2007. Seven privacy worries in ubiquitous social computing. In *ACM International Conference Proceeding Series; Proceedings of the 3rd symposium on Usable privacy and security*, 171-172.

[20] Motahari, S., Ziavras, S. and Jones, Q., 2009. Preventing Unwanted Social Inferences with Classification Tree Analysis. In *IEEE International Conference on Tools with Artificial Intelligence (IEEE ICTAI)*.

[21] Motahari, S., Ziavras, S., Naaman, M., Ismail, M. and Jones, Q. 2009. Social Inference Risk Modeling in Mobile and Social Applications *IEEE International Conference on Information Privacy, Security, Risk and Trust (PASSAT)*.

[22] Motahari, S., Ziavras, S., Schular, R. and Jones, Q. 2008. Identity Inference as a Privacy Risk in Computer-Mediated Communication. *IEEE Hawaii International Conference on System Sciences (HICSS-42)*, 1-10.

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

[24] O'Leary, D.E. 1995. Some Privacy Issues in Knowledge Discovery: The OECD Personal Privacy Guidelines. *IEEE Expert: Intelligent Systems and Their Applications 10* (2), 48-52.

[25] Robertson, S. 2004. Understanding inverse document frequency: on theoretical arguments for IDF, *Journal of Documentation.* Vol. 60 Iss: 5, pp.503 – 520

[26] Rodewig, C. March 7, 2012. Geotagging poses security risks. Retrieved Macrh 10, 2012. *The Official Homepage of the United States Army*. http://www.army.mil/article/75165/ Geotagging_poses_security_risks/

[27] Samarati, P. and Sweeney, L. 1998. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and cell suppression. *Technical report, SRI International*.

[28] Schechter, S., Bernheim Brush, a. J., Egelman, S.2009. It's no secret: Measuring the security and reliability of authentication via 'secret' questions. In 30th IEEE Symposium on Security and Privacy.

[29] Schuler, R.P., Laws, N., Bajaj, S., Grandhi, S.A. and Jones, Q. 2007. Finding Your Way with CampusWiki: A Location-Aware Wiki to Support Community Building *The ACM.s Conference on Human Factors in Computing Systems CHI2007*, San Jose California, USA.

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

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

[32] Sweeney, L. 2004. Uniqueness of simple demographics in the U.S. population. *Technical report, Carnegie Mellon University, Laboratory for International Data Privacy*.

[33] Tenenbaum, J.B., Griffiths T.L., 2001. Generalization, similarity, and Bayesian inference. Behav. Brain Sci. 24: 629-40; discussion 652-791.

[34] Terry, M., Mynatt , E.D., Ryall, K., Leigh, D., 2002. Social net: using patterns of physical proximity over time to infer shared interests, *CHI '02 extended abstracts on Human factors in computing systems*, (Minneapolis, Minnesota, USA, April 20-25, 2002),.

[35] Terveen, L., and McDonald D. 2005. Social Matching: A Framework and Research Agenda. In *ACM Transactions of Computer Human Interaction.*

[36] Weiser, M. 1991. The Computer for the 21st Century. *Scientific American*.

[37] Yam, M. Jan 22, 2010. Your Top 20 Most Common Passwords. Retrieved Nov 10, 2011. http://www.tomshardware.com/news/imperva-rockyou-most-common-passwords,9486.html

[38] ZabaSearch. Free People Search and Public Information Search Engine. Retrieved Nov 10, 2011. http://www.zabasearch.com/

[39] Zetter, K.,(2008, Sept 18) Palin E-Mail Hacker Says It Was Easy. *Wired.* Retrieved March 8, 2012. http://www.wired.com/threatlevel/2008/09/palin-e-mail-ha/

[40] Zhan, J. and Matwin, S. 2006. A Crypto-Based Approach to Privacy-Preserving Collaborative DataMining *Sixth IEEE International Conference on Data Mining Workshops*, 546-550.