Enhancing Privacy using Community Driven Recommendations: An Investigation with Facebook Data

M VenkataSwamy  
@WalmartLabs,  
California, USA  
venkat.taku@gmail.com

Nitin Agarwal  
University of Arkansas at Little Rock,  
Little Rock, AR, USA  
nxagarwal@ualr.edu

Sri Ramaswamy  
ABB Corporate Research,  
Bangalore, India  
srini@ieee.org

ABSTRACT

User specific information in social media is sensitive and subject to privacy. Continuously changing privacy policies and configuration procedures in social media require users to constantly educate themselves of the changes. A collective intelligence driven approach, known as Collective-Context Based Privacy Model (C-CBPM) has been developed that recommends privacy policies based on community and trust gleaned from social network information. By defining user-specified contexts, C-CBPM advances the existing content, user, or role-based privacy models. This research examines the efficacy of C-CBPM using Facebook data comprising of 957,359 users, 957,357 connections, and 32,176 communities. Objective trust and privacy risk assessment measures are developed. Results indicate promising findings with 83% correct recommendations. Out of the 17% incorrect recommendations, almost all (i.e., 99.24% of the incorrect recommendations) incur only 25% risk and only 0.018% incur 100% or maximum risk, in the worst-case scenario. The results demonstrate the feasibility of the C-CBPM in real-world for community driven privacy recommendations.

Keywords

Privacy; context; context based privacy model; CBPM; trust; collective intelligence; community; social media; Facebook.

INTRODUCTION

With the advent of social media websites such as Facebook, Myspace, Twitter, and social health websites including ‘Patientslikeme.com’ that help people with health conditions connect with others with similar conditions, user participation has skyrocketed. According to recent statistics, Facebook has 1.11 billion monthly active users1 as of March 2013. The social media sites have afforded extreme convenience in sharing information over the Internet leading to a vast ocean of user-generated content including even sensitive financial or health-related data. Although the websites provide their data sharing policies and allow the users to configure privacy settings, these configuration procedures are often quite intricate assuming the users at par with the technological know how (Dwyer, Hiltz, and Passerini, 2007). In recent years, data/information centers in industries are facing increasingly stringent federal regulations regarding information privacy, forcing them to constantly modify the privacy policies (Rizk, Marx, Schrepfer, Zimmerman, and Guenther, 2009). Further, social media sites frequently release new products and services that could leave even the most skillful social media gurus mystified. For instance, the much-debated Facebook’s new graph search feature is a case in point2.3 Evolving privacy policies and the paradigmatic shift of information ownership and accessibility pose interminable challenges for the users to keep up. As a result, users may inadvertently grant access to their personally identifiable information as mentioned in (Lo, 2010), leading to privacy threats. Consequently, malicious efforts to compromise users’ data are on the rise exacerbating the problem.

The existing works on developing privacy models are essentially content, user, or role-based. Such models are limited in adapting to the evolving privacy policies and the changing landscape of information ownership and accessibility, especially in the social media ecosystem. This warrants thinking beyond the existing privacy models and considers the context, in which the information is accessed, as a central tenet for enhancing privacy models (VenkataSwamy, Ramaswamy, and Agarwal, 2010). In this research, we implement a context based privacy model that allows users to define their own contexts and specify fine-grained policies. Considering the needs of a common user, we develop a framework extending the CBPM to

---

automate privacy recommendations in social media domain leveraging the principle of collective intelligence (Surowiecki, 2005) and trust measurement from social network information. The framework is known as Collective-Context Based Privacy Model (C-CBPM) (VenkataSwamy, Agarwal, and Ramaswamy, 2012). The C-CBPM framework allows users to selectively donate their privacy policies. Users looking for configuring their privacy policies could selectively adopt from the donated privacy policies. C-CBPM helps recommend privacy policies from the available donations. A collective intelligence based methodology is used to generate the recommendations. The methodology considers the user’s trust with the donors and the likelihood of adoption governed by the user’s self-assigned threshold while generating recommendations. The collective intelligence principle helps weed out donations from less trusted users making the model more efficient and less vulnerable to malicious donations. Further discussion on the CBPM and its extension C-CBPM model is provided in the next section.

The efficacy of C-CBPM is validated on a real world dataset from Facebook, which is one of the most popular social media sites. The data comprises of 957,359 users with 957,357 connections and 32,176 distinct communities. The data contains the privacy settings of all the users allowing us to study the behavior of C-CBPM for a large number of users ranging from liberal (‘share all’) to conservative (‘share nothing’) privacy outlook. Evaluation measures including recommendation accuracy and ‘risk factor’ are introduced to investigate the performance of the C-CBPM framework on the given dataset.

Key contributions of the paper are as follows,
- Motivated the need for thinking beyond content, user, or role-based privacy models and consider context of data access in enhancing privacy models for social media domain,
- Developed a Context Based Privacy Model (CBPM) and extended the framework to a Collective-Context Based Privacy Model (C-CBPM) to automate privacy recommendations,
- Developed a collective intelligence based methodology for recommendation generation leveraging user’s trust relationships,
- Studied the efficacy of the C-CBPM framework in real-world data from Facebook comprising of 957,359 users with 957,357 connections and 32,176 distinct communities, and
- Developed measure to assess trust in the Facebook network and objective evaluation metrics including recommendation accuracy and ‘risk factor’ to investigate the performance of the C-CBPM.

Next, we discuss the state-of-art privacy models, which led us to develop CBPM and C-CBPM frameworks.

RELATED WORK

State-of-art privacy models are discussed in this section to derive the benefits of the C-CBPM framework. The traditional models are primarily categorized into two groups, Lattice models and Take-Grant models.

Lattice Models

Lattice models assign a class to resources and each class is associated with set of policies. Following are two lattice models.

1. **Bell-LaPadula Model** (LaPadula and Bell, 1996): Objects are classified into four sensitivity levels: Top-Secret, Secret, Confidential and Unclassified. A subject can only access objects at certain levels determined by his/her security capability. The model is not capable in restricting illegal information flow within a sensitivity level.

2. **Ethical Wall Model** (Brewer and Nash, 1989): The main objective of the model is to avoid conflict of interest problems. No two entities in a class of conflict of interest exchange information. The model does not benefit in situations where mutual cooperation of objects in a class of conflict of interest is needed.

Take-Grant Models

In these models, a request to access a resource is authorized by requestor’s credentials. Following are some of the Take-Grant models.

1. **Information Flow Model** (Foley, 1990): Access to an object is granted for a subject only if the subject promises the information will not reach unauthorized entities from it. Subsequently the flow of the information is constrained. The model involves complex issues while incorporating in real time. The trace of the information is needed to make a decision.

2. **Role Based Access Control (RBAC)** (Ferraiolo and Kuhn, 1992): It uses the role of the requesting entity to decide whether to give access or to revoke. Identifying the role of the requester is not sufficient in taking decisions all cases.

3. **Context-Based Access Control**: Subject’s context plays major role rather than the subject or object’s security capability (Hong, Suh, and Kim, 2009; Kapsalis, Hadellis, Karelis, and Koubias, 2006). Context aware systems take advantage of a consultant’s context to analyze the activities at each layer of the communication.

Unlike user or role identity, privacy aware systems (pawS) in ubiquitous computing use privacy policies to make decision on information access (Langheinrich, 2002). According to Principle of minimum asymmetry, pawS should minimize asymmetry of information flow from and to a data owner (Jiang, Hong, and Landay, 2002). It is advised to reduce the asymmetry by
increasing information flow from the information owner (Kulkarni and Tripathi, 2008). CBPM leverages contexts to control the information to reduce data flow from users (VenkataSwamy, Ramaswamy, and Agarwal, 2010). Though the context aware systems are complex, CBPM made it simple by disseminating information based on their privacy policy. Such dissemination helps incorporating fine-grained privacy policies on each atomic piece of information. The systems discussed above do not provide comprehensible interface to users to define their privacy policies. C-CBPM offers safe privacy policies deriving from community support. C-CBPM is first of its kind to leverage the cognitive surplus of the community in strengthening the privacy policies. Next, we provide a brief discussion of CBPM and C-CBPM frameworks.

**CONTEXT BASED PRIVACY MODEL (CBPM)**

Social network services such as Facebook deals with sensitive information from individuals and subject to privacy mitigations. In this section, we discuss the context based privacy model (CBPM) that addresses the privacy issues. The model is defined for a set of contexts and a set of data sets. A data set is a collection of data elements where a data element is an atomic piece of information. A context is defined as an abstract state of a subject and is rich in expressing status of an entity compared to user or role identity. A privacy policy is defined as a rule to share or not share a data element within a given data set for a given context. The privacy policies for all the data sets and for all the contexts constitute a matrix called CBPM matrix. An entry in a CBPM matrix can be represented as:

\[(Ci,Dj) = 0 \text{ or } 1\]

where, \(Ci\) is a context, and \(Dj\) is a data set

The semantics of the entry is: if \((Ci,Dj)\) is ‘0’, no data element in data set \(Dj\) is accessible to a requester in \(Ci\) and if it is ‘1’, all the data elements in data set \(Dj\) are accessible in context \(Ci\). The CBPM approach assumes each user owns a matrix that defines his/her specified access policies for different contexts identified by the user. Creating a new matrix from scratch could be taxing for the user, especially for a naïve user, also known as the usability problem. A collective intelligence based approach could be used to help a naïve user by recommending a CBPM matrix. This technique is called as Collective Context Based Privacy Model, or C-CBPM, which is discussed next.

**COLLECTIVE- CONTEXT BASED PRIVACY MODEL (C-CBPM)**

The primary objective of C-CBPM is to recommend a CBPM matrix for a user through community support. Existing users voluntarily support a naïve user to develop his/her information sharing policies. Users are allowed to donate their privacy policies by sharing the whole or a part of their CBPM matrix with the community. Naïve or new users adopt a part of the privacy settings or completely imitate what donors have shared, thereby addressing the usability problem. Naïve users adopt such matrices depending on the combination of one or more characteristics of the donor, including affiliation, trust, social ties, structure, and role. Donors, who are ready to donate, bestow their CBPM matrices in C-CBPM donation pool. The donor could select parts (rows and columns) of his/her matrix to make a donation. The model assumes there are some donations available in C-CBPM donation pool for a naïve user when he/she joins a community. Adoption of a matrix from C-CBPM involves three phases.

Phase 1: Requesting Donations: Adopter who needs a CBPM matrix initiates the process by broadcasting request to available donations along with his credentials.

Phase 2: Processing Requests: Upon receiving a request, C-CBPM authorizes the requests looking at adopter’s credentials and grants whole or part of valid CBPM matrix donations.

Phase 3: Aggregation: Adopter expects responses from C-CBPM with CBPM matrix donations. Received matrices are aggregated to construct a new CBPM matrix. Default matrix is created when there is no donation available.

A typical scenario of C-CBPM is shown in Figure 1. Aggregation phase is critical in order to prevent malicious privacy policies derived from malicious donations. The aggregation matrix for an adopter \(X\) is obtained using the following formula.

\[M_X = M_X + \text{OR} [t(A,X)M_A + \text{OR} t(B,X)M_B + \text{OR} t(C,X)M_C]_p\] (1)

Where \(M_X\) is a donation matrix from user \(A\),

\(M_A\) is the masked matrix for the donation \(M_A\),

\(t(A,X)\) represents how much the user \(X\) trusts user \(A\),

and ‘+\text{OR}’ and ‘[ ]_p’ or matrix operators.

VenkataSwamy et al. Enhancing Privacy in Facebook


3
Eq (1) gives the derived matrix for a naïve user ‘X’ from available donations.

The masked matrix is defined as

\[ M_A(i,j) = \begin{cases} M_A(i,j) & \text{if adopter accepts donation for the entry of context ‘i’ and dataset ‘j’} \\ 0 & \text{otherwise} \end{cases} \]

The matrix binarization operation \([ \mu ]\) is defined as

\[ [t(A,X)M_A(i,j)]_{\mu} = \begin{cases} 1 & \text{if } t(A,X) * M_A(i,j) > \mu \\ 0 & \text{otherwise} \end{cases} \]

‘\( \mu \)’ is a binarization threshold. The threshold value ranges from ‘0’ to ‘1’.

And the matrix aggregation operator ‘+OR’ is defined as

\[ (M_A +_{OR} M_B)(i,j) = \begin{cases} 0 & \text{if } M_A(i,j) = M_B(i,j) = 0, \\ 1 & \text{otherwise} \end{cases} \]

It is trivial to show the ‘+OR’ addition operation a commutative, associative, transitive and closed.

Each donation is associated with a trust value denoted as \(t(A,X)\). Trust values always lie in \([-1,1]\).

Initial trust values computed using several attributes followed by evolving trust. Trust values can be learned after several donations and adoptions using the models proposed on trust (Agarwal and Liu, 2009) and expert identification (Agarwal, Liu, Tang, and Yu, 2009). Donors bequeath their matrices for adoption. The more their matrices are adopted, the more they will be rewarded by earning trust as recognition of donation. Users who are especially active in donating privacy matrices and are widely adopted are called “Expert Donors” or “Consultants”.

Eq (1) gives us derived matrix for user ‘X’ from available donations. The detailed discussion of the C-CBPM model is presented in (VenkataSwamy et al., 2012). Next section implements a platform to analyze behavior of C-CBPM in Facebook.

**EXPERIMENTS**

In this section, we present implementation details and study the performance of C-CBPM framework on a real world social network. Although the framework could be deployed in any social media system, we focused on Facebook due to its unparalleled popularity among social media users. Next, we discuss the details of the data.

**Dataset**

Founded in 2004, Facebook rapidly became one of the most popular social network services. There are one billion monthly active users as of December 2012, out of which approximately 81% are outside the U.S. and Canada. In an attempt to study such a massive online social network, researchers used sampling techniques that have been demonstrated to represent the
The entire data without loss of generality (Gjoka, Kurant, Butts, and Markopoulou, 2010). The sampled data comprises of 957,359 users with 957,357 connections and 32,176 distinct communities. The data is made available for public\(^4\).

The dataset has two aspects: (1) user profile information, and (2) user friendship information. The user profile information contains the id of the user, total number of friends, privacy settings, and ids of the communities to which the user is affiliated. The privacy settings of a user consist of a binary choice between shared or not-shared for four basic attributes:

1) Whether the user could be added as friend,
2) Whether the user shared his/her photo thumbnail,
3) Whether the user’s friends could be viewed, and
4) Whether a message could be sent to the user.

We construct the CBPM matrix of a given user by using the four privacy attributes. If an attribute were shared the CBPM matrix would have a ‘1’ otherwise it would have a ‘0’ for that attribute. To assess the privacy outlook of the users, we study the distribution of their shared attributes. We computed a histogram that illustrates the number of shared attributes for the users. The histogram is presented in Figure 2. It was found that 92.8% of the users shared all the 4 attributes. This indicates that the users in this data are more liberal than conservative in their privacy policies. A similar observation was made by the researchers in (Lo, 2010) from a survey of 80 Facebook users.

The community affiliations in the data correspond to the regional, school, or workplace Facebook communities of which the user is a member. It was observed that most of the users in the data were either members of a single community or no community at all. Further, no user had more than five different community affiliations. A histogram of the distribution of community affiliation is illustrated in Figure 3.

To further understand the community structure, we investigated the size of each of the 32,176 distinct communities. The size of a community is estimated by the number of its members. A log-log plot of the community size distribution is presented in Figure 4. We observed that most communities are relatively small (less than 100 members) and few communities have more than 1000 members. This finding suggests that community size distribution in this data also obeys a power law distribution, or the Long Tail distribution, like all other social networks as also reported in (Lancichinetti and Fortunato, 2009).

\(^4\) Data used for this research is made accessible by (Gjoka et. al., 2010) and can be downloaded from http://odysseas.calit2.uci.edu/doku.php/public:online_social_networks#facebook_social_graph (Last accessed: May 12, 2013).
The other aspect of this data consists of the users’ friendship information. For every user, the list of the friends is provided. For 957,359 users there are in total 957,357 connections indicating an extremely sparse network among users. We study the network structure using a log-log degree distribution presented in Figure 5, which demonstrates a power law degree distribution or a scale-free network structure for the Facebook data like other social networks. Both the social network information and the community affiliations for a user are used to measure trust and recommend privacy using the C-CBPM framework as described next.

![Figure 4. Distribution of the Community Size in the Facebook Data](image1)

![Figure 5. Distribution of the Users’ Social Network in the Facebook Data](image2)

**Parameters**

Here we introduce three measures; 1) Privacy Index to measure privacy outlook of a user, 2) Trust to measure trust between two users, and 3) Risk Factor to measure risk from the model.

**Privacy Index (PI)**

Each user in the dataset possesses a privacy setting (privacy policies) in terms of the CBPM matrix. Strength of the privacy policy of a user is measured through “Privacy Index” defined as follows,

\[
\text{Privacy Index (PI)} = \text{Fraction of zeros in the CBPM matrix}
\]

The value of PI lies between 0 and 1. A value of ‘0’ would correspond to the most liberal user (‘share everything’) and a value of ‘1’ would correspond to the most conservative user (‘share nothing’). In the analysis, we associate every user with his/her PI value.

---

5 A degree of a node denotes the number of connections or friends.
Trust

Trust between any two users, say ‘A’ and ‘X’, can be assessed using, (1) Implicit inferences – feedbacks, comments, user inputs (2) Explicit inferences – profile information, social network, communication patterns, or (3) Similarity – similarity between the users information. Due to the lack of implicit and explicit information, similarity between community affiliations of a pair of users is used to measure trust (Einwiller, Geissler, and Will, 2000). A trust score from the similarity of community affiliations is computed for every pair of connected users (friends) in the Facebook network. The more common community affiliations friends share the higher the trust between them. If a user A is a friend of X and shares community affiliations then A trusts X and vice versa. The strength of A’s trust on user ‘X’ is defined as,

\[
\text{Trust } (t(X,A)) = \frac{\text{Fraction of common communities between X and A}}{\text{Communities affiliated} \cap \text{Communities affiliated}}
\]

\[t(X,A) = \frac{\text{Communities affiliated} \cap \text{Communities affiliated}}{\text{Communities affiliated}}\]

The more common community affiliations friends share the higher the trust between them. If a user A is a friend of X and shares community affiliations then A trusts X and vice versa. The strength of A’s trust on user ‘X’ is defined as,

\[
\text{Trust } (t(X,A)) = \frac{\text{Fraction of common communities between X and A}}{\text{Communities affiliated} \cap \text{Communities affiliated}}
\]

\[t(X,A) = \frac{\text{Communities affiliated} \cap \text{Communities affiliated}}{\text{Communities affiliated}}\]

A trust score from the similarity of community affiliations is computed for every pair of connected users (friends) in the Facebook network. The more common community affiliations friends share the higher the trust between them. If a user A is a friend of X and shares community affiliations then A trusts X and vice versa. The strength of A’s trust on user ‘X’ is defined as,

\[
\text{Trust } (t(X,A)) = \frac{\text{Fraction of common communities between X and A}}{\text{Communities affiliated} \cap \text{Communities affiliated}}
\]

\[t(X,A) = \frac{\text{Communities affiliated} \cap \text{Communities affiliated}}{\text{Communities affiliated}}\]

The trust values of a user are normalized across its neighbors. Trust values always lies between 0 and 1. We computed trust scores for every relationship in the given Facebook dataset and construct a histogram of the trust values over relationships. The histogram is presented in Figure 6. From the Figure 6, it is observed that there are more relationships with relatively lower trust values and fewer relationships with higher trust.

**Risk Factor (RF)**

For every user, the C-CBPM framework generates a recommendation of privacy settings. The recommended CBPM matrix is called as the “Observed Matrix (OM)”’. The actual privacy policy of the user is labeled as the “Expected Matrix (EM)”. OM and EM are compared to assess the risk incurred by the user by adopting the recommended matrix. There would be no risk if OM and EM match perfectly. For all other cases, a risk factor (RF) is defined to quantify the risk incurred due to incorrect recommendation. Risk is incurred when an unshared item is recommended to be shared, meaning a ‘0’ in EM is converted to a ‘1’ in OM. Then, RF is defined as,

\[
\text{Risk Factor } (RF) = \frac{\text{Fraction of entries flipped to 1 from 0}}{\text{Number of shared entries}}
\]

In other words, Risk Factor is how many times an unshared data set is made accessible to all contexts. Mathematically,

\[
\text{Risk Factor } RF = \frac{\sum_{i,j} (\text{OM}(i,j) \land \text{EM}(i,j))}{\sum_{i,j} 1}
\]

**Runs**

To simulate the scenario of a user, called an adopter (A), joining Facebook and seeking a privacy policy (CBPM matrix) recommendation, we assume the adopter is from the given Facebook dataset itself. The adopter being a user from the given Facebook dataset embodies a CBPM matrix and the CBPM matrix is referred as an Expected Matrix (EM). There are two reasons behind selecting adopter from the given dataset. One, every user has a CBPM matrix (EM) which can serve as ground truth for validation. Other, trust scores can be computed between an adopter ‘A’ and donors using the above discussed technique. We run C-CBPM model for an adopter assuming the adopter do not have CBPM matrix and ready to

---

**Figure 6. Distribution of the Trust Values and Relationships among the Users in the Facebook Data**
accept a recommendation. Every friend of an adopter in the Facebook dataset is a donor and donates his/her CBPM matrix with some trust score. C-CBPM framework derives a recommendation CBPM matrix from the donations for a given threshold value (µ). The recommendation matrix is the Observed Matrix (OM) here. A recommendation (OM) is said to be “True” if OM is identical as EM and “False” otherwise. If a recommendation is false, then there is a possibility of risk for the adopter from accepting the recommendation, i.e. risk of OM over EM, which is measured through RF. Because of four elements in CBPM matrix for a user in the given Facebook dataset, there are only 5 possible RF values, which are 0, 0.25, 0.5, 0.75, and 1.0. Therefore, a recommendation can be True or False with a RF value.

The process is repeated for each user in the Facebook dataset by taking a user as an adopter at an instance for each possible threshold value ‘µ’ with an interval of 0.2. Each time a recommendation is computed. We observed cases where there is no recommendation from the model because of lack of trusted donations and/or the donations not meeting the given threshold limit. We discarded the runs yielding no recommendations in the following discussion. The distribution of the recommendations over True, False and different values of RF is computed and is discussed in the following section.

Table 1 is compiled to give a quick reference to the discussed parameters and evaluation measures. Next, we illustrate the experimental setup.

<table>
<thead>
<tr>
<th>Term</th>
<th>Notation</th>
<th>Value Range</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Privacy Index</td>
<td>PI</td>
<td>0-1</td>
<td>Fraction of zeros in matrix</td>
</tr>
<tr>
<td>Threshold</td>
<td>µ</td>
<td>0-1</td>
<td>Threshold limit for binarization</td>
</tr>
<tr>
<td>Risk Factor</td>
<td>RF</td>
<td>0-1</td>
<td>Fraction of entries flipped to 1 from 0</td>
</tr>
<tr>
<td>Observed Matrix</td>
<td>OM</td>
<td>-</td>
<td>CBPM matrix estimated using C-CBPM framework</td>
</tr>
<tr>
<td>Expected Matrix</td>
<td>EM</td>
<td>-</td>
<td>The actual CBPM matrix</td>
</tr>
</tbody>
</table>

### RESULTS AND DISCUSSION

In this section, we study the performance of C-CBPM by first examining the accuracy considering the true and false recommendations and further analyzing the risk incurred due to false recommendations. The distribution of true and false recommendations is presented in Figure 7. The analysis presented here is conducted on users with neighbors having at least one common community affiliation that amounts to 465,502 users. The remaining 491,857 users are discarded from this analysis. Following observations could be made from Figure 7.

1. The number of true recommendations is much higher than the number of false recommendations. It means the C-CBPM model recommendations are true for 83% to 90% of the users while the model failed for 17% to 10% users. The percentage of the users getting true recommendations proves the accuracy of the model in the given Facebook dataset.
2. The percentage of true recommendations slightly decreases with increasing threshold value. Simultaneously, the percentage of false recommendations increases with the same pace. The likelihood of recommending an accurate (true) recommendation for a user decreases with increasing threshold. The threshold represents the quantity of required support to adopt a privacy policy. Higher the threshold means higher the support needed from trusted donors. It is slightly unlikely to encounter such higher support from trusted donors when the threshold is high and the recommendations lead to be false. Therefore, higher the threshold implies slightly lower the accuracy.

Though the percentage of false recommendations is low (between 10% and 17% for different threshold values), we further investigated the false recommendations to study risk posed to the users who encountered false recommendations.

As discussed earlier, risk is measured by Risk Factor (RF), which is computed for each false recommendation. The histogram of recommendations for each possible value of RF is illustrated in Figure 8. Here, RF=0 denotes no risk from the recommendation, i.e. no 0’s are turned into 1’s or in other words none of the inaccessible data elements are made accessible. A false recommendation with RF=0 is encountered because there are 1’s that are turned into 0’s, which do not pose any privacy risk. We have 5 possible RF values that are 0, 0.25, 0.5, 0.75, 1.0 corresponds to 0%, 25%, 50%, 75% and 100% risk, respectively.
Let us consider extreme threshold value scenario for the discussion, i.e. threshold = 0 and threshold = 1,

1. **Threshold = 0**: From the chart in Figure 8, it can be observed that there is no false recommendation with zero risk. All the false recommendations pose at least 25% risk. The number of recommendations with 100% risk is only 8, i.e. only 8 users out of the 465,502 users experienced 100% risk and therefore the probability of experiencing a 100% risk recommendation is 0.0000172, which could be considered negligible.

2. **Threshold = 1**: In this scenario, the false recommendations with no risk are as many as the false recommendations with at least 25% risk. Though not clearly visible in the plot, there are less number of false recommendations with 25%, 50% and 75% risk compared to the case with threshold=0. We also observed that there is no increase in number of recommendations with 100% risk compared to scenario with threshold=0.

In summary, the number of false recommendation with zero risk increases with increasing threshold while the false recommendations with 25%, 50% and 75% risk decrease. The risk decreases with increasing threshold. From these observations, the increase in number of false recommendations in Figure 7 increased by the increase of the number of false recommendation with risk. Therefore, the increase in false recommendations in Figure 7 does not raise risk but decrease risk at the cost of accuracy.

Putting the observations from the both the charts in Figures 7 and 8 together, there is a tradeoff between having accurate recommendation over risk with varying threshold. And, the number of the true recommendations (83% in the worst case) is significantly higher than the false recommendations (17%), which demonstrates the accuracy of the C-CBPM model. Therefore, privacy policy recommendations learned from communities enhance privacy of the users in Facebook. Only the 8 out of 465,502 recommendations posed 100% risk and showed the negligible risk of the model.
CONCLUSION

In this research, we implemented a context based privacy model that allows users to define their own contexts and specify fine-grained policies. Considering the needs of a normal user, we develop a framework extending the CBPM to automate privacy recommendations in social media domain leveraging the principle of collective intelligence and trust measurement from social network information. The framework is known as Collective-Context Based Privacy Model (C-CBPM). C-CBPM framework allows selective donations and adoptions for the users. The recommendations are generated using a collective intelligence based methodology that considers the user’s trust with the donors and the likelihood of adoption governed by the user’s self-assigned threshold. The efficacy of C-CBPM is demonstrated on a real world dataset from Facebook comprising of 957,359 users with 957,357 connections and 32,176 distinct communities. The investigation further shows the framework’s ability to achieve high accuracy for a large number of users with diverse privacy outlook from liberal (‘share everything’) to conservative (‘share nothing’). The results indicate promising findings with 83% correct recommendations. Out of the 17% incorrect recommendations, almost all (i.e., 99.24% of the incorrect recommendations) incur only 25% risk and only 0.018% of the incorrect recommendations incur 100% or maximum risk, in the worst-case scenario.

As our immediate next steps, we plan to implement the framework in other social media sites and evaluate its performance. In the current framework, trust among users is estimated based on common community affiliations, which we plan to revisit and explore other complementary sources of measurements. Another intriguing direction for further investigation is to consider evolving trust relationship based on past donation-adoption behaviors of the users. We plan to analyze the feasibility of deploying the C-CBPM framework in cloud-based information systems and further study the implications of trust on user information privacy and security vis-à-vis service-level interactions.

ACKNOWLEDGMENTS

This work was supported, in part, by grants from the National Science Foundation (under Grant Nos. CNS-0619069, EPS-0701890, OISE 0729792, CNS-0855248, EPS-0918970, and IIS-1110868) and Office of Naval Research (Award number: N000141010091). The views expressed in this article are solely those of the authors and do not necessarily reflect the official policies or positions of their employers.

REFERENCES


