Connection Discovery using Big Data of User Shared Images in Social Media

Ming Cheung, *Student Member, IEEE,*, James She, *Member, IEEE,* and Zhanming Jie

***Abstract*—Billions of user shared images are generated by individuals in many social networks today, and this particular form of user data is widely accessible to others due to the nature of online social sharing. When user social graphs are only accessible to exclusive parties, these user shared images are proved to be an easier and effective alternative to discover user connections. This work investigated over 360,000 user shared images from two social networks, Skyrock and 163 Weibo, in which 3 million follower/followee relationships are involved. It is observed that the shared images from users with a follower/followee relationship show relatively higher similarities. A multimedia big data system that utilizes this observed phenomenon is proposed as an alternative to user generated tags and social graphs for follower/followee recommendation and gender identification. To the best of our knowledge, this is the first attempt in this field to prove and formulate such a phenomenon for mass user shared images along with more practical prediction methods. These findings are useful for information or services recommendations in any social network with intensive image sharing, as well as for other interesting personalization applications, particularly when there is no access to those exclusive user social graphs.**

***Index Terms*—big data, user shared images, connection, discovery, recommendation, social network analysis**

I. Introduction

USER connection is useful information for many personalised services or applications in online social networks. Such connections can be any type of online social relationship formed from some interactions between users in a social network, such as online friendship, a follower/followee relationship or a membership in the same community. Companies like Twitter and Pinterest, already have explicit information about user online friendships (i.e., social graphs) to improve their service relevance to users. Trending mobile social applications, such as Instagram (owned by Facebook from the US) and WeChat (owned by Tencent from China), keep the information of social graphs (SGs) only available to their related business services. Some users also hide or limit the information of their connections from the public in social media platforms due to privacy concerns. Accessing these SGs is getting more difficult and costly in today's online social networks, and novel applications using SGs become almost impossible to be offered independently by researchers, merchants, third-party practitioners and individuals. However, billions of user shared images are generated by individuals in many social networks daily, and this particular form of user data is indeed very accessible to others due to the nature of online image sharing. Hence, a common but unreliable alternative is using user annotated tags (or user tagging) associated with each shared image to discover user connections when the SG is not accessible. In this work, using user shared images directly to discover the user connections in follower/followee relationships through some signal processing technique (e.g., bag-of-features) is proved to be effective. Users with connections of follower/followee relationships are found to give relatively higher similarities of the visual features in their shared images. An extreme example of user generated images on Instagram is shown in Fig. 1: Both users A and B share images about cars and user C shared an image about a flower. The follower/followee relationship between users A and B can be possibly detected from the higher similarity of visual features in their shared images. When more shared images from each of users A, B and C are accessible for evaluation, the actual follower/followee relationships should become reliably and accurately detectable though becoming challenging to process when the number of shared images and user connections grows bigger and faster every day in social network.

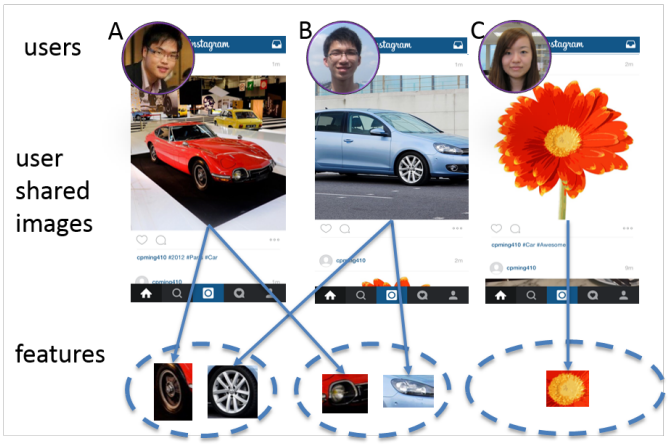

Fig. 1: Examples of the user shared image and their features

With the above motivations and challenges, this work has investigated over 360,000 user shared images and 3 million follower/followee relationships from 2 social networks - Skyrock from Europe and 163 Weibo from China. An interesting phenomenon of user shared images is observed from our intensive measurements, and this is formulated with a proposed method for a system to discover and recommend user connections in follower/followee relationships using user shared images directly. In summary, the contributions of this paper includes the following: 1) intensive measurements and characterizations of user shared images from two social networks, proved and formulated the phenomenon that two users with a higher similarity of their shared images are likely

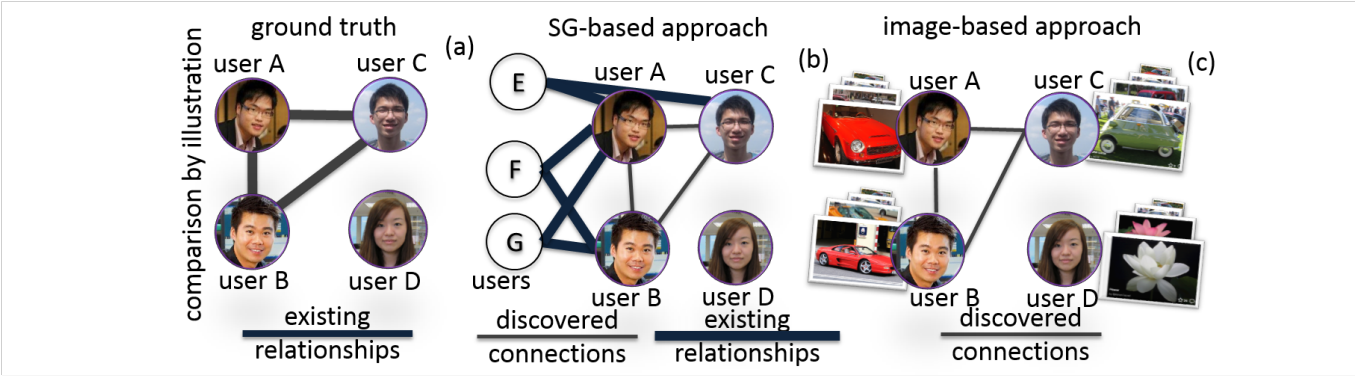

Fig. 2: Examples of connection discovery with different approaches, (a) ground truth, (b) SG-based, (c) image-based.

to have a connection in a follower/followee relationship; 2) methods using bag-of-features tagging (BoFT) are proposed as a recommendation system to discover user connections and recommend follower/followee relationships by their shared images; 3) extensive verifications of the proposed formulation, methods and system are provided with the datasets from two social networks (one from Europe, and one from China) and two practical use cases to prove the effectiveness of using user shared images through bag-of-features tagging as a better alternative to using user annotated tags for recommending follower/followee relationship and gender prediction.

This paper is organized as follows: section II presents the related works. Section III introduces the proposed method, BoFT, for connection discovery, while section IV shows the measurements of user shared images on the datasets. Section V proposes and formulates the follower/followee recommendation system, followed by the experimental results in section VI. Section VII concludes the paper and the future works.

II. Related Works

User behaviors in online social networks have been recently studied through the use of SGs in [1][2][3][4][5][6][7][8][9][10], and it was concluded that relationships (such as follower/followee and online friendships) in an SG are not formed randomly, but follow the power law distribution [11]. User connection can be therefore discovered and the connection strength can be obtained. Two users with a higher connection strength (more common related users) are more likely to be related, and the relationships are therefore predicted by their mutuality [12][13][14][15] from the strength of the discovered connections. An example is shown in Fig. 2 (b), where connection discovery is made via existing relationships, such as follower/followee, that users share in common. Users A, B and C share common related users, and connections among them can be obtained, while user D is alone.

Without access to SGs, follower/followee recommendation is also possible with the connection discovered by user common interests inferred from user input [15][16][17] or user generated content [18][19] and other personal information [1][3][20][21][22]. Analyzing shared images can help to understand users, and hence discover the connections among them [23][24]. Another common method to discover user connections is to analyze user annotated tags on shared images [23][24][25][26], in which a tag in the form of textual wording is provided by a user as meta-data to describe the shared image. These tags can represent users, and connections can be therefore discovered by calculating the similarity of the user annotated tags. However, user generated tags are unreliable [10][23][27][28] due to the use of different and inconsistent language, different levels of detail and even inaccurate or missing words, which results in noisy or low performing connection discovery. Collaborative filtering (CF) techniques [29][30][31] are used to improve the tag accuracy for better connection discovery. However, only some popular images are annotated by many users, while the rest are either not correctly annotated or missing annotation, which leads to a poor connection discovery performance [23]. An emerging image-based approach [28][32] applies computer vision techniques to produce non-user generated labels that reflect the context of the images, regardless of their popularity. Fig. 2 (c) is an example of how connections are discovered by user generated images. Users A, B and C generate car images, and their connections can be discovered by the similarity among their images.

As there is no model of the characteristics of user shared images and how BoFT similarity is distributed in [23], this paper has extended [23] on connection discovery using BoFT in the following way: 1) scraped two new datasets, Skyrock and 163 Weibo, for 360,000 user shared images and 3 million connections; 2) measured and modeled characteristic of user shared images and BoFT similarity distribution to explain how they are related to the connection discovery and follower/followee relationships; 3) formulated follower/followee recommendation based on the measurements and models, and verified that connection discovery is useful for recommendation and gender identification, even without using SGs and user annotated tags.

III. BoF Tagging and Similarity

This section introduces the proposed method, BoFT, that labels images with non-user generated labels, BoFT labels, and how

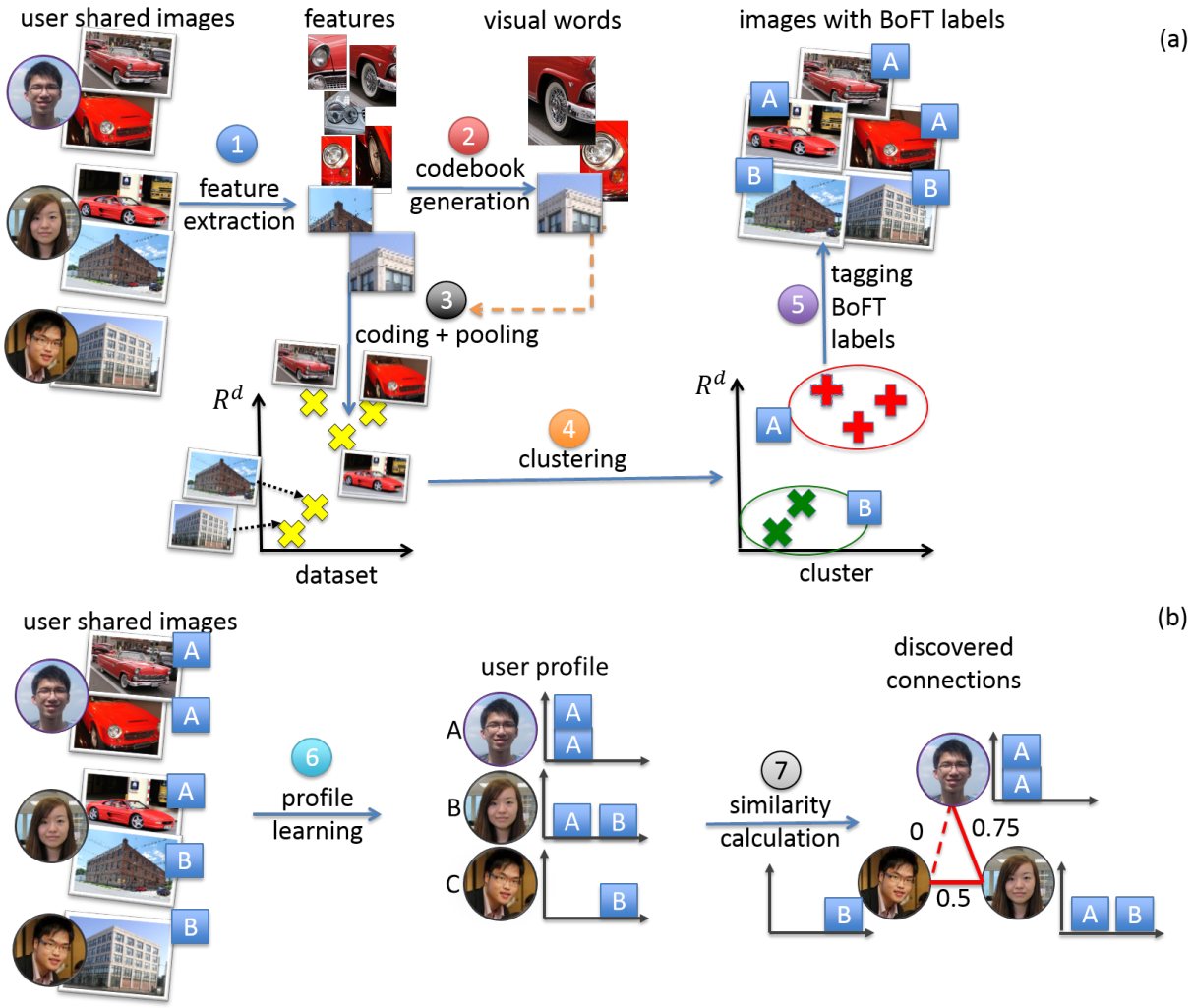

Fig. 3: BoFT: (a) annotation with BoFT labels, (b) user similarity calculation based on BoFT labels.

BoFT similarity, the pairwise similarity among users based on BoFT labels, is calculated.

A. BoF-Based Tagging

Images are analyzed using BoFT, which annotates each image with a BoFT label. BoF is a popular computer vision approach for analyzing images [33]. Fig. 3 shows the key steps involved: Fig. 3 (a) is the steps for BoF and Fig. 3 (b) is the method for connection discovery based on user shared images. The different steps of BoFT are introduced in this section below.

1) Feature Extraction: Feature extraction is a process to obtain the unique local features in step 1 of Fig. 3 (a). These unique features can be detected by feature detection, such as the Harris Affine detector, Maximally Stable Extremal Regions detector [33] and KadirBrady saliency detector [34]. 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) [35].

2) Codebook Generation: Codebook generation in step 2 of Fig. 3 (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 [36] and LindeBuzoGray algorithm [37]. A k-means clustering is used in our work.

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. 3 (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.

4) Clustering and BoFT Labeling: Clustering groups images that are visually similar through the similarity in their feature vectors, whihc is shown in step 4 of Fig. 3 (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

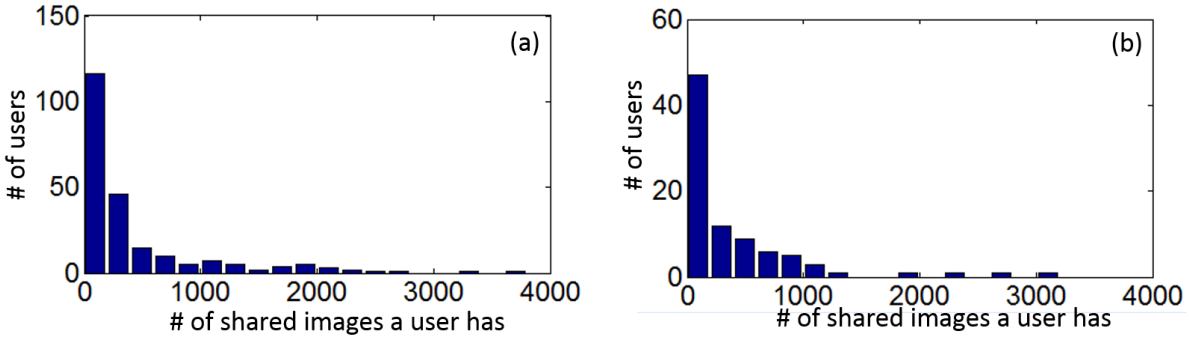

Fig. 5: Distribution of the number of shared images a user has: (a) Skyrock, (b) 163 Weibo.

Fig. 4: User interface of: (a) Skyrock, (b) 163 Weibo.

will first randomly generates 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. More discussion of this can be found in Section VI. 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. 3 (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 input and process.

B. Connection Discovery with BoFT labels

This section introduces how connection can be discovered through BoFT labels.

1) BoFT Labels and User Profile: A user profile, which reflects the content of a user's shared images, is the key in connection discovery. The proposed method uses the number of occurrences of the BoFT labels in step 5 of Fig. 3 (a) of the shared images of a user as his/her user profile, as in step 6 of Fig. 3 (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. The best value of K is subject to applications, and more discussion can be found in Section VI.

2) User Profile and User BoFT Similarity: When the user profile of each user is established, the next step is the connection discovery 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 set of BoFT labels of the shared image in the user profiles of users i and j, respectively.

IV. Measurements on User Shared Images

This section first describes the dataset, followed by the characteristics of the user shared images and follower/followee relationships. The third part of this section analyzes the BoFT similarity distribution by BoFT [23].

A. The datasets

Skyrock is a Western social networking site that allow users to create blogs, follow other users and exchange messages. Most of the users of Skyrock are from European countries like France, England, German, Holland, etc. 163 Weibo was a microblogging social network application from China with a similar mechanism to Twitter. As the user bases of the two social networks are different, it is interesting to observe whether the user behaviors in these social networks are similar. Fig. 4 shows the user interfaces of the two social networks, Skyrock and 163 Weibo. Skyrock, as shown in Fig 4 (a), users can share blogs with text, images and even video. On 163 Weibo, users share text and image content, as shown in Fig 4 (b). Similar to any social network, users of these networks can follow others

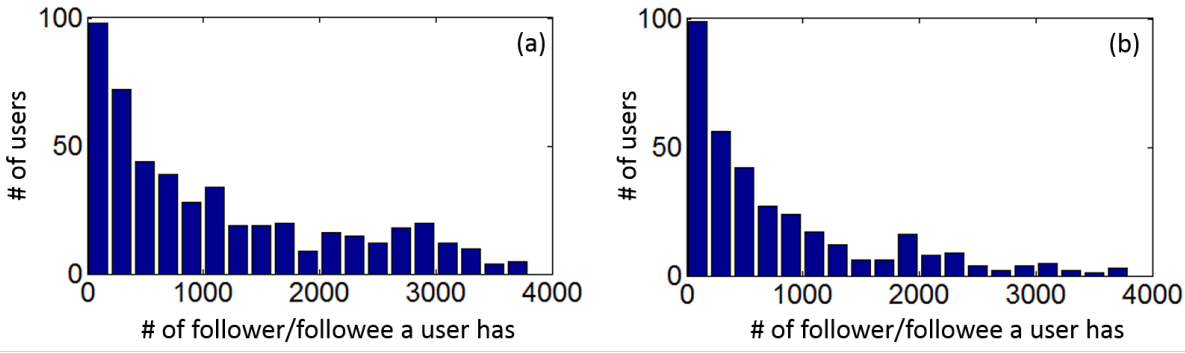

Fig. 6: Distribution of the number of follower/followee relationships a user has: (a) Skyrock, (b) 163 Weibo.

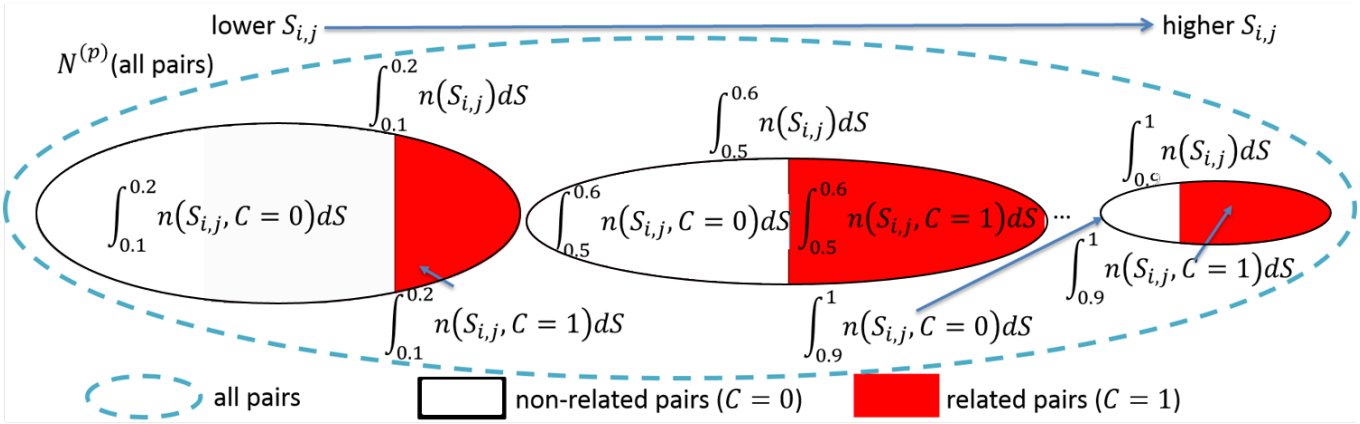

Fig. 7: The set of all pairs: the solid line ellipses formed are user pairs with a given $S_{i,j}$, with the white area is the non-related pairs and the coloured part the related pairs. The size of the ellipses represents the number of user pairs.

to receive notifications of newly shared content from those they follow. The experiments involve 176,547 images uploaded by 722 users on Skyrock, and 187,491 images uploaded by 493 users on 163 Weibo, which were collected by official APIs in mid-2014, before 163 Weibo was shut down at the end of 2014. All the users were selected randomly from a large set of users collected from follower/followee relationships, in which there are about 80,000 and 100,000 users collected for the selection on Skyrock and 163 Weibo, respectively. The datasets comprise more than 3 million follower/followee relationships, not only including those among the 722 users on Skyrock and 493 users on 163 Weibo, but also all the followers/followees those users have. Among the randomly selected users, the network densities are 1.41% and 0.71% on Skyrock and 163 Weibo, respectively. They are not likely to know others.

B. Characteristics of User Shared Images

This section describes the characteristics of the user shared images and the follower/followee relationships. Fig. 5 (a) and Fig. 5 (b) show the distribution of the number of user shared images a user has, and the frequency of this number, on Skyrock and 163 Weibo, respectively. It is observed that a few users share a large number of images, while most of the users share a few images only, and the same trend can be observed on both social networks. Fig. 6 (a) and Fig. 6 (b) show the distribution of the number of follower/followee relationships a user has, and the frequency of this number on Skyrock and 163 Weibo, respectively. The same trend that a few users have a large number of follower/followee relationships, while most of the users have a few follower/followee relationships only can also be observed. The same observation can be found in both social networks. It is obvious that the distribution of the number of shared images and follower/followee relationships on both social networks follows the power law distribution, as do most social networks. It is concluded the selected users are a good representation of the users in the two social networks.

C. BoFT Similarity Distribution

In this paper, there are two types of user pairs: related pairs, which are the pairs of users that are follower/followee, and non-related pairs, which are the pairs in which a follower/followee relationship does not exist between the two users. Related pairs and non-related pairs can be considered as two classes, and the class of each pair, C, can be defined as

$$C = \begin{cases} 1 & \text{if two users are a related pair} \\ 0 & \text{if otherwise,} \end{cases} \quad (3)$$

where $C = 1$ is the class in which the pair is a related pair, and $C = 0$ is the class in which the pair is a non-related pairs. Fig. 7 illustrates the idea. The ellipse in the broken line is the

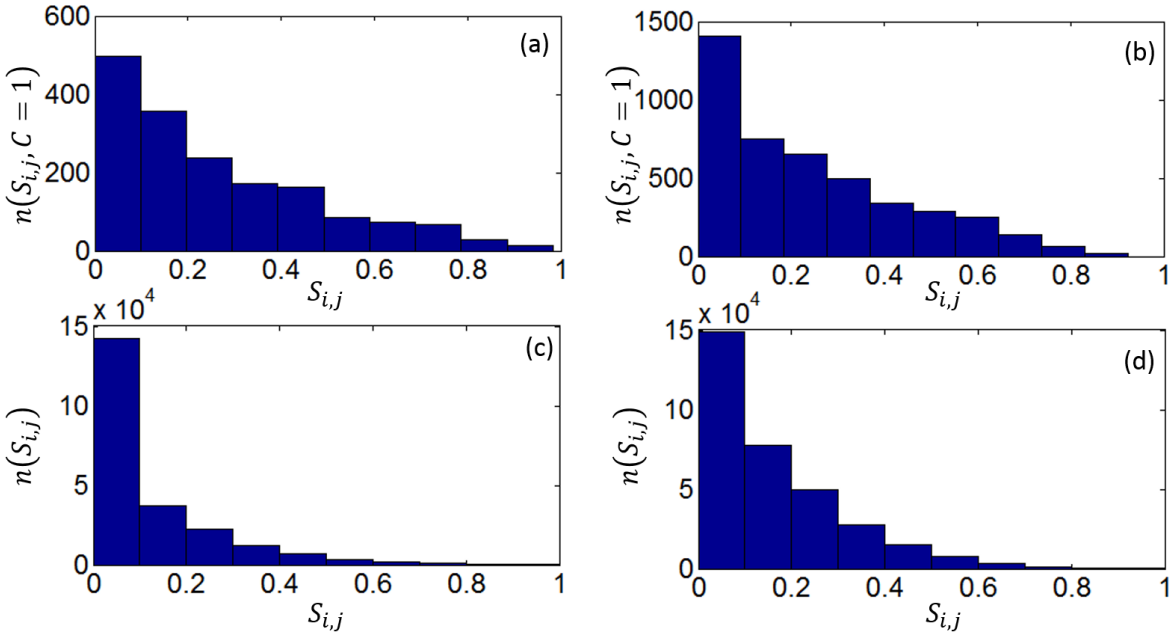

Fig. 9: Distribution of BoFT similarity among pairs of (a) follower/followee relationships on Skyrock, (b) follower/followee relationships on 163 Weibo, (c) all users on Skyrock, (d) all users on 163 Weibo.

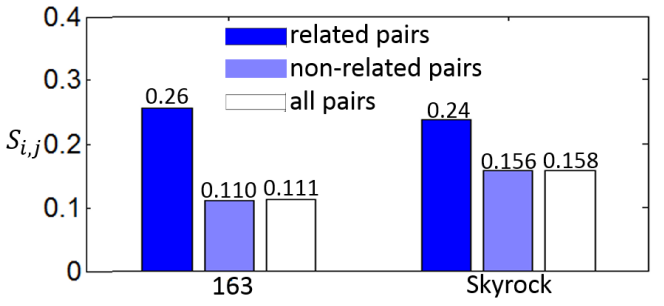

Fig. 8: Mean BoFT similarity: related, non-related and all pairs.

set of all pairs, with total $N^{(p)}$ pairs, which contains related and non-related pairs with different similarity values. The solid line ellipses are those with different ranges of $S_{i,j}$. The areas in white of the colored ellipses are the sets of non-related pairs with a given range of $S_{i,j}$, with $n(S_{i,j}, C=0)$ pairs. $n(S_{i,j}, C)$ is a function of $S_{i,j}$ and C that gives the number of non-related pairs when $C=0$, while giving the number of related pairs when $C=1$. The coloured areas are the sets of related pairs with a given range of $S_{i,j}$, with $n(S_{i,j}, C=1)$ pairs. Fig. 8 shows the average BoFT similarity, $S_{i,j}$, for related pairs, non-related pairs and all pairs. It is observed that both social networks have a similar trend: related pairs have a higher $S_{i,j}$, in which related pairs are 136% and 53.8% higher $S_{i,j}$ than non-related pairs on 163 Weibo and Skyrock, respectively. Since the number of all pairs, $N^{(p)}$, is much greater than the number of related pairs, $N^{(p)}_{C=1}$, $N^{(p)}$ is close to the number of non-related pairs, $N^{(p)}$, and gives a similar average $S_{i,j}$ for non-related and all pairs. It is interesting to investigate the distribution of $S_{i,j}$ of users, how it affects the mean value of $S_{i,j}$ and how it can be modeled. Fig. 9 (a) and (b) show the distribution of the number of related pairs, given a $S_{i,j}$, $n(S_{i,j}, C=1)$, of Skyrock and 163 Weibo, respectively. Fig. 9 (c) and (d) show the distribution of the number of all pairs, given a $S_{i,j}$, $n(S_{i,j})$, of Skyrock and 163 Weibo, respectively. It is observed that there are more related pairs with small BoFT similarities, while only a small number of them have a high $S_{i,j}$. The same observation can be found in the distribution of $S_{i,j}$ of all pairs, and can be modeled by exponential distributions:

$$f(S_{i,j}) = \gamma e^{-\lambda S_{i,j}}, \tag{4}$$

where $E[f(S_{i,j})]$ is equal to γ/λ^2. γ and λ are real numbers, and $E[f(S_{i,j})]$ is the mean of $f(S_{i,j})$. The probability that $S_{i,j}$ occurs in all pairs, $P(S_{i,j})$, can be calculated as:

$$P(S_{i,j}) = \frac{\int_a^b n(S_{i,j})ds}{N^{(p)}} \tag{5}$$

Similarly, the probability that $S_{i,j}$ occurs in related pairs, $P(S_{i,j}|C=1)$, can be calculated as:

$$P(S_{i,j}|C=1) = \frac{\int_a^b n(S_{i,j}, C=1)ds}{N^{(p)}_{C=1}} \tag{6}$$

where a equals $\lfloor BS_{i,j} \rfloor / B$ and b equals $\lceil BS_{i,j} \rceil / B$. B is the bin size, and $B = 10$ is used in this work. For example, $P(0.15)$ is equal to $\int_{0.1}^{0.2} n(S_{i,j})ds/N^{(p)}$. Fig. 10 shows $P(S_{i,j})$ and $P(S_{i,j}|C=1)$ of the best fit curves and the approximation of the two social networks. It is observed that the best fit curve goes well with the values. In the two social networks, related pairs have smaller γ and λ than all pairs, which implies that a higher $S_{i,j}$ means a higher probability

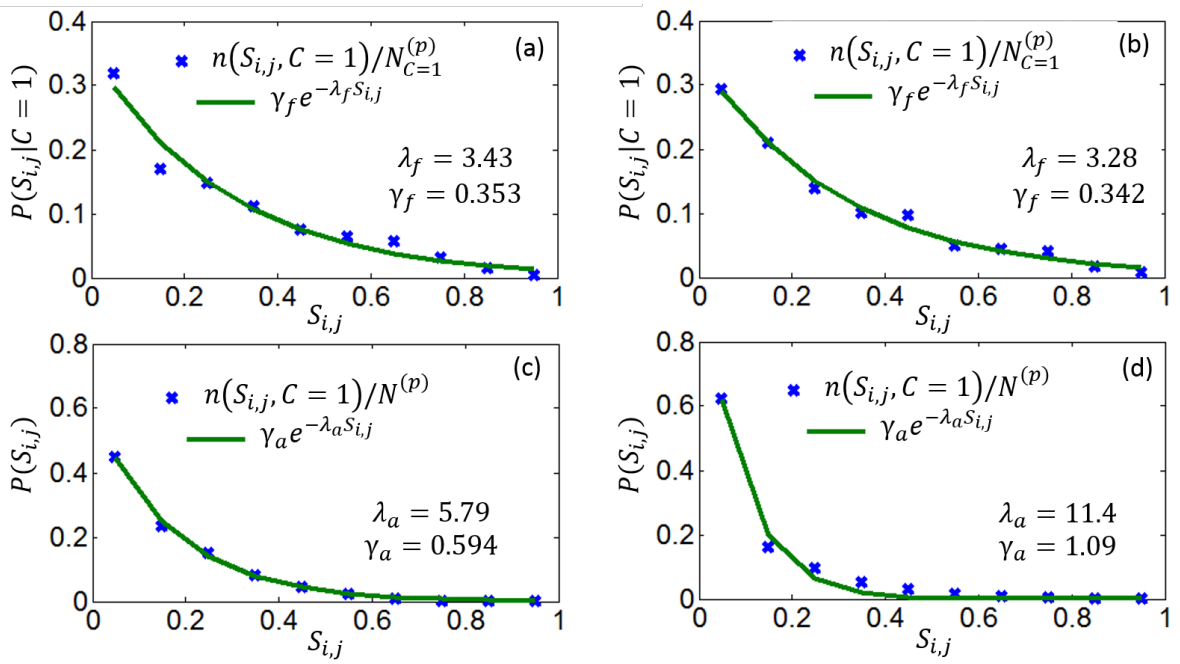

Fig. 10: $S_{i,j}$ against the probability of $S_{i,j}$, by the best fit curve and the measurement: (a) related pairs on Skyrock, (b) related pairs on 163 Weibo,(c) all pairs on Skyrock, (d) all pairs on 163 Weibo.

they are a related pair. This idea can be illustrated by the area of the colored ellipses in Fig. 7: although the area (number of pairs) is smaller when $S_{i,j}$ is higher, a larger area of the coloured part (a higher portion of related pairs) can be observed. The same observation can be obtained from both social networks.

V. Follower/Followee Recommendation using Discovered Connections

This section introduces the system flow and formulation of how follower/followee recommendation can be made with discovered connections. This is a 3-stage (stages A to C) systems as shown in Fig. 11. The first part is image collection, followed by connection discovery using BoFT. The third part focuses on how to recommend follower/followees based on the discovered connections and the BoFT similarity distribution. The stages are introduced one by one in this section.

A. Image Collection

The proposed system carries out data collection as shown in step A of Fig. 11, which shows the process to collect user generated images from social media applications, such as Skyrock and 163 Weibo. The images can be provided by the operators of the social media and mobile applications or collected through the API of the social networks. The user generated images can be shared in various forms, such as posted images on social media or images shared through instant messaging applications. On social networks such as Skyrock and 163 Weibo, user generated images are those images shared by users. This process is ongoing, which means that user shared images are collected continuously.

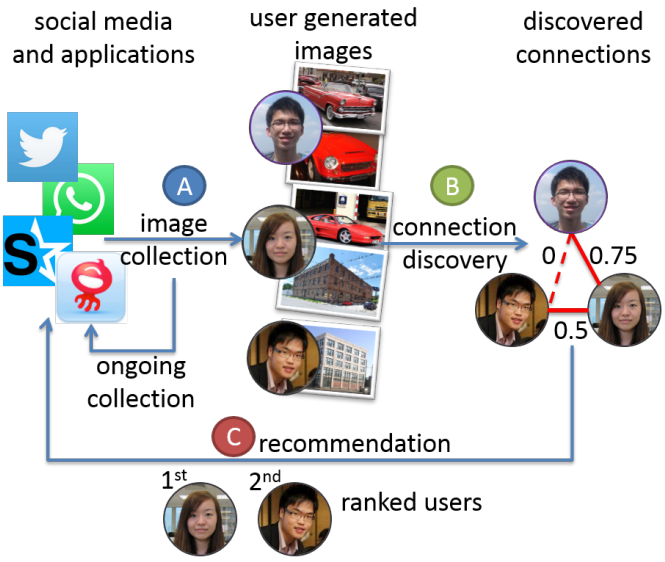

Fig. 11: System flow of the proposed system (a) image collection from social media; (b) connection discovery by collected images; (c) recommendation by discovered connections.

B. Connection Discovery using BoFT

The objective of the image understanding is to annotate user generated images with non-user annotated labels, as shown in step B of Fig. 11. The proposed system applies a computer vision approach to give a label to user generated images, which is not affected by the language, culture or other characteristics of the user who shares the image, but is based on the image's visual appearance only. The accuracy of the user generated tags is unreliable, sometimes even unavailable, and the per-

formance of connection discovery is affected. The proposed system applies BoFT to annotate user generated images with non-user annotated labels, called BoFT labels. The set of user shared images of user i is processed by the proposed method, and a set of BoFT labels, L_i, is generated to represent user i. As discussed, millions of images are generated every day, so a system that can process big data with scalable storage design is needed for collecting and processing these user shared images, such as a cloud-assisted system to handle profile learning and similarity calculation [38] for a scalable system. The feature vectors are first split into multiple blocks in the Hadoop Distributed File System (HDFS) and distributed to virtual machines (VMs) for the k-means clustering process. Each VM is in charge of computing the distribution of different labels for several users, and the BoFT similarity is also calculated in a distributed way.

C. Follower/followee Recommendation

Follower/followee recommendation is one of the most popular applications on social media.The probability that two users are a related pair, or $C = 1$, given the BoFT similarity of user i and j, $P(C = 1|S_{i,j})$ can be calculated by Eq. 2 based on L_i and L_j. Follower/followee recommendation should be made based on $P(C = 1|S_{i,j})$, from the highest to the lowest. This section starts with discussions on how $P(C = 1|S_{i,j})$, can be formulated and calculated by the proposed system based on the measurements followed by how recommendations can be made from the measurement. By Bayes' theorem, $P(C = 1|S_{i,j})$ can be written as:

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

$P(S_{i,j})$ is the probability that $S_{i,j}$ occurs, while $P(S_{i,j}|C = 1)$ is the probability that $S_{i,j}$ occurs given that two users are a related pair. $P(C = 1)$ is the probability that users i and j are a related pair. By Eq. 4, $P(S_{i,j})$ can be calculated as:

$$P(S_{i,j}) = \gamma_a e^{-\lambda_a S_{i,j}} \tag{8}$$

Similarly, $P(S_{i,j}|C = 1)$ can be calculated as:

$$P(S_{i,j}|C = 1) = \gamma_f e^{-\lambda_f S_{i,j}} \tag{9}$$

and $P(C = 1)$ can be obtained from:

$$P(C = 1) = \frac{N^{(p)}_{C=1}}{N^{(p)}} \tag{10}$$

where $N^{(p)}$ is the total number of possible user pairs and $N^{(p)}_{C=1}$ is the number of related pairs. $N^{(p)}$ can be calculated as:

$$N^{(p)} = \frac{N^{(u)}(N^{(u)} - 1)}{2} \tag{11}$$

where $N^{(u)}$ is the number of users. By putting Eq. 5, Eq. 9, Eq. 10 and Eq.11 into Eq. 7:

$$\begin{aligned} P(C = 1|S_{i,j}) &= \frac{\gamma_f e^{-\lambda_f S_{i,j}}}{\gamma_a e^{-\lambda_a S_{i,j}}} \frac{N^{(p)}_{C=1}}{N^{(p)}} \\ &= \frac{\gamma_f}{\gamma_a} \frac{2N^{(p)}_{C=1}}{N^{(u)}(N^{(u)} - 1)} e^{(\lambda_a - \lambda_f)S_{i,j}} \end{aligned} \tag{12}$$

and Eq. 12 can be written as:

$$P(C = 1|S_{i,j}) = \gamma_t e^{\lambda_t S_{i,j}} \tag{13}$$

where γ_t is equal to $(\gamma_f/\gamma_a)(2N^{(p)}_{C=1}/N^{(u)}(N^{(u)} - 1))$ and λ_t is equal to $\lambda_a - \lambda_f$. Symbols γ_t and λ_t are constants, α_f, α_a, $N^{(u)}$ and $N^{(p)}_{C=1}$ are non-zero positive numbers, and $\gamma_t \geq 0$ is expected. The sign of λ_t changes the distribution of $P(C = 1|S_{i,j})$. As a positive number will make $P(C = 1|S_{i,j})$ become a strictly increasing function, a higher BoFT similarity between a user pair implies a higher probability that the two users are a related pair. Based on the observation in Fig. 10, $\lambda_t = 8.09$ and $\lambda_t = 2.36$ are obtained on Skyrock and 163 Weibo, respectively. It can be concluded that a higher BoFT similarity, $S_{i,j}$ implies a higher $P(C = 1|S_{i,j})$, the probability that the two users, i and j, are a related pair.

With the measurement, follower/followee recommendation can be conducted based on the discovered connections. In follower/followee recommendation, a list of J users, $U_{i,j}$, is recommended to a user, i, given the BoFT similarities of all users and the list of users that are most likely to be related pairs with user i. The problem can be formulated as the following:

$$U^*_{i,J} = \arg\max_{U_{i,J}} P(U_{i,J}|S_{i,1}, ..., S_{i,N^{(u)}}) \tag{14}$$

where $P(U_{i,j}|S_{i,1}, ..., S_{i,N^{(u)}})$ is the probability that all users in $U_{i,j}$ are related pairs with user i, given the $S_{i,j}$ between user i and other users. Using Naive Bayes, Eq. 14 becomes:

$$U^*_{i,J} = \arg\max_{U_{i,J}} \prod_{j=1}^{J} P(C = 1|S_{i,j}),\ where\ j \subset U_{i,J} \tag{15}$$

Based on Eq. 12, $P(C = 1|S_{i,j})$ is a strictly increasing function with respect to $S_{i,j}$, and to find $U^*_{i,J}$ is equivalent to finding the list of J users that have the highest BoFT similarity to user i. Eq. 14 can be rewritten as:

$$U^*_{i,J} = \arg\max_{U_{i,J}} \prod_{j=1}^{J} S_{i,j},\ where\ j \subset U_{i,J} \tag{16}$$

The list of recommended users are those with the higher BoFT similarity to user i, and a recommendation system is built accordingly. The information can be sent to the social media and mobile applications when a list of follower/followee recommendations is needed for a given user.

VI. Experimental Results

This section introduces how the experiment is conducted, followed by the experimental results with two showcases. A discussion of the results concludes this section.

A. Setup

Based on the observation that user pairs with a higher BoFT similarity are more likely to be follower/followee, discovered connections can be evaluated as a follower/followee recommendation system using $S_{i,j}$. Fig. 12 shows the experiment setup for the evaluation with user shared images from Skyrock and 163 Weibo. As in step 1 of Fig. 12 (a), the user shared images are analyzed using BoFT, as in Fig. 3 (a), and users are

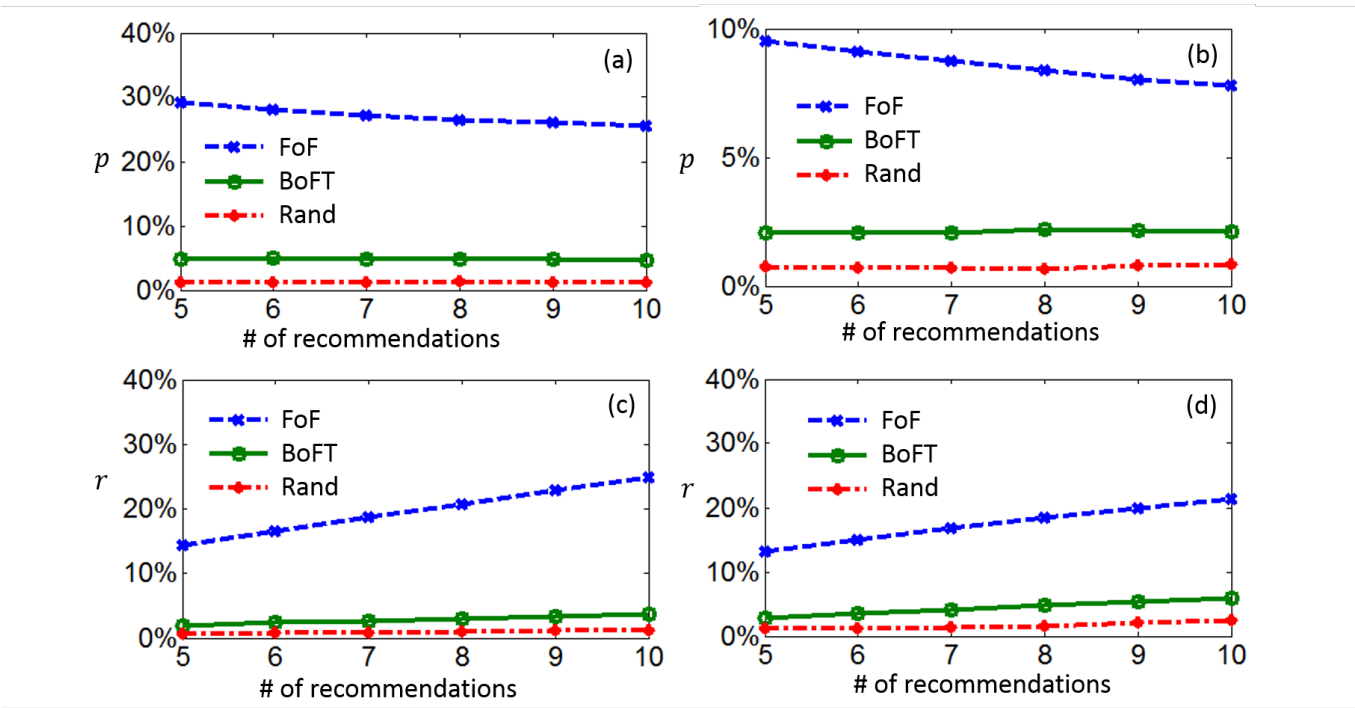

Fig. 13: Performance of follower/followee recommendations with connection discovery, in precision p (upper part) and recall r (lower part): (a) and (b) are the p from 5 to 10, for Skyrock and 163 Weibo, respectively, (c) and (d) shows the r from 5 to 10 for Skyrock and 163 Weibo, respectively.

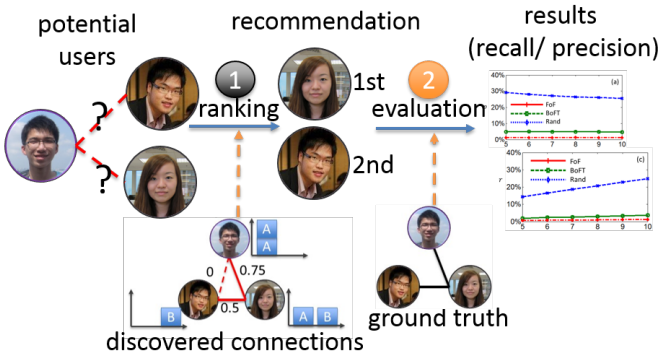

Fig. 12: Two steps in experimental setup: 1) ranking by $S_{i,j}$ of discovered connections, 2) evaluation by ground truth.

represented by user profiles and the distribution of BoFT labels that the user has, as in Fig. 3 (b). Connections are discovered by computing the pairwise $S_{i,j}$ by Eq. 2 with the user profiles. The list of users to be recommended to user i, are ranked by $S_{i,j}$ in the discovered connections. By Eq. 16, the set of J users are most likely to be follower/followees of user i if they are users with the highest similarities. As a result, the set of users with the highest similarities are recommended to user i, and the results are evaluated by two common metrics of prediction performance, recall rate r and precision rate p, for users with the highest similarities, as in step 2 of Fig. 12. The precision rate, p, measures the percentage of discovered follower/followee relationships that exist in the ground truth, while r is the percentage of existing follower/followee relationships in the ground truth that are recommended. A better discovery and recommendation method should give a higher value of p and r.

In order to evaluate the effectiveness of the proposed system, three connection discovery methods for follower/followee recommendation are implemented for comparison. The first method is FoF, which is an achievable upper bound when difficult and limited access SGs are available. The recommendation is from the similarity of the SGs. The second method uses user annotated tags (UserT), and the recommendation is based on the connection discovered based on the similarity among the user annotated tags on shared images between two users. Each user is represented by a user profile of the occurrence of each user annotated tag (152,938 unique tags) used in the dataset. The connection strength in both methods is calculated by cosine similarity, and the same methods are used as in Fig. 12. The third method is a random method (Rand), in which follower/followee relationships are recommended randomly. This serves as a baseline, or the lower bound for the evaluation. The ground truth, the follower/followee relationship, is hidden in that it is not used as an input to the calculation.

B. Results

The number of recommendations is set to be 5 to 10, to simulate a normal recommendation system; however, the same trend can be found even when a smaller or a bigger number of recommendations is used. As the selected users are scraped randomly, the network densities are low, with 1.41% and

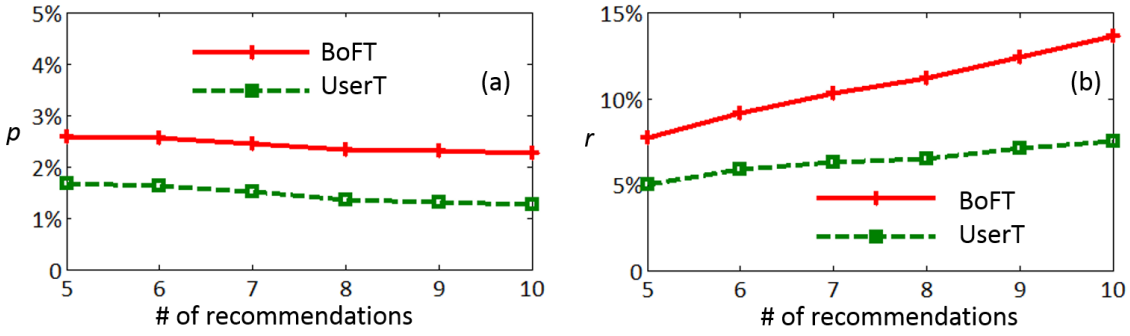

Fig. 14: Follower/followee recommendation results in Flickr with the proposed system and user annotated tags: (a) p, (b) r.

0.71% on Skyrock and 163 Weibo, respectively. The low network densities make the recommendation challenging, with Rand only able to achieve 1.41% and 0.71%. Fig. 13 (a) and (b) show the p of different methods against the number of recommendations, on Skyrock and 163 Weibo, respectively. It is observed that the proposed method is at least 4 times better than Rand, and achieves 25% of the performance of FoF, the achievable bound of the discovery. Fig. 13 (c) and (d) show the r of different methods against the number of recommendations from 5 to 10 on Skyrock and 163 Weibo, respectively. The same observation is found for r. BoFT achieves 20% and 41% of the r value of FoF for Skyrock and 163 Weibo, respectively. The results show that the proposed approach can indeed recommend follower/followees using discovered connections.

Table I shows the runtime of the proposed system. The user shared images are processed by two different calculations. The first one is a Matlab program running on a machine with 8Gb of memory and an i5-4570 CPU, and running feature extraction, codebook generation and feature coding and pooling The runtime is shown in Table I(a). The second one is a cloud-assist calculation with 8 virtual machines, using m3.xlarge on Amazon EC2. It runs clustering, profile learning and similarity calculation, and the runtime is shown in Table I(b). It is observed that feature extraction and clustering spend most of the runtime, of which it takes 583s and 265s for each user on 163 Weibo and Skyrock, respectively. A longer runtime is expected when more users and user shared images are involved, and a big data system is needed to handle the data.

TABLE I: Runtime for BoFT

(a)		
Process	Time (in sec)	
	163 Weibo	Skyrock
Feature Extraction	215,662	130,051
Codebook Generation	43,772	42,351
Feature Coding & Pooling	606	398
(b)		
Clustering	71,723	61,586
Profile Learning	1,038	854
Similarity Calculation	788	1,089
Total	333,589	236,329

C. Showcase 1: User Annotated Tags on Flickr

Tagging is a popular feature on many social networks today, such as Flickr. Flickr is an image oriented social network, that focuses on the sharing of images, and all content shared involves at least one image. Follower/followee relationships can be predicted with the discovered connections from the similarity of the user annotated tags on the shared images between two users. It is interesting to compare the effectiveness of the proposed method on an image oriented social network, using user annotated tags to calculate the user similarity, UserT. A comparison between UserT and the proposed method on Flickr is evaluated with over 201,006 user shared images from 562 users with 902 relationships, as used in [23], and is shown in Fig. 14. Fig. 14 (a) shows p and Fig. 14 (b) shows r for the top 5 to 10 highest similarity users. It is proved that the proposed BoFT method performs better than UserT, meaning it could be a better alternative for today's social networks when an SG is not accessible. The results proved that the proposed method is 65% better than UserT in terms of both p and r on Flickr.

D. Showcase 2: Discovering Gender on Flickr

Profiles on social media are also important for applications but are not always available. Among the information in profiles, gender is interesting, as it is useful for recommendation. Another showcase using the same Flickr dataset is conducted to show how gender can be identified with discovered connections. 445 out of the 562 users provide their gender, and of these there are 79 females and 366 males. In each trial, 50 females and 50 males are selected randomly and $S_{i,j}$ is calculated. The experiment uses a 5-NN approach, in which the gender of the top 5 users, with the highest $S_{i,j}$ with user i, is predicted as the gender of user i, and this is verified by the ground truth. 1000 trials were conducted and the averaged result is shown in Fig. 15. It is observed that gender identification is possible with the discovered connections, being 22% better than that random guessing. This showcase proved that the proposed method can be an alternative to SGs.

E. Discussions

This paper has successfully proved and characterized the phenomenon that related pairs share similar images by mass

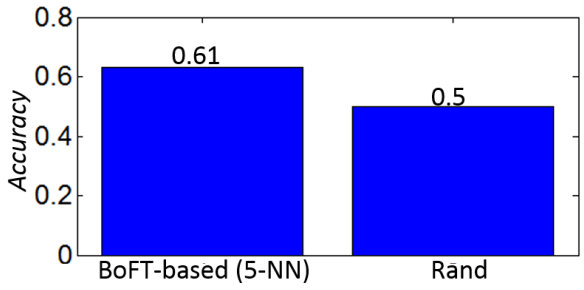

Fig. 15: Gender discovery results: (a) measurement on $S_{i,j}$, (b) discovery using 5-NN with $S_{i,j}$.

user shared images from real-world social networks, and then formulated and developed the results into practical methods to discover user connections. There are other methods, such as that in [13], which proved that using a completed SG performs better than FoF. However, it would be more useful to compare the proposed method with other inputs under the same experimental procedures to prove the possibility of using user shared images for recommendation. Further investigation will be needed to maximize the effectiveness for certain applications, such as recommendation and gender identification, of the proposed method, for instance, by giving a higher weight to a less frequent BoFT label.

There are two possible directions for improving the precision. As the features extracted from images are a low level description, even 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 distribution of HSV color, or other feature extraction techniques, such as GIFT. Another highlighted direction is the value of K, the number of labels. As discussed previously, the number of labels, K, or the number of clusters in clustering, has to be pre-defined. A too small value could make two dissimilar images be annotated with the same label, while a too large number will cause two similar images with different labels. Strategies such as that in [39] combine the advantages of on-line clustering [40] and mean-shift [41] in an under sampling framework [42]. The method does not require knowing k in advance, and performs better than k-means in image categorization [39]. Those directions can help to improve the results of connection discovery. With the connections among users, many interesting applications, such as centrality analysis, recommendation, virality prediction, and many other applications [43][44][45][46] become possible.

VII. Conclusion and Future Work

This work has proposed a connection discovery method and system for follower/followee recommendation based on user shared images. A practical method, BoFT, is discussed to label user shared images with BoFT labels on over 360,000 user shared images. The characteristics of user shared images are then investigated and modeled as exponential distributions based on the analysis of 3 million follower/followee relationships from two social networks with different origins, Skyrock and 163 Weibo, for which similar observations are found. Based on the observations, a practical follower/followee recommendation system is proposed and formulated with the discovered connections, which are extensively verified with ground truth. It is concluded that follower/followee recommendation using discovered connections by user shared images is possible, and the recommendation is 60% better than UserT and achieves 25% of the performance of FoF, a method used when limited access SGs are available. The discovered connections are also proven to be able to identify user gender. These findings create a potential long term impact and contribution to scientific research and commercial applications, especially when access to SGs is difficult or limited. This work enables the use of social network analysis on any social media with image sharing mechanisms, for which many interesting applications, such as centrality analysis, recommendation, virality prediction, and many other applications become possible.

Acknowledgment

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

References

[1] J. Chen, W. Geyer, C. Dugan, M. Muller and I. Guy, "Make new friends, but keep the old: Recommending people on social networking sites," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 201-210 (2009).

[2] V. Leroy, B. B. Cambazoglu and F. Bonchi, "Cold start link prediction," in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 393-402 (2010).

[3] I. Kahanda and J. Neville, "Using transactional information to predict link strength in online social networks." in Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media, pp. 74-81 (2009).

[4] L. Backstrom and J. Leskovec, "Supervised random walks: Predicting and recommending links in social networks," in Proceedings of the 4th ACM International Conference on Web Search and Data Mining, pp. 635-644 (2011).

[5] D. Krackhardt and M. Kilduff, "Whether close or far: Social distance effects on perceived balance in friendship networks." Journal of Personality and Social Psychology, vol. 76, pp. 770 (1999).

[6] J. Leskovec, D. Huttenlocher and J. Kleinberg, "Predicting positive and negative links in online social networks," in Proceedings of the 19th International Conference on World Wide Web, pp. 641-650 (2010).

[7] Y. Zheng, L. Zhang, Z. Ma, X. Xie and W. Ma, "Recommending friends and locations based on individual location history," ACM Transactions on the Web (TWEB), vol. 5, pp. 5 (2011).

[8] A. Rae, B. Sigurbjrnsson and R. van Zwol, "Improving tag recommendation using social networks," In Proceedings of the International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information, pp. 92-99 (2010).

[9] R. Ottoni, J. P. Pesce, D. B. Las Casas, G. Franciscani Jr, W. Meira Jr, P. Kumaraguru and V. Almeida, "Ladies first: Analyzing gender roles and behaviors in pinterest." in Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, (2013)

[10] L. Kennedy, M. Naaman, S. Ahern, R. Nair and T. Rattenbury, "How flickr helps us make sense of the world: Context and content in community-contributed media collections," in Proceedings of the 15th International Conference on Multimedia, pp. 631-640 (2007).

[11] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel and B. Bhattacharjee, "Measurement and analysis of online social networks," in Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, pp. 29-42 (2007).

[12] Jin, Emily M., Michelle Girvan, and Mark EJ Newman. "Structure of growing social networks." Physical review E 64.4 (2001): 046132.

[13] L. L and T. Zhou, "Link prediction in complex networks: A survey," Physica A: Statistical Mechanics and its Applications, vol. 390, pp. 1150-1170 (2011).

[14] I. Guy, N. Zwerdling, I. Ronen, D. Carmel and E. Uziel, "Social media recommendation based on people and tags," in Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 194-201 (2010).
[15] I. Guy, N. Zwerdling, D. Carmel, I. Ronen, E. Uziel, S. Yogev and S. Ofek-Koifman, "Personalized recommendation of social software items based on social relations," in Proceedings of the 3rd ACM Conference on Recommender Systems, pp. 53-60 (2009).
[16] R. Parimi and D. Caragea, "Predicting friendship links in social networks using a topic modeling approach," in Advances in Knowledge Discovery and Data Mining, Springer, pp. 75-86 (2011).
[17] W. H. Hsu, A. L. King, M. S. Paradesi, T. Pydimarri and T. Weninger, "Collaborative and structural recommendation of friends using weblog-based social network analysis." in AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp. 55-60 (2006).
[18] X. Xie, "Potential friend recommendation in online social network," in IEEE International Conference on Cyber, Physical and Social Computing (CPSCom), pp. 831-835 (2010).
[19] Chow, Wing S., and Lai Sheung Chan. "Social network, social trust and shared goals in organizational knowledge sharing." Information & Management 45.7: pp. 458-465 (2008).
[20] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 211-220 (2009).
[21] S. A. Golder and S. Yardi, "Structural predictors of tie formation in twitter: Transitivity and mutuality," in Social Computing (SocialCom), 2010 IEEE Second International Conference on, pp. 88-95 (2010).
[22] S. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng and H. Zha, "Like like alike: Joint friendship and interest propagation in social networks," in Proceedings of the 20th International Conference on World Wide Web, pp. 537-546 (2011).
[23] M. Cheung and J. She. "Bag-of-features tagging approach for a better recommendation with social big data." in Proceedings of the 5th International Conference on Advances in Information Mining and Management. (2014).
[24] H. Kim, J. Jung and A. El Saddik, "Associative face co-occurrence networks for recommending friends in social networks," in Proceedings of 2nd ACM SIGMM Workshop on Social Media, pp. 27-32 (2010).
[25] X. Li, L. Guo and Y. E. Zhao, "Tag-based social interest discovery," in Proceedings of the 17th International Conference on World Wide Web, pp. 675-684 (2008).
[26] T. C. Zhou, H. Ma, M. R. Lyu and I. King, "UserRec: A user recommendation framework in social tagging systems." in Proceedings of the 24th AAAI Conference on Artificial Intelligence, (2010).
[27] A. Shepitsen, J. Gemmell, B. Mobasher and R. Burke, "Personalized recommendation in social tagging systems using hierarchical clustering," in Proceedings of the 2008 ACM Conference on Recommender Systems, pp. 259-266 (2008).
[28] X. Zhang, X. Zhao, Z. Li, J. Xia, R. Jain and W. Chao, "Social image tagging using graph-based reinforcement on multi-type interrelated objects," in Signal Process, vol. 93, no. 8, pp. 21782189 (2013).
[29] X. Zhang, Z. Li and W. Chao, "Tagging image by merging multiple features in a integrated manner," in Journal of Intelligent Information Systems, vol. 39, pp. 87-107 (2012).
[30] B. Sigurbjrnsson and R. Van Zwol, "Flickr tag recommendation based on collective knowledge," in Proceedings of the 17th International Conference on World Wide Web, pp. 327-336 (2008).
[31] J. Sang, C. Xu and J. Liu, "User-aware image tag refinement via ternary semantic analysis," IEEE Transactions on Multimedia, vol. 14, no. 3, pp. 883895 (2012).
[32] E. Moxley, J. Kleban, J. Xu and B. Manjunath, "Not all tags are created equal: Learning flickr tag semantics for global annotation," in Proceedings of IEEE International Conference on Multimedia & Expo, pp. 1452-1455 (2009).
[33] Z. Wang, L. Sun, W. Zhu, S. Yang, H. Li and D. Wu, "Joint Social and Content Recommendation for User-Generated Videos in Online Social Network," IEEE Transactions on Multimedia, Vol.15, no. 3, pp. 698709 (2013).
[34] T. Kadir and M. Brady, "Saliency, scale and image description," International Journal of Computer Vision, vol. 45, pp. 83-105 (2001).
[35] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, pp. 91-110 (2004).
[36] A. McCallum, K. Nigam and L. H. Ungar, "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, pp. 169-178 (2000).
[37] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," IEEE Transactions on Communications, vol. 28, pp. 84-95 (1980).
[38] Z. Jie, M. Cheung and J. She, "A cloud-assisted framework for bag-of-features tagging in social networks" in IEEE 4th Symposium on Network Cloud Computing and Applications, (2015).
[39] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in 10th IEEE International Conference on Computer Vision, 2005. ICCV 2005, pp. 604-610 (2005).
[40] A. Meyerson, "Online facility location," in Proceedings. 42nd IEEE Symposium on Foundations of Computer Science, 2001. pp. 426-431 (2001).
[41] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pp. 603-619 (2002).
[42] A. Estabrooks, T. Jo and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," Computing Intelligent, vol. 20, pp. 18-36 (2004).
[43] L. Weng, A. Flammini, A. Vespignani and F. Menczer, "Competition among memes in a world with limited attention," Scientific Reports, vol. 2, (2012).
[44] Freeman, Linton C. "Centrality in social networks conceptual clarification." in Social networks, Vol. 1, Issue 3, pp. 215-239 (1979).
[45] F. E. Walter, S. Battiston and F. Schweitzer, "A model of a trust-based recommendation system on a social network," in Autonomous Agents and Multi-Agent Systems, vol. 16, no. 1, pp. 57-74 (2008).
[46] I. Konstas, V. Stathopoulos, and J. M. Jose. "On social networks and collaborative recommendation." Proceedings of the 32nd international ACM SIGIR conference on Research and Development in Information Retrieval. ACM, (2009).

Ming Cheung was born in Hong Kong. He received his B.Eng. and M.Phil in Electronic and Computer Engineering at Hong Kong University of Science and Technology (HKUST) in 2010 and 2012 respectively. He joined the HKUST-NIE Social Media Lab, Asia's first social media lab, in 2012 as a research assistant, and currently is a Ph.D. candidate at HKUST. His research interests include social media analytics, information diffusions and user behavior predictions.

James She is an assistant professor in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology (HKUST), and a visiting research fellow at the University of Cambridge. He is also the founding director of Asia's first social media lab, HKUST-NIE Social Media Lab, and spearheads multidisciplinary research and innovation in cyber-physical social media systems, viral media analytics and mobile media broadcast systems. Celebrated as a thought leader in new media and emerging cyber-physical societies, James is a member of the World Economic Forum's Global Agenda Council (Social Media) and joins other government and business leaders to develop solutions to key social media issues on the global agenda.

Zhanming Jie was born in Guangdong, China. He received his B.Eng. in School of Automation Engineering at University of Electronic Science and Technology of China in 2014. During his final-year undergraduate period, he worked at Singapore University of Technology and Design as an intern. He is currently a Ph.D student at HKUST-Nie Social Media Lab. His research interest includes machine learning, social network and recommender system.