Privacy Attacks in Social Media Using Photo Tagging Networks: A Case Study with Facebook

João Paulo Pesce
UFMG
Brazil
jpesce@dcc.ufmg.br

Diego Las Casas
UFMG
Brazil
diegolascasas@psi.grad.ufmg.br

Gustavo Rauber
UFMG
Brazil
rauber@dcc.ufmg.br

Virgílio Almeida
UFMG
Brazil
virgilio@dcc.ufmg.br

ABSTRACT

Social-networking users unknowingly reveal certain kinds of personal information that malicious attackers could profit from to perpetrate significant privacy breaches. This paper quantitatively demonstrates how the simple act of tagging pictures on the social-networking site of Facebook could reveal private user attributes that are extremely sensitive. Our results suggest that photo tags can be used to help predicting some, but not all, of the analyzed attributes. We believe our analysis make users aware of significant breaches of their privacy and could inform the design of new privacy-preserving ways of tagging pictures on social-networking sites.

Categories and Subject Descriptors

H.1.2 [**Models and Applications**]: User / Machine systems—*human factors, human information processing*; K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy*; K.6.5 [**Management of Computing and Information Systems**]: Security and protection

General Terms

Measurement, Human factors, Security

Keywords

Privacy, Online social networks, Photo tags

PSOSM'12 April 17 2012, Lyon, France

1. INTRODUCTION

The first decade of the 21st century saw the popularization of the Internet and the growth of web services that facilitate participatory information sharing and collaboration. Specifically, Social Network Sites (SNS), allow users to interact with others in an unprecedented way. Recently, SNSs, more than just web applications, have become part of human culture and how society interacts. News agencies, big and small companies, governments, famous personalities and the general population all use SNSs to interact with each other. With over 600 million users and 30 billion pieces of content shared each month, Facebook has stood out as the most popular SNS in the world[1] and the website where people spend the most time[2]. Users spend 700 million minutes / month on Facebook[3]. Sharing news, photos, personal taste and information with friends and family has never been so easy. This luxury of technology and services comes along with concern of user privacy. Privacy-related issues with Facebook have been constantly appearing on international press either because of the company's privacy policy or because of user's unawareness of content sharing consequences.

As shown by researchers, a simple exposure of date and place of birth of a profile in Facebook can be used to predict the Social Security Number (SSN) of a citizen in the U.S. [9]. Sometimes, by simply revealing their friends list, users might be revealing much more. For example, through the use of prediction algorithms it is possible to infer private information that was previously undisclosed [10, 18, 24]. Photo albums may also contain sensitive information about the user, like places she usually goes to, whether or not she is on vacation and who are some of her closest friends and family members. Sometimes sensitive information even comes embedded in the photo as metadata [6]. They may also be accompanied by more information that could be exploited, like captions, comments and photo tags; marked regions that identify people on the photo. Even if the individuals in a photo are not explicitly identified by photo tags, the combination of publicly available data and face recognition software can be used to infer someone's identity [1].

[1]http://www.ebizmba.com/articles/social-networking-websites
[2]http://techcrunch.com/2011/02/07/comscore-facebook-keeps-gobbling-peoples-time/
[3]http://www.facebook.com/press/info.php?factsheet

This kind of problems are defined as collateral damage: users unintentionally put their friends or even their own privacy at risk when performing actions on SNSs such as Facebook [13].

In this paper we expose some of the privacy issues with photo albums and, more specifically, the use of photo tags to predict information about Facebook users. Our objective is to show that the use of photo tagging can enhance accuracy of attackers aiming to predict personal user attributes. We also analyze a possible reason for it by linking the higher prediction levels to the presence of Simmelian ties. Our results may raise awareness of the kinds of information transmitted by photo tags in SNSs, thus avoiding collateral damages.

In Section 2, we present some related works to privacy and attribute prediction in SNSs. In Section 3, we introduce the photo tagging feature and list some direct privacy-related issues with it. In Section 4, we introduce our data collection effort. In Section 5, we present the inference algorithms that could be used by attackers and in Section 6 we show our results. Finally, in Section 7 we discuss these results in terms of social relations.

2. RELATED WORK

With the growth of SNSs in the last decade, study and characterization of privacy in these services has been a central theme in many studies on the field of network science, computational social science, and human-computer interaction. Gross and Acquisti [9] studied patterns of information revelation on Facebook and revealed that, even though personal data is widely provided, users do not tend to limit the SNS's permissive default privacy settings. Krishnamurthy and Wills [14] characterized and measured various privacy aspects across different SNSs using the concept of bits (or pieces) of shared information. They also exposed that, much like traditional websites, third-party domains track users activities in SNSs. Contrary to widespread assumptions, Boyd and Hargittai [3] showed that youth do care about privacy and actively use privacy settings in SNSs like Facebook.

More related to our work, some researchers have studied ways of inferring otherwise private bits of information from users. He et al. [10] used Bayesian network to model the prediction problem and developed a variety of algorithms to predict private user attributes. Zheleva and Getoor [24] took in consideration SNS's groups and used classification algorithms to improve prediction accuracy. Mislove et al. [18] have looked at the problem from two distinct perspectives: globally (given some attributes on the network, infer the attributes of the rest of the network) and locally (given some people in a certain community, infer who else is on that community). Lindamood and Kantarcioglu [16] focused on finding the political affiliation of their subjects based on either profile information or friendship links. They concluded that it is best, from a privacy stand point, to conceal more profile attributes than friendship links.

Social network analysis usually works with either: *complete networks*, *e.g.*, a network containing all ties in a defined population or *egocentric networks* (ego-networks for short), *e.g.*, all ties that certain individuals may have [4]. In this work, we model ego-networks from user's friendship links and use photo tags to improve the prediction of private attributes (such as nationality, age, city and ideally any attribute the SNS might provide), and we do so for the first time: there is no prior study that exploits photo tagging feature for attribute prediction.

3. PHOTO TAGGING

The ability to upload and share photo albums on Facebook was launched in October 2005, when the site had about 5 million users[4]. By then, photo hosting was already exploding on the Internet and other sites which offered photo hosting services were already quite popular, like MySpace and Flickr. Nonetheless, the simple interface and mass adoption allowed Facebook to become the number one online photo service by late 2009, with more than 30 billion photos uploaded on the site [11].

Since its start, the service allowed users to tag their friends and comment on photos. An example of the feature is depicted in Figure 1. When tagged, friends received e-mail alerts, driving lots of traffic to the website. Within one month of its launch, 85% of the subscribed users were tagged at least once [11].

It is not hard to imagine the privacy consequences made possible by this service. People upload nowadays more than 3 billion photos each month[5] and add more than 100 million tags to photos on Facebook every day[6]. Facebook also reached a new record with 750 million photos uploaded over the first weekend of 2011[7]. Despite its enormous popularity, photo tagging has been the target of many critics regarding users privacy. Even though researchers have proposed improvements to the photo tagging process [2] and policies have changed since its first implementation, the system is far from ideal. Next, we describe direct privacy breaches related to photo tagging.

3.1 Privacy Concerns

3.1.1 *Who took this picture?*

Most Facebook users want to keep some level of controlled exposure so they can interact better with others. Seeing that tagging is a popular and convenient feature, users have tended not to disable it. It seems reasonable to assume that users are not willing to expose their friends to obviously embarrassing situations, and those who do, are doing it with their own responsibility. However, it is not always clear which kind of impression a user is trying to convey by his Facebook profile. So, for example, if Alice has added some co-workers as her friends, and wants to convey a professional identity, she might feel embarrassed when her friend Bob tags her in a picture of her long-forgotten freshman years. Of course, regular users can promptly delete their tags, but this might lead to socially unpleasant situations [2].

3.1.2 *Identification of users without a profile picture*

Some users might not want to be identified visually on Facebook. To do that, Bob chooses not to upload a profile picture and not to upload any pictures of him on his albums. However, his friend Carl uploads a photo of their last encounter and adds a tag of him. Without being asked, he can be instantly identified by anyone with access to the album, which may include all of Carl's friends or even everyone on

[4]http://www.facebook.com/press/info.php?timeline
[5]http://www.insidefacebook.com/2010/02/15/new-facebook-statistics-show-big-increase-in-content-sharing-local-business-pages/
[6]http://blog.facebook.com/blog.php?post=467145887130
[7]http://techcrunch.com/2011/01/03/facebook-users-uploaded-a-record-750-million-photos-over-new-years/

Figure 1: Example of photo tags in a Facebook photo. Three out of the four people are tagged. When the user hovers the mouse over the tagged faces their names are revealed and clicking them redirects to the corresponding user's profile.

Facebook, depending on the album's privacy setting. Besides, it will appear on his own profile, frustrating whatever intention he had of staying out of sight. Supposing he does not log in so often, it may take a while before he deletes the tag. Of course, this is a case in which completely disabling photo tags is a plausible choice. But to infrequent users, the constant changes in Facebook's interface and features often make them unaware of their profile's exposure.

3.1.3 Unintended audience

Seeing that the tagged picture conveys personal information directly linked to its tagged profile, one can ask to whom the "ownership" of the photo should be attributed. Anyone with access to the album has also the option to share it, thus broadening its exposure. Comments, links and tags can flow through the news feed of a reasonable amount of unknown or unwanted users. In the current tagging mechanism, the tagged user has no means to control the degree of exposure her pictures are getting, since the "owner" is another user. Thus, although photo tags point exactly to the fact that the user is personally linked to the photo, this does not grant the user any right over it besides denying this link. Tags can also grant access to private information in the opposite way: someone without access to a certain picture can see it if she is tagged in it, even when the tag does not actually point to anything linked to her (e.g., her face does not appear in the photo).

3.1.4 Naming non-users

It is possible to tag someone who has no Facebook profile. Unlike tagging users, this does not link to any other information source, so it should not be as relevant. However, it is important to state that, in this situation, the tagged person has no direct control of his exposure (i.e. cannot delete the tag), and the tag may still brings valid information to third party users and systems.

3.2 Social Ties

There is still another issue brought forth by photo tagging in its relation to user privacy. Photo-tags can threaten privacy burdens in an indirect way, by pinpointing the nodes in the social graphs on which privacy-attacking algorithms may extract information, thus enhancing their accuracy. Previous work has already shown that the social graphs drawn by the Facebook friend's list are not the most accurate way of representing the user's social environment [22]. It is reasonable to claim that our ability to keep social connections are limited in some account. Previous work has given a properly scientific explanation for this fact by linking the extent of our social networks to limits in our cognitive power and social abilities [20]. We would naturally try to restrain social permissions to inner circles in our cognitive limits, which would lead to a hierarchical structure of the social network, directly related to the time spent with others. This would justify the intuitive assumption that we have different degrees of friendships, and we would deal differently with each degree (which has been recently reflected in Google+ Circles). The innermost layers of the ego-network structure are what matters for us here, which we will roughly consider as being close relationships[8].

These relationships should tell more about one's characteristics than distant relationships, as it is implied in the concept of homophily [17]. Thus, filtering close relationships from a social network may refine attempts to predict those characteristics by looking at the social graph. Although interaction graphs have already shown interesting results for that matter [22], it is not always possible to access information about all users interactions. Hence, we try to show that looking at the photo tags is an easy and reliable way of filtering social ties. It must be clear that we are not claiming that tagged friends are exactly the user's close friends, but we believe they are closer than an average Facebook "friend". The reasons why we believe that are listed below.

1. Tagged friends normally share or have shared the same physical environment for a reasonable amount of time (at least for the during of taking the picture). Some Facebook friends are distant acquaintances that barely meet the user, so are less likely to share social characteristics. Wilson et al. [22] states physical closeness as an important element for keeping relationships.

2. Taking a photo with someone is slightly more socially relevant than merely knowing this person. We cannot rule out situations in which a person appears in a photo unintentionally, without knowing the owner. However, it is not likely that this person will be tagged in the owner's album.

3. Photo tagging someone shows a higher degree of interaction than simply adding someone. Also, the features in Facebook lead users to interact more as they are linked by a tag, by notifying them of all comments in the photo and displaying the photo in their album.

3.2.1 Simmelian Ties and Clustering Coefficient

The Simmelian tie earns its name to the 19th century sociologist Georg Simmel [23], who thoroughly described the different properties of network ties as they leave being dyads by the addition of a third element. It is nowadays seen as

[8]Note that "closeness" will not be used as a graph-theoretic term, but as emotional closeness.

Table 1: Demographics of the study participants (dataset P). The dataset comprehends participants from 25 countries. Data was not available for some users, these are presented with N/A in the table.

	Total N = 744	**Brazil** N = 382	**India** N = 251	**Others** N = 111
Gender (%)				
Female	26.75	31.41	18.33	28.46
Male	67.88	64.66	71.71	67.69
N / A	5.37	3.93	9.96	3.85
Age (%)				
Under 18	1.20	0.26	2.39	1.54
18 - 25	44.76	37.43	62.95	30.00
26 - 35	33.87	39.00	20.32	46.15
36 - 45	6.32	7.07	3.59	8.46
46 - 55	1.88	2.62	1.59	1.54
Over 55	0.67	0.52	0.80	0.77
N / A	11.30	13.10	8.36	11.54
User degree				
Average	207.17	155.2	252.39	287.56
Median	159	122	202	236
User content				
#Albums	3,653	1,530	1,294	829
#Photos	78,801	28,531	22,514	27,756
#Tags	55,855	7,891	38,811	9,153

ties embedded in other ties [19]. This relation can be directly linked to the concept of clustering coefficient in graph theory. A simple way of calculating the clustering coefficient (CC) of a node is defined as follows. Suppose that a node u has k_u neighbours; then at most $k_u(k_u-1)/2$ edges can exist between them (this occurs when every neighbour of u is connected to every other neighbour of u). The clustering coefficient CC_u of node u is, then, the fraction of these allowable edges that actually exist [21]. Because CC is directly proportional to the number of existing closed triplets it is also proportional to the number of Simmelian ties in the network.

Simmelian ties have interesting properties in a social context: they promote trust between members, reduce individuality and allow for the creation of norms, increase cohesion in a clique, and finally, promote homophily [12, 19, 23]. This leads us to believe that Simmelian ties will be more frequent as one approaches the closer relationships in an individual's network. If this is true, and if photo tags do tackle more accurately the inner layers of social relationships, then we should notice an increase in the CC of the network containing only the user's photo tagged friends as compared to the CC of the complete friend list network.

4. METHODOLOGY

To analyze the impact of photo-tagging on privacy attacks, we developed a Facebook application to collect profiles from various users. To recruit participants to the study, emails and Facebook messages were sent, fliers were affixed on university notice boards, media posts were published and, due to the prizes offered, there was a lot of word-of-mouth advertising. Most of the campaigning about the study was done in Brazil and India. A domain was registered *http : //www.theprivacystudy.org/* to host all information (link to the Facebook application, details about prizes) about the study.

Table 2: Demographics of the study participants and their friends (dataset PF). The dataset comprehends users from 142 countries. 74,623 users didn't reveal their country and are accounted in *Total* column only. Data was not available for some users, they are presented with N/A in the table.

	Total n=119,091	**Brazil** n=12,704	**India** n=17,287	**Others** n=14,477
Gender(%)				
Female	36.32	40.47	22.30	34.59
Male	57.15	51.05	70.10	50.31
N/A	6.53	8.48	7.60	15.10
Age(%)				
Under 18	1.92	0.64	2.37	2.39
18 - 25	29.69	20.65	45.88	20.87
26 - 35	23.75	32.18	11.54	21.39
36 - 45	5.33	7.26	1.21	5.16
46 - 55	2.25	2.62	0.77	1.91
Over 55	1.49	1.57	0.29	1.28
N/A	35.57	35.08	37.94	47.00

A total of 664 participants installed the application in their Facebook account. When they installed the *Privacy Study* application[9], they were presented with the study privacy policy, which explains what kind of data will be collected from them and for what purpose. Participants were then invited to authorize the application access to their Facebook data. The application worked in conformity with the Terms of Service of Facebook and ethical ways of studying social media [5]. We were not required, neither in Brazil nor in India, to go through an IRB-type approval process before collecting the data through the Facebook application. However, authors of this paper have previously been involved in studies with U.S. Institutional Review Board (IRB) approvals, and have applied similar practices in this study.

Our objective was to infer some of the user's missing attributes by looking at their graph neighbours. We developed three algorithms for this type of prediction: i) A naive approach, that would pick the most frequent attribute in the user's friend list; ii) The same naive approach, but applied to the user's friends that have been photo tagged; iii) a weighted approach, using the frequency of photo tags to give more importance to friends that have been photo tagged more often.

We chose four attributes te be predicted: *Gender*, *Current Country*, *Current City* and *Age*. Our choice was motivated by the fact that homophily is divided in two categories: *selection*, *i.e.*, that similarities contribute to the formation of ties, and *influence*, *i.e.*, that close relationships tend to cause the ego to change his ideas and cause similarities. Our study was based on one snapshot of the Facebook profiles, which did not allow us to create timestamps and control for influence effects. Thus, we chose variables that would only be marked by selection.

[9]http://apps.facebook.com/privacystudy

4.1 Datasets

Two sets of data were collected, one of participants who installed the application and one of the friends of the participants. Dataset P represents the users who installed the application and is summarized in Table 1. Both in Brazil and India, male participants were more frequent than female participants, which is opposite to the data that Boyd et al. collected in their study [3]. According to Facebook Ads[10], there are far more male users (72%) in India compared to female users (28%), whereas there are more female users (54%) in Brazil than male (46%). Users between 18 - 25 years old were the most common (44.8%) in our study, which is likewise the most active age group on Facebook. Young adults in this specific age group are the most prevalent users of most Online Social Networks on the Internet[11]. Dataset PF represents the participants along with their friends; some details about this dataset is given in Table 2. It is important to point out that the application has only the permission to collect private information of its users, so the availability of friend information are dependent on their privacy settings.

5. INFERENCE ALGORITHMS

Consider the ego-network E_u of a certain user u. This consists of all his friends $f_1 ... f_n$ on the OSN. We define E_u as the *Friends Ego Network* and it can be obtained from the user's friend list. This is a common network used in privacy attack studies. We also consider a second network $T_u \subset E_u$, which consists of all the linked profiles in photo tags in the photo albums of u. We call this network *Tagged Friends Ego Network* and it can be obtained from the user's albums and photos.

5.1 Friends Ego-Network Lookup (FEL)

For this algorithm we take a simple approach by analyzing the *Friends Ego Network*. Given the attributes of some nodes in network E_u, the probability $P(a = a_i|E_u)$ of the sensitive attribute a being a_i is given by

$$P(a = a_i|E_u) = \frac{N_{a_i}}{N} \quad (1)$$

where N_{a_i} is the number of nodes with attribute $a = a_i$ and N is the total number of nodes in the network. In other words, the sensitive attribute a of query node u is assumed as the most frequent value of this attribute in the network. On the example in Figure 2 (a), u would have the same attribute as friends f_1, f_2 and f_6 (50% of probability). Considering it is the most basic algorithm for privacy attacks on ego networks, we consider it to be our baseline.

5.2 Tagged Friends Ego-Network Lookup (TFEL)

For this approach we use the *Friends Ego Network* as the scope of the privacy attack. The calculation of the probability for the sensitive attribute of the attack is analogous to Equation 1 presented in the previous algorithm, differing only by the network used. On the example depicted on Figure 2 (b), only friends f_2, f_3, f_4 and f_5 are present on the network since they are the only friends tagged in u's photos. This way, u would be assumed to have the same attribute as friends f_3 and f_4 (50.0% of probability).

[10]http://www.facebook.com/ads, as of July 13th , 2011

[11]http://social-media-optimization.com/2008/05/social-network-user-demographics/

5.3 Weighted Tagged Friends Ego-Network Lookup (WTFEL)

In this algorithm, we transform the *Friends Ego Network* into a weighted graph, *i.e.*, each edge in the graph has a label (weight) associated with it. The weight is the number of times a friend is tagged on the user's album. As marked in the edges in Figure 2 (c), friends f_2 and f_3 have been tagged in two pictures, f_4 in one picture and f_5 in four pictures. The weight is now taken in consideration when calculating the probability $P(a = a_i|T_u)$ of sensitive attribute a being equal to a_i:

$$P(a = a_i|T_u) = \frac{\sum w_{f_{a_i}}}{\sum w_{f_n}} \quad (2)$$

where $\sum w_{f_{a_i}}$ is the sum of all weights in edges connecting to friends with attribute $a = a_i$ and $\sum w_{f_n}$ is the sum of all weights in the graph. For the example in Figure 2, the predicted attribute for u would be the same as f_5 (44.4% of probability).

6. RESULTS

To compare how photo tags may facilitate privacy attacks, we trimmed our dataset by selecting only users who had at least one photo tag of a friend revealing the attribute to be predicted. We used the three algorithms and calculated the accuracies in predicting the four attributes. When the probability for two or more attributes are equal, we randomly choose one of them as the predicted attribute. Results are discussed below.

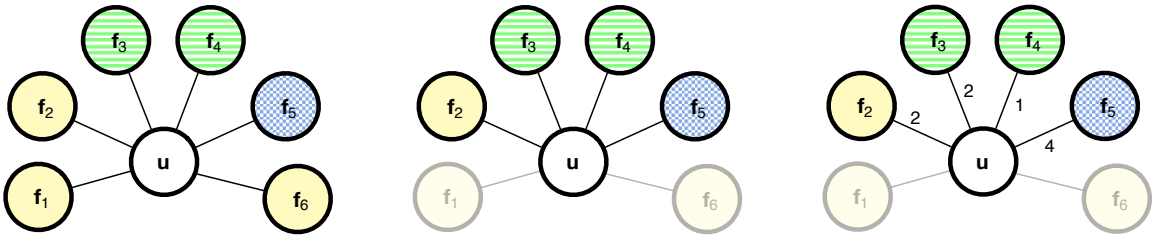
We used two metrics for calculating the accuracy of the algorithms for the age attribute: *Mean Absolute Error* (MAE) and *Cumulative Score* (CS). The MAE is calculated by the mean of the absolute difference between every prediction and the actual value (ground truth). Let a'_k be the predicted age by the algorithm, a_k be the real value of the attribute and n the sample size, MAE=$\sum_{k=1}^{n} |a'_k - a_k|/N$. It is a widely used metric for age prediction algorithms [15]. The CS measures the results in a discrete way (hit or miss) by comparing the absolute error to a predetermined threshold value. Given a T threshold value, $CS(T) = N_{e<T}/N \times 100\%$ where $N_{e<T}$ is the number of cases where the absolute error between the predicted value and the real value is no more than T and N is the sample size. For example, $CS(4)$ considers as a right prediction all predicted values with absolute error smaller than 4 years. This metric has also been used before [7] and could be generalized for any numerical attribute.

6.1 Non-numerical Attributes

We ran a Chi-squared analysis on the prediction of the attributes *City*, *Country* and *Gender*. Results are shown in Table 3. Results were statistically significant ($p<0.05$) with all the algorithms taken together, and also significant when analysing the two algorithms FEL x WTFEL. In the FEL x TFEL analysis, Gender was not significant. No difference was significant in the TFEL x WTFEL analysis.

6.2 Numerical Attributes

Even though the algorithms are defined in a way that it is possible to generalize them to any sensitive attribute



(a) Friends Ego Network (b) Tagged Friends Ego Network (c) Weighted Tagged Friends Ego Network

Figure 2: Examples of networks used in privacy attack algorithms. User u is the subject of the attack, $f_1..f_6$ are friends of u and $f_2..f_5$ have been photo tagged in u's photos. f_2 and f_3 have been tagged twice, f_4 once and f_5 four times. The different colors and patterns represent different values for a hypothetical sensitive attribute.

		FEL	TFEL	WTFEL
Current City	n=111	54.0%	**65.8%**	**77.5%**
Current Country	n=111	80.2%	**88.3%**	**93.7%**
Gender	n=238	73.9%	85.7%	**89.1%**

Table 3: Accuracy of the algorithms as measured by the percentage of right predictions for attributes *CurrentCity* *Current Country* and *Gender*. **Boldface values** are statistically significant (p<0.5) when tested against FEL.

the SNS might provide, numerical attributes can take advantage of the possibility of arithmetic averages. The algorithms in Section 5 assign probabilities to each possible value of sensitive attribute a. For non-numerical attributes, the value with the highest probability is selected. Next we present three ways of computing the result for numerical attributes. Let a be the numerical sensitive attribute of the attack. Suppose the network has two nodes with $a = 10$, one with $a = 20$ and one with $a = 30$. In the way it is presented, the algorithm would assume $a = 10$ to the targeted subject. This assumption is analogous to the mode of the values because it uses the most frequent number on the list, but it is also possible to use the mean ($a = 35$) or the median ($a = 20$). Table 4 presents the comparison of these three variants (mean, median and mode) for age prediction. To measure their performance, we used both MAE and CS with a threshold of 4 years.

As can be seen, the results are almost the same in each algorithm, with the FEL even performing slightly better than the TFEL in some cases. A one-way ANOVA on the MAE evaluation shows no statistically significant difference between the algorithms. A Chi-Squared analysis for the CS(4) also reveals no significant difference. The reason for this will be discussed further in Section 7.

6.3 Clustering Coefficient

To compare the existence of closer social relations in the networks it is necessary to measure its clustering coefficient (Section 3.2.1). The average clustering coefficient found for the friends ego-networks was 0.10 and for the tagged friends ego-network was 0.52. The average CC is lower than values found by previous studies, 0.16 [8] and 0.164 [22]. This might be explained by the fact that the set of friendships available is incomplete due to the concealment of the friend list by some part of the users. When considered only nodes with

	FEL	TFEL	WTFEL
MAE (Years)			
Mean	2.38 ± 0.44	2.28 ± 0.45	2.01 ± 0.44
Median	1.51 ± 0.45	2.01 ± 0.46	1.64 ± 0.44
Mode	1.50 ± 0.47	1.93 ± 0.43	1.58 ± 0.43
CS(4) (%)			
Mean	86.8	83.6	87.9
Median	92.6	85.2	88.9
Mode	92.6	84.7	88.9

Table 4: Accuracy of the algorithms for sensitive attribute age as measured by the two metrics MAE and CS. Neither of them show statistical significance between the algorithms.

full information, the average clustering coefficient is raised to 0.13. Figure 3 depicts an example of photo tagging ego-network where a lot of Simmelian ties can be seen.

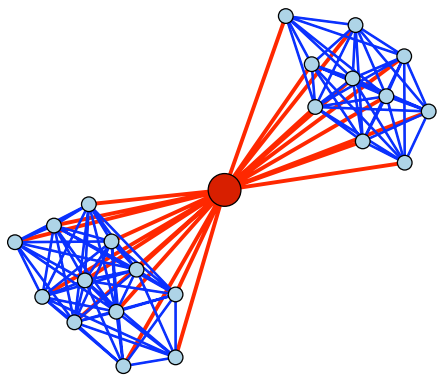


Figure 3: Example of the tagged friends ego-network of a certain user. The user who tags is represented by the larger node at the center, while the friends tagged are shown at the two extremes. A lot of links between tagged friends can be seen. The actual clustering coefficient for this example is 0.42.

7. DISCUSSION

Although our data enables us to make a compared estimation of the network CC, it cannot be treated as the actual CC of the ego-network. Since we could not access the full network of all the users that were friends with the ones who installed the application, there were missing links in both the friend-list and the photo tag network. We did, however, have an equal amount of missing links in both networks, which allowed us to compare them.

Photo tags enhanced prediction of attributes like *Gender*, *Current City* and *Current Country*. We believe that homophily is higher among close friends than among distant ones. However, tags had no effect in inferring age. One possible explanation is that, since a Facebook ego-network does not completely translate one's actual social network and is adopted mostly by a specific age group, age is a biased attribute that escapes social sensitiveness. The most common age group in SNSs are users between 18 and 25 years old (44.8% in our sample). However, overall, all algorithms had a reasonably good accuracy in age prediction.

Recent updates on Facebook features included a face recognition engine that suggests photo tags to any newly uploaded photo[12]. As seen in our data, the majority of users does not use photo tags, but this can change as the new feature starts to get more popular.

We studied only attributes which would not be significantly affected by homophily-based *influence*. The main reason is that we could not control for this effect, since we had only one time stamp of the network. However, we believe that the filtering made by photo tag will be stronger in attributes affected both by selection and influence, such as *Music*, *Books*, *Political Views* and *Favourite Teams*.

Since our goal was to take into account the effects of using photo tags in inferring attributes, we compared fairly simple algorithms. The principle, however, can be combined to data-mining and machine-learning techniques, iterative methods, finer statistical analyses, sample selection, and combination between tags and friendship networks.

Photo tagging is an important interaction in social media. SNSs other than Facebook employ it as a central feature (such as Google+ and Orkut). Linking users to photos promote a higher degree of interaction and allow users to know when there is information about themselves that they did not put on the Internet. However, tags also convey information that can be used by third parties to breach one's privacy. More specifically, we have found that they can be used to predict information about the user's *Gender*, *Current City* and *Current Country*. It is important for users to be aware of this possibility, so they can make informed decisions when exposing information and controlling their privacy settings. Also, SNSs might benefit from this kind of information by incorporating features that avoid user's unintended loss of privacy. For example, there could be a "hiding" feature for photo tags. Users would be able to hide their tags instead of deleting it. Thus, they would still keep track of the photos they have online and would still keep a high degree of interaction with the album owner (e.g., by receiving updates on comments on the photo), but without directly linking the photo to their profiles.

[12]http://www.facebook.com/blog.php?post=467145887130

Acknowledgments

The Indo-Brazil Research Council supported this work. The authors would like to thank Prof. Ponnurangam Kumaraguru and his students at PreCog at IIIT-Delhi, India for their support in collecting data and hosting two of the student authors at IIIT-Delhi in India. The authors also thank all the users who participated in the study and helped to promote it.

8. REFERENCES

[1] R. G. Alessandro Acquisti and F. Stutzman. Faces of facebook: Privacy in the age of augmented reality. *BlackHat USA*, 2011.

[2] A. Besmer and H. R. Lipford. Moving beyond untagging: photo privacy in a tagged world. *Conference on Human Factors in Computing Systems*, pages 1563–1572, 2010.

[3] D. Boyd and E. Hargittai. Facebook privacy settings: Who cares? *Journal on the Internet*, 15(8), 2010.

[4] R. J. Brym and B. Wellman. Social structures: A network approach. Cambridge University Press, 1988.

[5] D. Fisher, D. W. McDonald, A. L. Brooks, and E. F. Churchill. Terms of service, ethics, and bias: Tapping the social web for cscw research. *Computer Supported Cooperative Work (CSCW), Panel discussion*, 2010.

[6] G. Friedland and R. Sommer. Cybercasing the joint: On the privacy implications of geo-tagging, 2010.

[7] X. Geng, Z. H. Zhou, Y. Zhang, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. In K. Nahrstedt, M. Turk, Y. Rui, W. Klas, and K. M. Patel, editors, *ACM Multimedia*, pages 307–316. ACM, 2006.

[8] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. A Walk in Facebook: Uniform Sampling of Users in Online Social Networks. Technical report, arXiv.org, 2009.

[9] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80, 2005.

[10] J. He, W. W. Chu, and Z. V. Liu. Inferring privacy information from social networks. In *IEEE International Conference on Intelligence and Security Informatics*, 2006.

[11] D. Kirkpatrick. *The Facebook Effect: The Inside Story of the Company That Is Connecting the World*. Simon and Schuster, 2010.

[12] D. Krackhardt. *Simmelian Ties - Super Strong & Sticky*, chapter 2, pages 21–37. Sage, Thousand Oaks, CA, 1998.

[13] B. Krishnamurthy. I know what you will do next summer. *ACM SIGCOMM Computer Communication Review*, 40(5):65–70, Oct. 2010.

[14] B. Krishnamurthy and C. Wills. Characterizing privacy in online social networks. *Proceedings of the first workshop on Online social networks*, (37–42), 2008.

[15] A. Lanitis, C. Taylor, and T. Cootes. Toward automatic simulation of aging effects on face images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4):442 –455, apr 2002.

[16] J. Lindamood and M. Kantarcioglu. Inferring private information using social network data. Technical report, 2008.

[17] M. McPherson, L. Smith-Lovin, and J. M. Cook. BIRDS OF A FEATHER : Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, Aug. 2001.

[18] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: Inferring user profiles in online social networks.

[19] W. D. Nooy. Social Network Analysis , Graph Theoretical Approaches to, 2009.

[20] S. G. Roberts, R. I. Dunbar, T. V. Pollet, and T. Kuppens. Exploring variation in active network size: Constraints and ego characteristics. *Social Networks*, 31(2):138–146, May 2009.

[21] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[22] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the fourth ACM european conference on Computer systems - EuroSys '09*, page 205, New York, New York, USA, 2009. ACM Press.

[23] K. H. Wolff. *The Sociology of Georg Simmel*. The Free Press, Glencoa, 1950.

[24] E. Zheleva and L. Getoor. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In *18th International World Wide Web Conference*, pages 531–531, April 2009.