Evaluating the Privacy Risk of User Shared Images

Ming Cheung, HKUST-NIE Social Media Lab
James She, HKUST-NIE Social Media Lab

User shared images are shared on social media about a user's life and interests that are widely accessible to others due to their sharing nature. Unlike for online profiles and social graphs, most users are unaware of the privacy risks relating to shared images, as they do not directly disclose characteristics such as gender and origin. Recently, however, user shared images have been proven to be an accessible alternative to social graphs for online friendship recommendation and gender identification. This paper evaluates 1.6M user shared images from an image-oriented social network, Fotolog, and concludes how they can create privacy risks by proposing a system for de-anonymization, as well as inferring information on online profiles with the user shared images. It is concluded that given user shared images, using social graphs is 2 and 2.5 times more effective in de-anonymization than using origins or genders. With two showcases, it is also proven that using user-shared images is effective in online friendship recommendation, gender identification, and origin inference. To the best of our knowledge, this is the first paper to evaluate the privacy issue qualitatively with big multimedia data from a real social network.

CCS Concepts: •**Security and privacy** → **Social aspects of security and privacy;** •**Networks** → **Online social networks;** •**Human-centered computing** → *Social networking sites;* Social networks;

General Terms: Big data analytic system, Images

Additional Key Words and Phrases: big data, privacy, de-anonymization, user shared images, social network analysis

1. INTRODUCTION

A huge amount of content is generated daily from our mobile devices as they have become an essential part of our lives. Advances in devices such as smartphones and wearables, as well as wireless and cloud technologies, make taking, sharing and analyzing high-quality images much easier than before. User shared images are images shared on social media that relate a user's life and interests. They are widely accessible to friends, and on popular mobile social applications such as Instagram. User shared images are available to everyone, even those who are unknown to the user. In contrast to images, users often hide or limit their online profiles and social graphs (SGs) from the public on social media platforms due to privacy concerns. That information is only available to the partnered companies of these platforms, usually in an anonymized form, in which the identity of users is removed. Applications and social networks, such as Instagram (owned by Facebook from the US) and WeChat (owned by Tencent from China), keep the online profile and SGs of the user privately. This is the trend of today's social networks for preserving user privacy, and users believe that they can enjoy image sharing without risking their privacy. However, most users are unaware that there are similar privacy risks relating to shared images, as such images do not directly disclose their online profile characteristics, such as gender and origin.

Recently, using user shared images has been proven to be a more accessible alternative to SGs for follower/followee relationship recommendation and gender identifi-

This work is supported by the HKUST-NIE Social Media Lab.
Author's addresses: M. Cheung and J. She, Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

DOI: 0000001.0000001

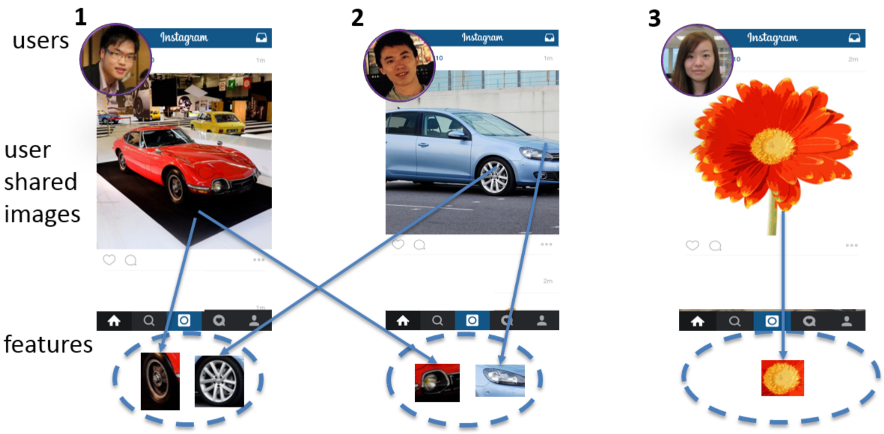

Fig. 1. Examples of the user shared image and their features

cation [Cheung et al. 2015a], even without the access to the SGs. It is confirmed in [You et al. 2015] the interests of a user can be inferred from their shared images. Researches have also confirmed that gender can be identified from image similarity, the similarity of the visual features in their shared images [You et al. 2014], as users of the same gender are more likely to share similar images. An example of user generated images on Instagram is shown in Fig. 1. Both users 1 and 2 have shared images of cars and user 3 has shared an image of a flower. As the features of cars are similar, the similarity between users 1 and 2 is higher than that between users 1 and 3 or 2 and 3. Users 1 and 2 are therefore more likely to be of the same gender as they have a higher image similarity in their shared images. As the user shared images provide a way for identifying users, the sharing nature of such images raises a privacy concern about sharing images on social media.

Social media operators often share potentially sensitive information about users and their relationships, including information about users' genders, interests, origins and SGs, with other organisations. The information is anonymized, such as by replacing names and interests with a number or in such a way that the online profile of a particular user cannot be identified from the data. In this way, the privacy of users is protected. However, there is an incentive for these organisations to try to de-anonymize the data for various purposes, for example, for academic and government data-mining to locate a group of specified users, for advertisers and applications to obtain more information for marketing and sales promotion, and for personal uses, such as background checks and requests to obtain information about a particular user [Narayanan and Shmatikov 2009]. In de-anonymizing data, these organisations become attackers. Most attackers have previously focused on de-anonymizing user identity based on the structure of SGs. Since access to SGs is getting harder, there is a motivation to investigate other means of de-anonymizing user identity, including using user-shared images. As shared images can reflect user characteristics such as gender and follower/followee relationships, applications such as recommendation and marketing may be possible, even without access to an SG. It is therefore interesting to investigate whether the widely accessible user-shared images can be used to invade the privacy of users by allowing anonymized information, such as SGs and online profiles, to be de-anonymized, or for online profile information to be inferred, even without any anonymized information.

With this motivation, this work investigates 1,598,769 user shared images from

6,036 users from an image-oriented social network, Fotolog, using a novel image processing technique, bag-of-features tagging (BoFT) [Cheung et al. 2015a]. BoFT is a bag-of-features-based technique that annotates images with non-user generated labels, so that the connections among users can be discovered. Intensive measurements show an interesting phenomenon between user profile information and their shared images: two users with a higher similarity in their shared images are likely to have similar online profile information and have an online friendship. This phenomenon is nicely formulated with a proposed analytic system to de-anonymize user identity from user shared images on social media. To the best of our knowledge, this is the first paper to evaluate how user shared images can disclose a user's private information, and to propose a system to de-anonymize user identity by matching shared images, profile information and friendships. In summary, the contributions of this paper include the following:

- measured intensively and characterized user shared images from an image-oriented social network, Fotolog, proving the phenomenon that two users with a higher similarity between their shared images are likely to have similar online profile information and have an online friendship between them;
- proposed and verified extensively a formulation and an analytic system for de-anonymization with over 1.6 million images from 6,036 users from Fotolog to prove the effectiveness of using user shared images for de-anonymization through bag-of-features tagging;
- concluded that social graphs are the most sensitive information in terms of privacy protection, as compared to origin and user gender for sharing images on social media;
- demonstrated 2 showcases to understand the effectiveness of using user shared images for origin inference and gender identification.

This paper is organized as follows: Section 2 discusses the related works. Section 3 introduces the image-based method for connection discovery, while Section 4 shows the measurements of user shared images on the datasets. Section 5 proposes and formulates the de-anonymization analytic system, Section 6 is the experimental results of the system, and Section 7 concludes the paper.

2. RELATED WORKS

User online profiles contain names, user-associated locations, age, e-mail and habits [Campisi et al. 2009], as well as SGs and user shared content, which is all valuable information for many different applications. It has been shown that online friendship and follower/followee recommendations can be made from SGs [Jin et al. 2001][L and Zhou 2011], user generated content [Cheung and She 2014][Parimi and Caragea 2011] and other personal information [Xie 2010][Gilbert and Karahalios 2009][Golder and Yardi 2010][Yang et al. 2011]. Other recommendations and services are also possible using such information [Guy et al. 2009][Hsu et al. 2006] and location information [Yu et al. 2016][Guo et al. 2015]. These services make social media users' lives more convenient than ever, as the information they want and people they may know are always available, without the needs for manual searching and efforts. However, people are often concerned about their privacy on social media. On one hand, they want to share more information to enjoy the convenience, but on the other hand, they are afraid that their shared content and information will be used in an unintended way. Such information can be stored to analyze a user's changes in behaviors over time, and where users do not have any control over it [Campisi et al. 2009]. This data, even if not personal, could also be used in the future to identify an individual.

Data is regularly provided from social media to other parties, such as advertisers, application developers, and researchers [Narayanan and Shmatikov 2009]. However, the data is anonymized [Zhang et al. 2014], so that the user's sensitive information, such as user IDs, is represented by a string. For example, the SGs are anonymized in this way so that they cannot be mapped to the original user ID to obtain more information about a user. However, there is a risk that recipients of this data will become attackers and try to de-anonymize it [Narayanan and Shmatikov 2009] by matching the anonymized user data provided by the social media operators to the actual online profiles of users to provide more information about them. Those parties who try to de-anonymize are called attackers, and this has long been researched [Thomas et al. 2010][Calandrino et al. 2011] based on user online profiles such as from email address [Balduzzi et al. 2010] and other information [Wondracek et al. 2010][Zheleva and Getoor 2009]. Based on information on social media [Jain et al. 2013][Korayem and Crandall 2013], it is possible to match profiles on 2 social media platforms to gain access to more personal information. It is demonstrated in [Kleinberg 2007] that an attacker can conduct a passive attack, in which fake users are created and connected to a social graph before anonymization by a social media operator [Narayanan et al. 2011]. By locating the fake nodes as seeds, it is able to use a propagation method to de-anonymize all users. To protect privacy, k-anonymity [LeFevre et al. 2005], in which each anonymized user shares exactly the same properties with other $k-1$ users, has been proposed to measure with how much difficulty an anonymized social graph can be de-anonymized. By adding some connections while deleting others, k-anonymity can be achieved and the difficulties of de-anonymization are increased [Song et al. 2011].

Protecting privacy is especially important for shared content such as user shared images, which are widely accessible due to their sharing nature [Ahern et al. 2007]. [Campisi et al. 2009] were able to identify co-occurring faces in shared images, and also, retrieve objects and tags associated with the shared images. This raises a concern about how user shared images can be used to invade a user's privacy. For example, if the faces of users are known, their relationships may be able to be discovered based on the co-occurrence of their faces on the same image. An emerging image-based approach [Zhang et al. 2012][Moxley et al. 2009] applies computer vision techniques to produce non-user generated labels that reflect the context of images. One of the techniques for tagging images with non-user generated labels is bag-of-features tagging (BoFT), which has recently been proven to be an alternative to social graphs for connection discovery by [Cheung et al. 2015a]. This is an unsupervised method, in which the objects in images are not required to be recognized, hence, there is no need for well-tagged images for training. Two users who are follower/followee have a higher image-based similarity between them. De-anonymization can be conducted by matching their shared images with SGs, by using the fact that friends have a higher similarity in terms of shared images. As previous research has provided no model of the characteristics of user shared images for de-anonymization, this paper uses BoFT to annotate user shared images with non-user generated labels, BoFT labels, for de-anonymization. This paper goes beyond the previous works in the following ways: 1) evaluated in depth a massive dataset from Fotolog, in which there are 1.6 million user shared images from 6,036 users, which has never been reported before; 2) measured intensive and characterized user shared images on the dataset to prove the phenomenon that two users with a higher similarity between their shared images are likely to have similar online profile information; and 3) formulated de-anonymization and concluded that SGs are the most sensitive information in terms of privacy protection, as compared to origin and user gender, for sharing images on social media.

3. BOF-BASED TAGGING FOR CONNECTION DISCOVERY

This section introduces the proposed method, BoFT, which labels images with non-user generated labels, BoFT labels, and present how BoFT similarity, the pairwise similarity among users based on BoFT labels, is calculated at the discovered connections.

3.1. BoF-Based Tagging

The images are analyzed using the method, BoFT, in which each image is annotated with a BoFT label. BoF is a popular computer vision approach for analyzing images [Vedaldi and Fulkerson 2010]. Fig. 2 shows the key steps involved: Fig. 2 (a) shows the steps for BoF, and Fig. 2 (b) shows the method for similarity calculation based on user shared images. The different steps of BoFT are introduced below.

3.1.1. Feature Extraction. Feature extraction is a process to obtain the unique local features, as in step 1 of Fig. 2 (a). These unique features can be detected by feature detection, such as the Harris Affine detector, Maximally Stable Extremal Regions detector [Vedaldi and Fulkerson 2010] and KadirBrady saliency detector [Kadir and Brady 2001]. The extracted features are relatively consistent across images taken under different viewing angles and lighting conditions. In this work, the images representation is independent of the size and orientation by scale-invariant feature transform (SIFT) [Lowe 2004].

3.1.2. Codebook Generation. Codebook generation, in step 2 of Fig. 2 (a), is a clustering process to obtain a set of visual words, a representative and distinct set of unique visual features. This step starts with clustering extracted visual features into groups by clustering techniques, such as K-means clustering, based on their visual similarity, and the mean vectors of each group are defined as a visual word. Other possible techniques are the Canopy clustering algorithm [McCallum et al. 2000] and LindeBuzoGray algorithm [Linde et al. 1980]. A K-means clustering is used in this work.

3.1.3. Feature Coding and Pooling. Feature coding represents each visual feature by the closest visual word. Each image is represented by a feature vector in the feature pooling, as shown in step 3 of Fig. 2 (a). One of the most common approaches is counting the number of occurrences of each unique visual word on an image as the feature vector.

3.1.4. Clustering and BoFT Labeling. Clustering groups images that are visually similar through the similarity in their feature vectors, as shown in step 4 of Fig. 2 (a). For example, when two images contain cars in the countryside, the feature vectors of the two images are similar in terms of the number of occurrences of each unique visual word. As a result, the two images will be assigned the same BoFT label to indicate that they are visually similar. BoFT applies one of the most popular clustering algorithms, K-means, which will first randomly generate K cluster centroids. It then iteratively assigns points to their nearest centroids, followed by a recomputing of the centroids until it converges. However, K-means does have its drawbacks in that the points lying far from any of the centers can significantly distort the position of the centroids and the number of centers must be known in advance. The next step, BoFT labeling, assigns each cluster a BoFT label so that those images with the same BoFT label are visually similar, and this is shown in step 5 of Fig. 2 (a). The set of BoFT labels of user shared images of user i, L_i, is obtained. L_i is a vector, with each element being the set of occurrences of a BoFT label in the shared images of user i. The step is an unsupervised operation that analyzes user shared images without any manual inputs or processes.

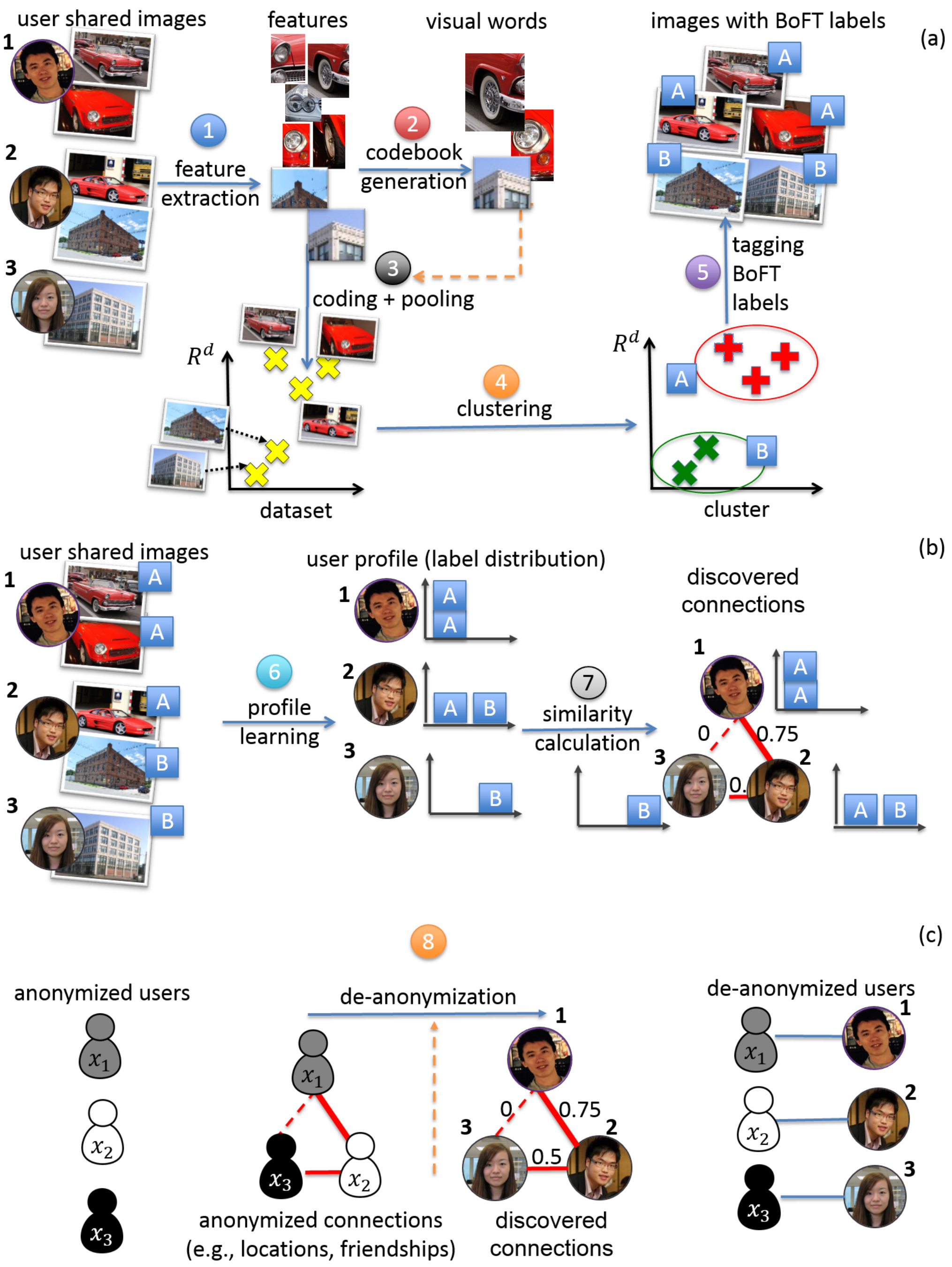

Fig. 2. BoFT for de-anonymization: (a) annotation with BoFT labels, (b) user similarity calculation based on BoFT labels, (c) de-anonymization by known profiles and the discovered connections.

3.2. Similarity Calculation with BoF Labels

This section introduces how similarity between two users can be calculated through BoF labels.

3.2.1. BoFT Labels and User Profile. The distribution of BoFT labels, which reflects the content of user shared images, is the key in similarity calculation. The proposed method uses the number of occurrences of the BoFT labels, as in step 5 of Fig. 2 (a), of the shared images of a user as his/her user profile, as in step 6 of Fig. 2 (b). A user i is represented by his/her user profile, L_i, and the distribution of the BoFT labels that the user has is defined as:

$$L_i = \{l_1, ...l_k, ...l_K\} \tag{1}$$

where l_k is the number of occurrences of the k-th label among the shared images of user i, and K is the total number of labels, which is set to 500.

3.2.2. User Profile and User BoFT Similarity. When the user profile of each user is established, the next step is to identify user genders based on the BoFT similarity, $S_{i,j}$, of users i and j, in which users who share highly similar images will have a high BoFT similarity. This requires a pairwise similarity comparison among user profiles based on the number of occurrences of BoFT labels, and this is calculated using the following formula:

$$S_{i,j} = S(L_i, L_j) = \frac{L_i \cdot L_j}{||L_i|| \cdot ||L_j||} \tag{2}$$

where L_i and L_j are the sets of BoFT labels of the shared image in the user profiles of users i and j, respectively. The similarity, $S_{i,j}$, among all anonymized users is then calculated and the connections among them are discovered, as in step 7 of Fig. 2 (b).

3.3. De-anonymization using Discovered Connections

De-anonymization, in this paper, is the process of matching a set of publicly and widely accessible user shared images to an anonymized dataset provided by a social media operator. An example is Instagram, where SGs and other pieces of online profile information are not publicly available but user-shared images are. An anonymized dataset could include SG, and online profile information such as origin and gender of users. Based on the anonymized online profiles and SGs, connections can be found among the unknown users. The de-anonymization can be based on the similarity reflected by user characteristics, as users with similar characteristics have a higher $S_{i,j}$, and one of the examples of this is online friendships. An example is shown in Fig. 2 (c). This is an optimization problem, in which the matching maximizes users with similar characteristics.

However, even if the set of user shared images cannot be matched to a user profile, user privacy can still be invaded if some of the personal information is inferred or identified from user shared images [Narayanan and Shmatikov 2009][Zheleva and Getoor 2009]. This paper also investigates how privacy can be invaded, even without the needs to use anonymized data, such as SGs, provided by the social network. For example, the names of some users can be used to identify their gender [Liu and Ruths 2013], and information can also be found from accounts linked to other social networks [Jain et al. 2013][Korayem and Crandall 2013]. With this information, it is possible to infer the gender and other online profile information of other users, as 2 users with similar private profile information have a high $S_{i,j}$. This work will show through experiments how gender identification and origin inference are also possible using shared images, and the details are discussed in the coming sections.

4. USER SHARED IMAGES AND SIMILARITY DISTRIBUTIONS

This section first describes the dataset, followed by an analysis of the BoFT similarity distribution by BoFT. Lastly, it introduces how de-anonymization is possible based on the distribution.

4.1. The Datasets

Fotolog is an image-oriented social network, originating in the West, that has developed into a global social network, with users from many different parts of the world. As an image-oriented social network, it allows users to share images with others, and images are the only or the primary form of sharing. Fig. 3 shows the user interfaces of Fotolog. As shown in Fig 3 (a) and (b), users can share images via the user interface of the web page and the mobile application. Users can also decide if they want to share information, such as gender, and origin (circles in broken lines), on their online profiles, as shown in Fig 3. Unlike on social networks such as Twitter and Flickr, users of Fotolog form online friendships not follower/followee relationships. Similar to Facebook, a user first accepts a request from another user before they can form an online friendship.

The experiments in this paper involve 1,598,769 user shared images by 6,036 users from Fotolog, which were collected by Ruby-based scrapers during mid-2015. All the users were selected randomly from a large set of users collected from online friendships, in which 4342 users provided their gender. Among the 4342 users who provided their gender, there are 1810 males and 2532 females. 5805 users also provided their origins, the country that a user has specified in their online profile that they originate from, and there are 153 origins among the 5805 users. Finally, there are 11,432 existing online friendships among the 6,036 users, and there are in total 496,931 online friendships listed in their profiles. Note that the 496,931 online friendships include users who are and who are not among the 6,036 users. Those online friendships are considered to be the SGs of users. Measurements of the dataset can be found in the appendix.

4.2. BoFT Similarity Distribution

A pair of users is 2 users who are connected if they share some similarity. For example, a pair of users can be considered to be connected if they share the same gender. By

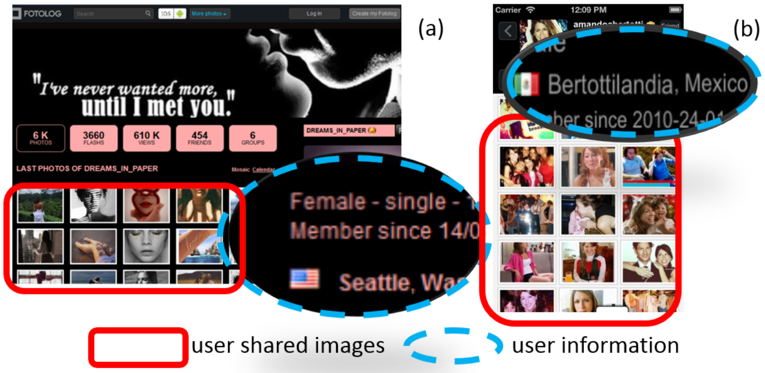

Fig. 3. The user interface of Fotolog: (a) web page, (b) mobile application.

considering the origin, gender and online friendships, 3 types of connections can be found in the datasets. For each type of connection, the connection between a pair of users can be considered as of two classes: connected and unconnected pairs. Note that a user pair can be connected through one of the types, but may be either connect/unconnected through the other types. For example, for user gender, there are 2 classes of connection. The first class of connection is user pairs with the same gender, that is, the two users are both males, or are both females. The second class of user pairs is made up of pairs of different genders, that is, one of them is male and the other is female. Considering user genders, a user pair can be classified into 2 types, C_g, defined as:

$$C_g = \begin{cases} 1 & \text{if two users are of the same gender} \\ 0 & \text{if otherwise,} \end{cases} \tag{3}$$

where $C_g = 1$ is the class of connected pairs, in which the two users of the pair are of the same gender, and $C_g = 0$ is the class of unconnected pairs, in which the two users of the pair are of different genders. Similarly, considering SGs, there are two classes of user pairs, C_f, defined as:

$$C_f = \begin{cases} 1 & \text{if two users are related, i.e., have an online friendship} \\ 0 & \text{if otherwise,} \end{cases} \tag{4}$$

where $C_f = 1$ is the class of connected pairs in which there is an online friendship between a user pair, and $C_f = 0$ is the class of unconnected pairs, in which a user pair does not have an online friendship. Finally, considering the user origin, C_o can be defined as:

$$C_o = \begin{cases} 1 & \text{if two users are from the same origin} \\ 0 & \text{if otherwise,} \end{cases} \tag{5}$$

where $C_o = 1$ is the class of connected user pairs, in which the two users of the pair have the same origin, and $C_o = 0$ is the class of unconnected user pairs, in which the two users of the pair have different origins.

Based on the definition of connected and unconnected pairs, it is interesting to measure and compare their average $S_{i,j}$. The result is shown in Fig. 4. It is observed that connected users have a higher $S_{i,j}$, regardless of the type of connection. Among these connections, the differences between connected and unconnected users based on SGs is highest. It is 53% higher for connected than for unconnected users. The value of $S_{i,j}$ for SGs is also 6.1% and 9.1% higher for connected users based on origins and gender, respectively. The distribution of $S_{i,j}$ can be found in the appendix.

The probability that a pair of users are connected and are of the same gender for a given $S_{i,j}$, $P(C_g = 1|S_{i,j})$, can be calculated as:

$$P(C_g = 1|S_{i,j}) = \frac{n(S_{i,j}, C_g = 1)}{n(S_{i,j}, C_g)} \tag{6}$$

where $n(S_{i,j}, C_g = 1)$ and $n(S_{i,j}, C_g)$ are the numbers of pairs of the same gender and the total number of pairs who provide their genders, given a similarity $S_{i,j}$, as obtained in Fig. 14. Similarly, the probability that a pair of users are connected by SGs, i.e., the 2 users have an online friendship, and have the same origin, for a given

$S_{i,j}$, $P(C_f = 1|S_{i,j})$ and $P(C_o = 1|S_{i,j})$, can be calculated as:

$$P(C_f = 1|S_{i,j}) = \frac{n(S_{i,j}, C_f = 1)}{n(S_{i,j}, C_f)} \tag{7}$$

$$P(C_o = 1|S_{i,j}) = \frac{n(S_{i,j}, C_o = 1)}{n(S_{i,j}, C_o)} \tag{8}$$

where $n(S_{i,j}, C_f)$ and $n(S_{i,j}, C_o)$ are the numbers of possible pairs who provide their SGs, and origin, given a similarity $S_{i,j}$, respectively. As there are different numbers of users who provide their gender, SGs, and origin, $n(S_{i,j}, C_g)$, $n(S_{i,j}, C_f)$ and $n(S_{i,j}, C_o)$ are not necessarily the same for a given $S_{i,j}$. However, if all users provide their SGs, genders and origins, the three values are the same. Based on the above definition, Fig. 5 (a), (b) and (c) show the measurements of $P(C_f = 1|S_{i,j})$, $P(C_g = 1|S_{i,j})$ and $P(C_o = 1|S_{i,j})$ of the social network, respectively. It is observed that when a user pair has a low $S_{i,j}$, they are less likely to have an online friendship, be of the same gender, and be from the same origin than a user pair with a high $S_{i,j}$. Note that in Fig. 5 (b), $P(C_g = 1|S_{i,j})$ is about 0.5 when $S_{i,j}$ is below 0.4. This is because there are only 2 classes, so $P(C_g = 1) = 0.5$, even 2 users are selected randomly. Motivated by the above observations, an analytic system that utilizes these observations for de-anonymization and obtaining user online profile information is proposed, and the details are discussed in the next section.

5. DE-ANONYMIZATION

This section introduces a formulation to conduct de-anonymization based on the observations in the previous section. The goal is to use easily accessible shared images to de-anonymize data. This is a 4-stage (stages A to D) system, as shown in Fig. 6. The first stage is image collection, followed by similarity calculation using BoFT. The third stage is profile collection, and then the use of the profile information to de-anonymize user identity. The last two parts of this section discuss how to implement the annotation of 1.6 million user shared images with non-user generated labels, BoFT labels, and the system design information to analyze image big data.

5.1. Image Collecting

The proposed analytic system carries out image collection, as shown in step A of Fig. 6, which shows the process of collecting user generated images from social media applications such as Fotolog. The images can be provided by the social media operators and mobile applications themselves, or scraped from them, but do not include private information such as the online profiles of the users. The images can be shared in various

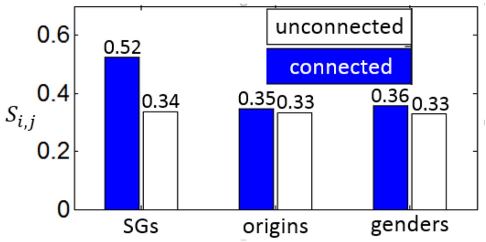

Fig. 4. Similarity measurement of connected and unconnected users in terms of SGs, origins and genders.

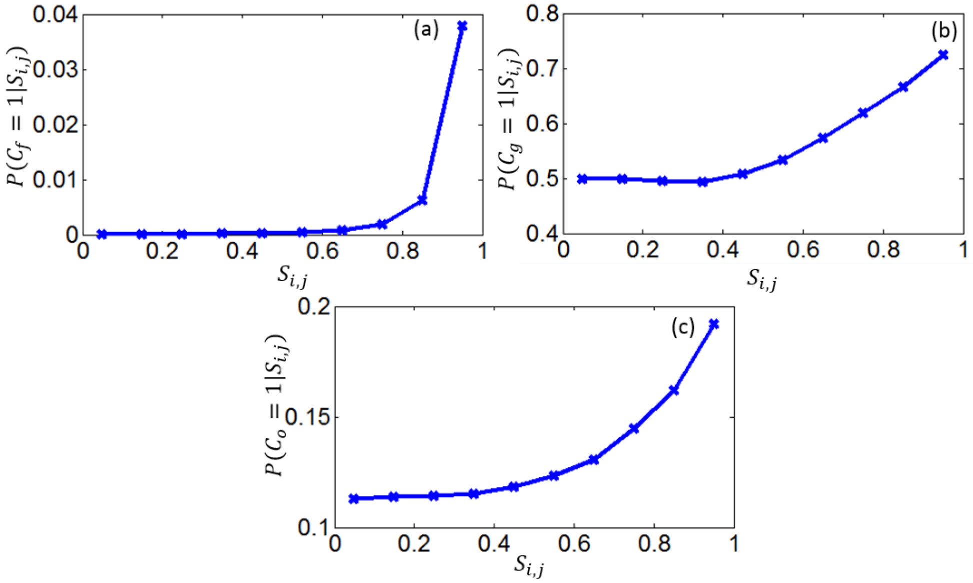

Fig. 5. Distribution of probability for a given similarity: (a) online friendships, (b) gender, (c) origins.

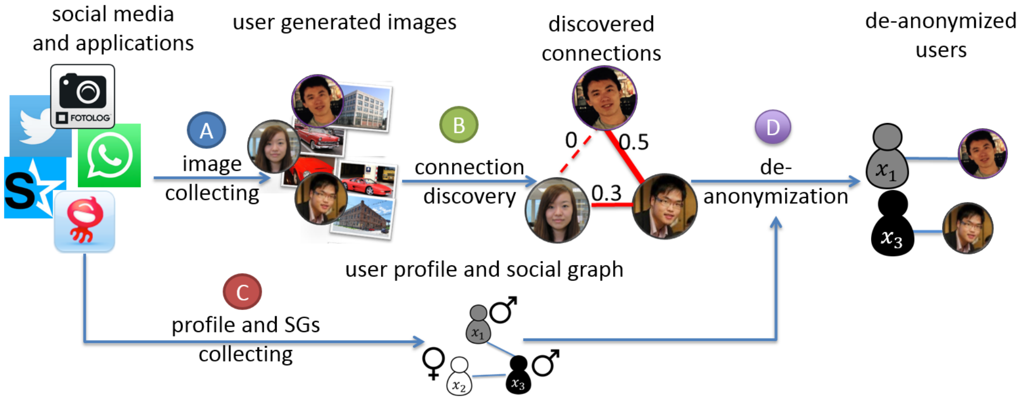

Fig. 6. System flow of the proposed analytic system, (a) collecting images from social media; (b) similarity calculation by collected images; (c) profile scraping from social media; (d) de-anonymization.

forms, such as images posted on social media or images shared through instant messaging applications. On Fotolog, the images are shared by users. In this process, a user who has shared a set of user shared images is represented by user i, and the process is ongoing, which means that the user shared images are collected continuously.

5.2. Similarity Calculation using BoFT

To understand the user generated images, non-user generated labels are generated based on BoFT, as shown in step B of Fig. 6. The accuracy of user annotated tags is unreliable, sometimes even unavailable, so the accuracy of similarity calculation is affected [Cheung and She 2014]. Therefore the proposed system applies a computer vision approach to give a label to user shared images, that is not based on language, culture or other characteristics of the user who shares the image, but is based on the visual appearance of the images only. BoFT is applied to annotate user generated images with non-user generated labels, BoFT labels. The set of user shared images of user x_i is processed by the proposed method, and a set of BoFT labels, L_i, is generated to represent user i. With L_i, the $S_{i,j}$ among users, can be calculated by Eq. 2 and the connections can be discovered.

5.3. Profile and SGs Collection

As introduced in the previous section, the images collected do not contain any personal information about the users who share them. The SGs and user online profile information are assumed to be provided by social network operators after anonymization. The users are represented by x_i. The online profile information considered in this study is user gender and origins. The profile and social graph collection process is an on-going operation, in which the available profile information grows with time.

5.4. Formulation of De-anonymization

De-anonymization is conducted to obtain an N by N assignment matrix A to match anonymized users to the sets of easily accessible images on social media. Given a user i in U_{known}, for which the image-based similarity, $S_{i,j}$, is known for the users within the set, the user is matched to another set of users, x_i in U_x, in which user online profiles or SGs among the users in the set are known but anonymized. The element at position i and x_i of matrix A, $A(i, x_i)$, is defined as:

$$A(i,x_i) = \begin{cases} 1 & \text{if user } i \subset U_{known} \text{ is matched to user } x_i \subset U_x \\ 0 & \text{otherwise} \end{cases}$$
$$\text{where} \sum_{i=1}^{N} A(i,x_i) = 1 \text{ and } \sum_{x_i=1}^{N} A(i,x_i) = 1 \tag{9}$$

The two conditions, $\sum_{i=1}^{N} A(i,x_i) = 1$ and $\sum_{x=1}^{N} A(i,x_i) = 1$, guarantee that it is a one-to-one mapping. Based on the assignment, $C_{i,j}$, the connection between 2 users, i and j, who are assigned to anonymized users x'_i and x'_j, respectively, can be obtained as:

$$C(i,j) = \begin{cases} 1 & \text{if user } x'_i \text{ and } x'_j \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

$C(i,j)$ indicates whether the 2 known users, i and j, matched to x'_i and x'_j, are connected, such as by being of the same gender. $C(i,j)$ is 0 otherwise. De-anonymization then becomes a discrete optimization problem:

$$A^*_x = \max_A \sum_{i=1}^{N} \sum_{j=1}^{N} C(i,j) S_{i,j} \tag{11}$$

However, the complexity of this optimization problem is $O(N!)$, which cannot be solved by a full search. The next subsection introduces a greedy approach to solve the optimization problem.

5.5. Computation for De-anonymization

This section discusses how to solve the discrete optimization problem introduced in the previous section, and proposes a system to process the data for de-anonymization.

As mentioned in the previous section, there are $N!$ possible matches. In order to match the users, a rejection sampling is used to estimate the gain that an unknown user x_i is user i, G_{i,x_i}. In each iteration, a random matching for a user is first generated, and the average $S_{i,j}$ is calculated based on the connections, $C(i,j)$, from the matched users, x_i' and x_j'. If users x_i' and x_j' are connected, e.g., they are of the same gender, then $S_{i,j}$ is for connected user pairs. The average $S_{i,j}$ for connected user pairs and unconnected user pairs is calculated accordingly. If the average $S_{i,j}$ of unconnected user pairs is greater than that of connected pairs, this randomly generated sample will be rejected, as it contradicts the measurements in the previous section. This process is repeated until enough samples are gathered, and the gain that an unknown user x_i is assigned to user i, is estimated by:

$$G_{i,x_i} = \quad \frac{1}{N_s}\sum_{s=1}^{N_s}\sum_{j=1}^{N^{(u)}} C_s(i,j)S_{i,j} \tag{12}$$

where N_s is the number of samples taken, and $C_s(i,j)$ is the connection of i and j for the sample s. Based on the estimated gain, a greedy algorithm is proposed, starting from the highest G_{i,x_i}. User i is then matched to x_i. For those users that are unmatched, the highest G_{i,x_i} is located and the process is repeated until all users are matched. The result is evaluated with the ground truth, the actual correspondence of i and x_i. Fig. 7 shows the algorithm for de-anonymization. Given the $S_{i,j}$ of two users and the connections of user i and user j, $C_{i,j}$, G_{i,x_i} for matching each user to all known users is estimated, as in step 1 in Fig. 7. This process is repeated until 100 samples are generated. The maximum G_{i,x_i} is then located and is considered as a matched user, as in step 2 in Fig. 7. The matched user is then removed, and the process is repeated until there are no more unmatched users. The experiment is then repeated 50 times and the average accuracy is recorded. This greedy algorithm is summarized in Alg. 1.

5.6. Analyzing Image Big Data

This section describes how the 1.6 million images collected from Fotolog are analyzed. The processes of BoFT, such as extracting features from user shared images, become a challenge with millions of images. In the experiment, a Matlab-based analytic system is implemented and deployed on Amazon. It contains a master and a number of slaves; the master assigns images to the slaves and the feature vectors of the images are returned to the master for clustering to annotate each image with a non-user generated label. In this work, slaves are run on an Amazon EC2, with Ubuntu Server 14.04 LTS

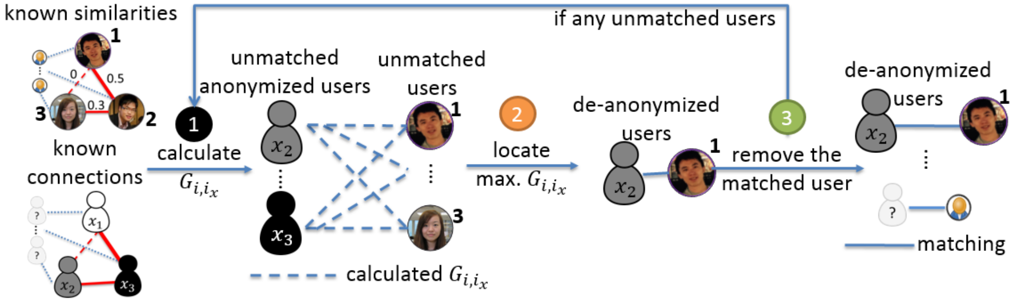

Fig. 7. Computation of de-anonymization

ALGORITHM 1: matching sets of user shared images and user online profiles

```
Data: $S_{i,j}$, user online profile of user $i$
Result: matched online profiles of user $i$ to unknown user $x_i$
# comment: estimate the gain, $G_{i,x_i}$;
set all $G_{i,x_i} = 0$ $sampleCount = 1$;
arraySample = [];
while sampleCount< $N_s$ do
    $A$ = rand(assignment);
    if mean($S_{i,j}$, connected)>mean($S_{i,j}$, unconnected) then
        arraySample<< $A$;
        $sampleCount++$;
    end
end
for $i = 1, i < N^{(u)}$ do
    for $i_x = 1, i_x < N^{(u)}$ do
        compute $G_{i,x_i}$ by Eq. 12
    end
end
# comment: matching starts;
unknownUnmatchedUser = all users;
knownUnmatchedUser = all users;
matching = [];
while UnknownUnmatchedUser.size>0 do
    $user_{x_i}$ = maxElementinMatrix($G_{i,x_i}$ in unknownUnmatchedUser);
    $user_i$ = maxElement($user_i$ in knownUnmatchedUser);
    match($user_{x_i}$)=$user_i$;
    remove $user_{x_i}$ in unknownUnmatchedUser;
    remove $user_i$ in knownUnmatchedUser;
end
```

(HVM) using Compute optimized instances (c3.xlarge). Each machine consists of 4 virtual CPUs (vCPUs), and 7.5 GB of memory.

6. EXPERIMENTAL RESULTS

This section shows the experimental results based on the 1.6 million images from the image-oriented social network, Fotolog. Section 6.1 describes the setting of the experiments, while the next part is the results of the de-anonymization. Section 6.3 and Section 6.4 are 2 showcases that demonstrate how to use user shared images for origin inference and gender identification. Section 6.5 ends this section with a discussion.

6.1. Experimental Settings

The goal of the experiment is to investigate which information, origins, genders or SGs, is the most sensitive to de-anonymization with user shared images. There are 3 steps to the experiment. In the first step, the 1.6M user shared images from 6,036 users on Fotolog are input to the system for BoFT, and each image is labeled with a non-user generated label, a BoFT label. The connections based on the targeted user profile information are discovered from the user shared images through the distribution of the BoFT labels. Then the user online profiles of the 6,036 users are matched with the connections discovered with the user shared images for de-anonymization, as proposed in the previous section, in step 2. The user identity is hidden, which means that the correspondences between the user shared images and the user online profiles are hidden to the system. The de-anonymized users are then compared with the ground truth, the actual correspondences of user shared images and the users, which is measured by accuracy, in step 3. The different types of information are evaluated one by one in the next subsection. This is followed by 2 showcases that demonstrate how user shared images are useful in origin inference and gender identification. The showcases demonstrate how user privacy can be invaded, even without matching online profiles with shared images.

6.2. De-Anonymization Results

This section shows the results in terms of accuracy for partial origins and genders information from online profiles as well as SGs. As mentioned in Section 6.1, the goal of this experiment is to evaluate which of the types of information is more sensitive to de-anonymization by matching the users with that type of information with the user shared images. Some users only provide their gender (4,342 out of 6,036), while some only provided their origin (5,806 out of 6,036). As the origin and gender are not available for some users, the matching rate for information that is only available among a small number of users could be higher as the possible matching sets are smaller. The matching accuracy is normalized for a fair comparison, by dividing the accuracy by that of the random approach, giving the normalized accuracy, and the results are shown in Fig. 8. Another approach using SGs is also implemented to compare with user shared images only when the SG is available online for scraping. The same procedure is applied, but instead of using the distribution of BoFT labels as the user profile information for similarity calculation, it makes use of the friendships two users shared for cosine similarity calculation.

It is observed that matching the anonymized SGs with the publicly and widely available user shared images gives a normalized accuracy of 11.5, while the accuracy is 5.15 for anonymized origin and 4.4 for anonymized gender. The performances of SGs and the proposed method are similar. It can be concluded that, given user shared images, anonymized SGs are the most sensitive information to de-anonymization, i.e, de-anonymization is most effective. SGs are shown to be 2 times and 2.5 times more sensitive than origins and genders, respectively. In privacy protection, one of the goals is for multiple users to have the same characteristics, such as k-anonymity, in anonymized SGs [LeFevre et al. 2005]. As there are more than 100 possible origins, and only 2 possible genders, it would be expected that origin should be much more sensitive information than gender. However, the results show that they are equally sensitive.

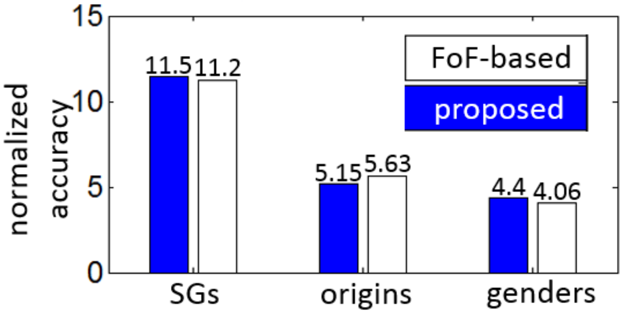

Fig. 8. Normalized accuracy of matching with different information with user shared images.

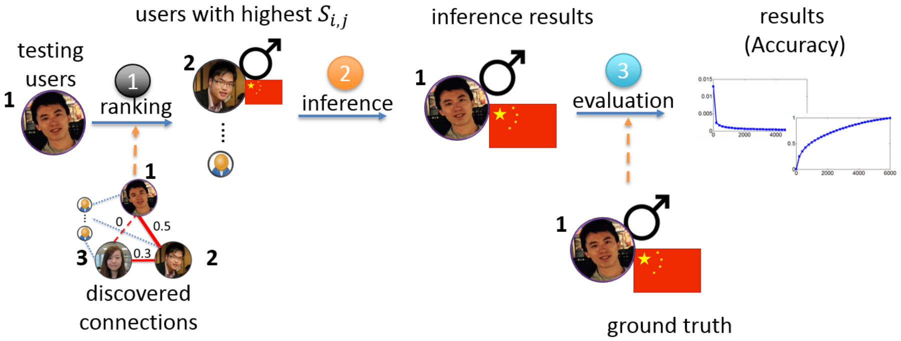

Fig. 9. Key steps in origin inference and gender identification.

6.3. Showcase 1: Origin Inference

This section shows the effectiveness of using user shared images to infer the origins of users based on $S_{i,j}$. As mentioned in [Narayanan and Shmatikov 2009], user privacy can be invaded even if profile information and the anonymized data cannot be matched. For example, for a marketing company, users' origins and genders are helpful to make recommendations and sell products. From the company perspective, if the gender and origin are known, it is as good as matching the anonymized online profile to users in some application, while from the user perspective, one would like to protect all information. It is, therefore, useful to investigate how much information user generated images alone can tell. In the second experiment, user shared images are used as the input for recommendation, user origin inference and gender identification, even without any anonymized online profile information or SGs from social media.

Although the origin may not be available in a user's publicly available online profile, it is possible to obtain this information. It is not necessary that those users with known origin to be the interested user, or have online friendships with interested users. According to Fig. 5 (b), two users are more likely to be from the same origin if they have a higher $S_{i,j}$. Based on this observation, a K-NN based classifier is built to demonstrate how the origin of a user can be inferred from $S_{i,j}$. In the experiment, a set of users is selected randomly as the training set, and the remaining users are the testing set. For each user, the inferred origin is the most frequently occurring origin of the top J users, which is the users who have the highest $S_{i,j}$ with user i, and the results are measured in terms of accuracy, the percentage of inferred origins that are the actual origins of the users. Fig. 9 shows the key steps of origin inference. The experiment is repeated 100 times, and the accuracy is shown in Fig. 10. In Fig. 10 (a), the accuracy is measured by the percentage of users in the training set. When there are more users in the training set, a better inference can be obtained. In Fig. 10 (b), the accuracy is measured by selecting the origin of J users who have the highest $S_{i,j}$ with user i. It is observed that when J grows, the accuracy saturates at a certain level, which means that a certain accuracy can be achieved with a sufficiently large J. As most of the origins apply to only a few users, as shown in Fig. 13, a large J may push the results towards those origins that apply to a lot of users. A larger dataset may be able to solve the problem.

6.4. Showcase 2: Gender Identification

This section shows the effectiveness of using user shared images to identify gender based on $S_{i,j}$. Similar to user origin inference, two users are more likely to be of the same gender with a higher $S_{i,j}$ as shown in Fig. 5 (c). The experiment in this subsection

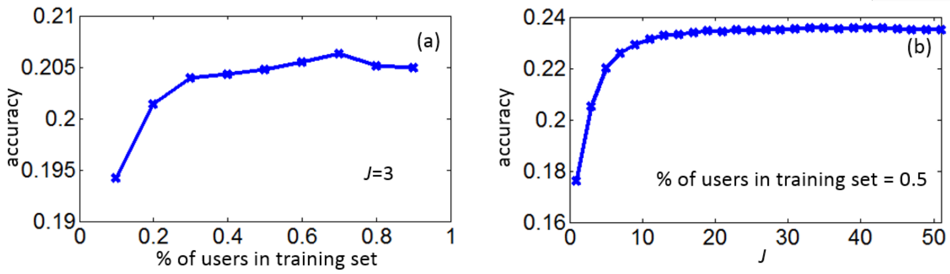

Fig. 10. The accuracy of user origin inference with different: (a) % of users in the training set, (b) J.

uses the same settings as the origin inference experiment. A K-NN based classifier is built and users are divided into a training set and a testing set randomly. For each user in the testing set, the identified gender is the most frequently occurring gender of a set of J users who have the highest $S_{i,j}$ with user i, and the results are measured in terms of accuracy, which is the percentage of inferred genders that are the actual genders of the users. The experiment is repeated 100 times, and the accuracy is shown in Fig. 11. In Fig. 11 (a), the accuracy is measured by the percentage of users in the training set. When there are more users in the training set, a better inference can be obtained. This information is easily available, as it can be scraped from social networks, and it is not necessary for those users to be included in the testing dataset. In Fig. 11 (b), the accuracy is measured by the J users, the number of users who have the highest $S_{i,j}$ with user i. It is observed that when J grows, the accuracy saturates at a certain level, which means that a certain accuracy can be achieved with a sufficiently large J.

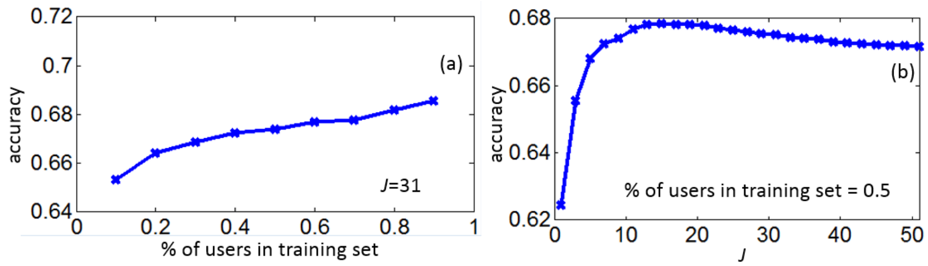

Fig. 11. The accuracy of gender identification with different: (a) % of users in training set, (b) J.

6.5. Discussion

This paper has successfully proved and characterized the phenomenon that two users are likely to be friends, be from the same origin and be of the same gender, if their shared images are similar. As the features extracted from images are a low-level descriptor, two images with exactly the same feature vector could be two completely different images. This could be solved by combining other forms of feature vectors, such as the distributions of color-based, or other feature extraction techniques, such as GIFT [Cheung et al. 2015b]. The use of different techniques could indicate different dimensions of the images, such as texture, color and more. Besides using different techniques, how to handle billions of images generated everyday is another challenge. A big data system, such as a cloud-assisted system to handle profile learning and similarity calculation [Jie et al.] for a scalable system, is a must to handle such images.

In the future, it would be interesting to investigate if the same observations could be formed for other online profile information, such as interests and occupation. Although it is proven in [Cheung et al. 2015a] that a higher $S_{i,j}$ implies a higher probability that 2 users are friends, it is not clear if the same phenomenon can be observed in different social networks and techniques. Further investigation is needed. Besides matching the profile directly using $S_{i,j}$, the 2 showcases above also demonstrate how a user online profile can be reconstructed from shared images, and this information could potentially help social graph de-anonymization.

7. CONCLUSIONS

This work has investigated 1,598,769 user shared images by 6,036 users on Fotolog, an image-oriented social network. Based on intensive measurements and characterizations of these user shared images, this work has proved the phenomenon that two users with a higher similarity between their shared images are likely to be of the same gender or origin or have an online friendship between them. From this phenomenon, an analytic system using bag-of-features tagging to de-anonymize a user's identity using their shared images is proposed and verified by nearly 1.6 million shared images. It is observed that friendship is the most sensitive information for disclosing user identity. This paper has also presented 2 showcases to demonstrate the effectiveness of using user shared images for gender identification and origin inference. The experiments show that using user shared images is effective to disclose user identity. To the best of our knowledge, this is the first paper to evaluate how user shared images can be used to invade user privacy, and to propose a system to de-anonymize user identity by matching their shared images with anonymized profile information and friendships. With the advances in wearable devices and smart mobile devices, sharing images on social media has become a norm, so how to protect user privacy in shared images will become more important.

APPENDIX

The appendix measures the characteristics of user shared images, online friendships, and origins. Fig. 12 (a) and Fig. 12 (b) show the distributions of the number of user shared images and friendships users have, respectively. It is observed that a few users share a large number of images and have many friendships, while most users share a few images or have a few friendships only, and the same trend can be observed on most social networks [Mislove et al. 2007]. Another measurement of the origins of the users was conducted, and the results are shown in Fig. 13. It is observed that most of the origins have only a few users, while a few origins have a large number of users. This represents the fact that Fotolog is more popular in certain countries.

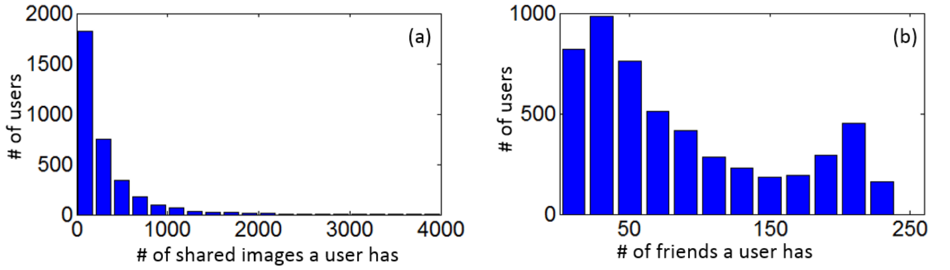

Fig. 12. Distribution of the number of shared images and friendships a user has: (a) shared images, (b) friendships.

Besides the average $S_{i,j}$, it is also interesting to investigate the distribution of $S_{i,j}$

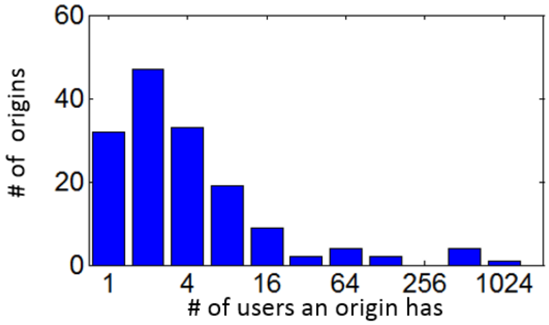

Fig. 13. Distribution of the number of users for an origin.

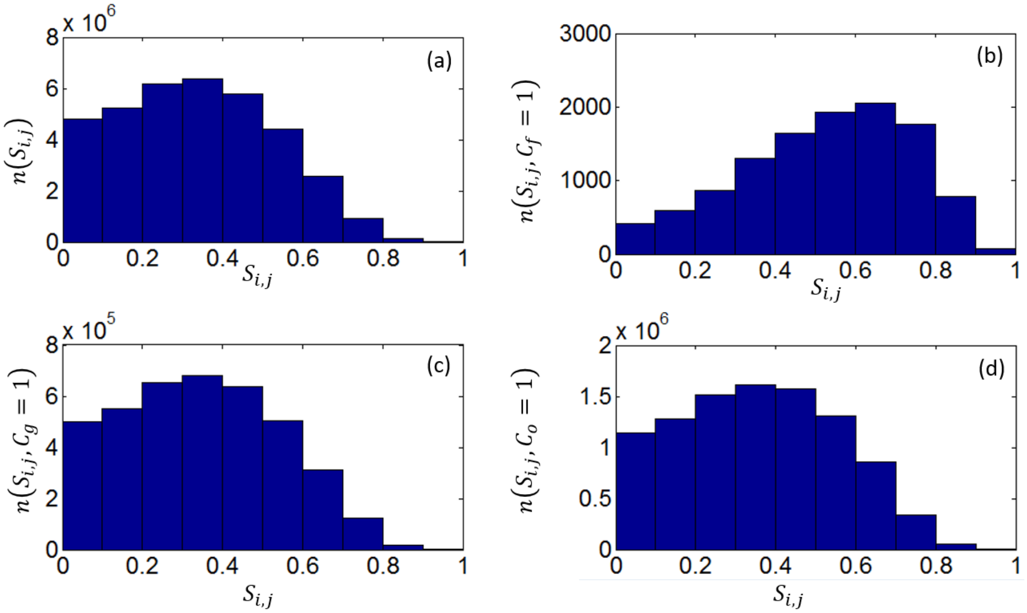

Fig. 14. Distribution of $S_{i,j}$ among pairs of: (a) all pairs, (b) related pairs ($C_f = 1$), (c) pairs with of same gender ($C_g = 1$), (d) pairs with same origin ($C_o = 1$).

among users with respect to different types of connections of user pairs. Fig. 14 (a) shows the distribution of $n(S_{i,j})$, which is the number of all pairs given a $S_{i,j}$, from Fotolog. This is the distribution of all pairs, regardless of the types of connections of the pairs. Fig. 14 (b), (c) and (d) show the distributions of the number of connected pairs based on SGs, gender and origin, given a $S_{i,j}$, and they are respectively represented by $n(S_{i,j}, C_f = 1)$, $n(S_{i,j}, C_g = 1)$ and $n(S_{i,j}, C_o = 1)$. It is observed that the distributions are similar. They reach a peak value and decrease gradually, and there are only a few pairs that have a high $S_{i,j}$. Note that as the numbers of users who provide their gender or origin are different, the numbers of possible pairs in Fig. 14 (b), (c) and (d) are also different. There are 4,342 and 5,805 users, as well as 9.4M and 17M possible pairs who provide their gender and origin, respectively.

ACKNOWLEDGMENTS

This work is supported by the HKUST-NIE Social Media Lab., HKUST.

REFERENCES

Shane Ahern, Dean Eckles, Nathaniel S Good, Simon King, Mor Naaman, and Rahul Nair. 2007. Overexposed?: privacy patterns and considerations in online and mobile photo sharing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 357–366.

Marco Balduzzi, Christian Platzer, Thorsten Holz, Engin Kirda, Davide Balzarotti, and Christopher Kruegel. 2010. Abusing social networks for automated user profiling. In *Recent Advances in Intrusion Detection*. Springer, 422–441.

Joseph Calandrino, Ann Kilzer, Arvind Narayanan, Edward W. Felten, and Vitaly Shmatikov. 2011. " You Might Also Like:" Privacy Risks of Collaborative Filtering. In *Security and Privacy (SP), 2011 IEEE Symposium on*. IEEE, 231–246.

Patrizio Campisi, Emanuele Maiorana, and Alessandro Neri. 2009. Privacy protection in social media networks a dream that can come true?. In *Digital Signal Processing, 2009 16th International Conference on*. IEEE, 1–5.

Ming Cheung and James She. 2014. Bag-of-features tagging approach for a better recommendation with social big data. In *Proceedings of the 4th International Conference on Advances in Information Mining and Management (IMMM'14)*. 83–88.

Ming Cheung, James She, and Jie Allan. 2015a. Connection Discovery using Big Data of User Shared Images in Social Media. *Multimedia, IEEE Transactions on* (2015).

Ming Cheung, James She, and Li Xiaopeng. 2015b. Non-user Generated Annotation on User Shared Images for Connection Discovery. In *Proceedings of The IEEE International Conference on Cyber, Physical and Social Computing (CPSCom 15)*.

Eric Gilbert and Karrie Karahalios. 2009. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 211–220.

Scott A. Golder and Sarita Yardi. 2010. Structural predictors of tie formation in twitter: Transitivity and mutuality. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*. IEEE, 88–95.

Bin Guo, Huihui Chen, Zhiwen Yu, Xing Xie, Shenlong Huangfu, and Daqing Zhang. 2015. FlierMeet: A mobile crowdsensing system for cross-space public information reposting, tagging, and sharing. *Mobile Computing, IEEE Transactions on* 14, 10 (2015), 2020–2033.

Ido Guy, Naama Zwerdling, David Carmel, Inbal Ronen, Erel Uziel, Sivan Yogev, and Shila Ofek-Koifman. 2009. Personalized recommendation of social software items based on social relations. In *Proceedings of the third ACM conference on Recommender systems*. ACM, 53–60.

William H. Hsu, Andrew L. King, Martin S. Paradesi, Tejaswi Pydimarri, and Tim Weninger. 2006. Collaborative and Structural Recommendation of Friends using Weblog-based Social Network Analysis.. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. 55–60.

Paridhi Jain, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. @ i seek'fb. me': identifying users across multiple online social networks. In *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 1259–1268.

Zhanming Jie, Ming Cheung, and James She. A Cloud-Assisted Framework for Bag-of-Features Tagging in Social Networks. In *Network Cloud Computing and Applications. Proceedings. 4th IEEE Symposium on*. IEEE.

Emily M Jin, Michelle Girvan, and Mark EJ Newman. 2001. Structure of growing social networks. *Physical review E* 64, 4 (2001), 046132.

Timor Kadir and Michael Brady. 2001. Saliency, scale and image description. *International Journal of Computer Vision* 45, 2 (2001), 83–105.

Jon M. Kleinberg. 2007. Challenges in mining social network data: processes, privacy, and paradoxes. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 4–5.

Mohammed Korayem and David J. Crandall. 2013. De-Anonymizing Users Across Heterogeneous Social Computing Platforms.. In *ICWSM*.

Linyuan L and Tao Zhou. 2011. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications* 390, 6 (2011), 1150–1170.

Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. 2005. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 49–60.

Yoseph Linde, Andres Buzo, and Robert M. Gray. 1980. An algorithm for vector quantizer design. *Communications, IEEE Transactions on* 28, 1 (1980), 84–95.

Wendy Liu and Derek Ruths. 2013. What's in a Name? Using First Names as Features for Gender Inference in Twitter.. In *AAAI Spring Symposium: Analyzing Microtext*.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

Andrew McCallum, Kamal Nigam, and Lyle H. Ungar. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 169–178.

Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2007. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 29–42.

Emily Moxley, Jim Kleban, Jiejun Xu, and BS Manjunath. 2009. Not all tags are created equal: Learning Flickr tag semantics for global annotation. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 1452–1455.

Arvind Narayanan, Elaine Shi, and Benjamin I. Rubinstein. 2011. Link prediction by de-anonymization: How we won the kaggle social network challenge. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 1825–1834.

Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing social networks. In *Security and Privacy, 2009 30th IEEE Symposium on*. IEEE, 173–187.

Rohit Parimi and Doina Caragea. 2011. *Predicting friendship links in social networks using a topic modeling approach*. Springer, 75–86.

Yi Song, Sadegh Nobari, Xuesong Lu, Panagiotis Karras, and Stphane Bressan. 2011. On the privacy and utility of anonymized social networks. In *Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services*. ACM, 246–253.

Kurt Thomas, Chris Grier, and David M. Nicol. 2010. unfriendly: Multi-party privacy risks in social networks. In *Privacy Enhancing Technologies*. Springer, 236–252.

Andrea Vedaldi and Brian Fulkerson. 2010. VLFeat: An open and portable library of computer vision algorithms. In *Proceedings of the international conference on Multimedia*. ACM, 1469–1472.

Gilbert Wondracek, Thorsten Holz, Engin Kirda, and Christopher Kruegel. 2010. A practical attack to de-anonymize social network users. In *Security and Privacy (SP), 2010 IEEE Symposium on*. IEEE, 223–238.

Xing Xie. 2010. Potential friend recommendation in online social network. In *Green Computing and Communications (GreenCom), 2010 IEEE/ACM Int'l Conference on & Int'l Conference on Cyber, Physical and Social Computing (CPSCom)*. IEEE, 831–835.

Shuang-Hong Yang, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng, and Hongyuan Zha. 2011. Like like alike: joint friendship and interest propagation in social networks. In *Proceedings of the 20th international conference on World wide web*. ACM, 537–546.

Quanzeng You, Sumit Bhatia, and Jiebo Luo. 2015. A picture tells a thousand wordsAbout you! User interest profiling from user generated visual content. *Signal Processing* (2015).

Quanzeng You, Sumit Bhatia, Tong Sun, and Jiebo Luo. 2014. The eyes of the beholder: Gender prediction using images posted in online social networks. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*. IEEE, 1026–1030.

Zhiwen Yu, Huang Xu, Zhe Yang, and Bin Guo. 2016. Personalized Travel Package With Multi-Point-of-Interest Recommendation Based on Crowdsourced User Footprints. *Human-Machine Systems, IEEE Transactions on* 46, 1 (2016), 151–158.

Kuan Zhang, Xiaohui Liang, Xuemin Shen, and Rongxing Lu. 2014. Exploiting multimedia services in mobile social networks from security and privacy perspectives. *Communications Magazine, IEEE* 52, 3 (2014), 58–65.

Xiaoming Zhang, Xiaojian Zhao, Zhoujun Li, Jiali Xia, Ramesh Jain, and Wenhan Chao. 2012. Social image tagging using graph-based reinforcement on multi-type interrelated objects. *Signal Processing* (2012).

Elena Zheleva and Lise Getoor. 2009. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th international conference on World wide web*. ACM, 531–540.

Received July 2015; revised April 2016; accepted April 2016

Online Appendix to:
Evaluating the Privacy Risk of User Shared Images

Ming Cheung, HKUST-NIE Social Media Lab
James She, HKUST-NIE Social Media Lab

DOI: 0000001.0000001