9-2014

A study of age gaps between online friends

Lizi LIAO
Singapore Management University, lzliao@smu.edu.sg

Jing JIANG
Singapore Management University, jingjiang@smu.edu.sg

Ee Peng LIM
Singapore Management University, eplim@smu.edu.sg

Heyan HUANG

Citation

LIAO, Lizi; JIANG, Jing; LIM, Ee Peng; and HUANG, Heyan. A study of age gaps between online friends. (2014). *HT'14: Proceedings of the 25th ACM Conference on Hypertext and Social Media: September 1-4, 2014, Santiago, Chile*. 98-106.
Available at: https://ink.library.smu.edu.sg/sis_research/2416

A Study of Age Gaps between Online Friends

Lizi Liao
School of Computer Science
Beijing Institute of Technology
liaolizi.llz@gmail.com

Jing Jiang
School of Information Systems
Singapore Management
University
jingjiang@smu.edu.sg

Ee-Peng Lim
School of Information Systems
Singapore Management
University
eplim@smu.edu.sg

Heyan Huang
School of Computer Science
Beijing Institute of Technology
hhy63@bit.edu.cn

ABSTRACT

User attribute extraction on social media has gain considerable attention, while existing methods are mostly supervised which suffer great difficulty in insufficient gold standard data. In this paper, we validate a strong hypothesis based on homophily and adapt it to ensure the certainty of user attribute we extracted via weakly supervised propagation. Homophily, the theory which states that people who are similar tend to become friends, has been well studied in the setting of online social networks. When we focus on age attribute, based on this theory, online friends tend to have similar age. In this work, we take a step further and study the hypothesis that the age gap between online friends become even smaller in a larger friendship clique. We empirically validate our hypothesis using two real social network data sets. We further design a propagation-based algorithm to predict online users' age, leveraging the clique-based hypothesis. We find that our algorithm can outperform several baselines. We believe that this method could work as a way to enrich sparse data and the hypothesis we validated would shed light on exploring the proximity of other user attributes such as education as well.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Algorithms, Experimentation

Keywords

Social Network Analysis; Age Prediction; Homophily

1. INTRODUCTION

With the fast adoption of social media, more and more people have moved their social activities online. Large online social networks such as Facebook and Twitter allow users to make friends and form communities beyond the physical boundaries that offline social networks have. To better understand these online social networks, there have been many studies on online user behaviors and properties of online communities, some relating to hypotheses and theories developed from offline social networks. In particular, researchers have studied homophily [17] in the online setting [18, 24]. Homophily is the theory that people similar to each other tend to become friends, or in other words, "birds of a feather flock together." Here similarity between people may be based on various attributes including location, age, education, social status, interest, etc. Researchers have studied to what extent this theory is true in online social networks and whether this theory can be exploited for prediction tasks [15, 4, 22, 19, 1].

In this paper, we are interested in the particular attribute of age of online users and the age gaps between online friends. Based on the notion of homophily, we expect that users of similar age are more likely to become friends than users with a larger age gap. If this hypothesis is true, then presumably we can make use of the friendship links in online social networks and a small number of users' age information to predict other users' age. This age prediction task can be useful for many applications such as user profiling and targeted advertising, especially considering that age information is often unknown for many online users.

While it is not new to leverage the theory of homophily to infer user attributes, including age, in online social networks [18, 24], in this work we focus on a stronger hypothesis than the general notion of homophily. With the huge amount of link information provided by social media, we could obtain useful information about a user's age via his or her friends. At the same time, by observing links between

HT'14, September 1–4, 2014, Santiago, Chile.
Copyright 2014 ACM 978-1-4503-2954-5/14/09 ...$15.00.
http://dx.doi.org/10.1145/2631775.2631800.

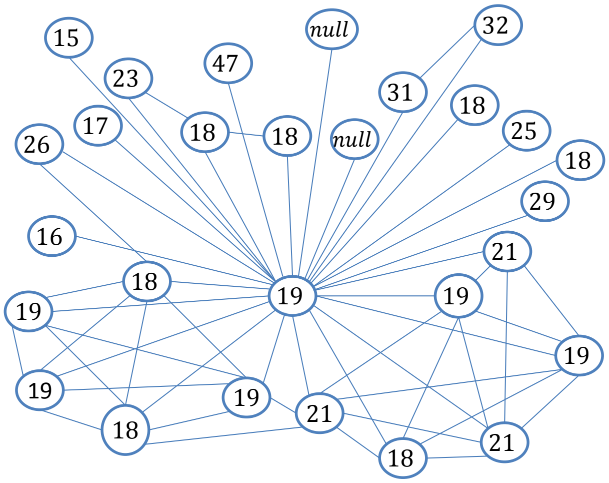

Figure 1: An example of an ego network with age information. There are some public accounts without age information denoted as *null*.

his or her friends, we could further infer the certainty of the information we get. In fact, the friendship links among a user's friends provide latent but precious information (see Figure 1). We hypothesize that users who form a large clique in an online social network (i.e. users who are pairwise friends with each other) are more likely to have similar ages. Using data from two real online social networks, we validate our hypothesis based on various statistics gathered from the data sets. We further propose a maximal clique based age propagation algorithm to predict users' age. Our algorithm is based purely on the topology of the social network and a small percentage of users' age; it does not use any other information such as users' profile, online behavior or user-generated content. We find that our algorithm can perform better age prediction compared with baselines that use random prediction or only immediate friends' age for prediction. On a Twitter network with over 25K of users, by observing about 10% users' age, our algorithm can predict the other users' age (within an error gap of 5) with an accuracy of over 79%. We expect that our clique-based hypothesis can be combined with other age prediction methods to further improve the accuracy of predicting online users' age.

Our work has the following contributions: (1) We propose a new hypothesis that the age gap between online friends is smaller in larger cliques and empirically validate this hypothesis. (2) We design a scalable algorithm to predict a user's age using maximal cliques and label propagation, which could help with the data sparsity problem. (3) We empirically evaluate our algorithm and show that the performance is better than several baselines. (4) Our method could work as a way to enrich sparse data and the hypothesis we validate would shed light on exploring the proximity of other user attributes such as education.

The rest of the paper is organized as follows. We first discuss some related work in Section 2 and provide information about our MG and Twitter data sets in Section 3. We explore the property of homophily in Section 4. In Section 5, we take a step further to validate out stronger hypothesis for the relation between age gap and clique size. The formulation of our task and details of our clique age propagation algorithm are shown in Section 6, followed by the evaluation of our algorithm in Section 7. Finally our work is concluded in Section 8.

2. RELATED WORK

As we aim to profile a user's age attribute accurately via heuristics based on homophily, our work is related to homophily in social networks, user attributes inference and age prediction. We briefly summarize related research below.

2.1 Homophily in Social Networks

The hypothesis that people similar to each other tend to become friends dates back to at least the 70s in the last century. In social science, there is a general expectation that individuals develop friendships with others of approximately the same age [23]. In [16] the authors study the interconnectedness between homogeneous composition of groups and the emergence of homophily. In more recent years, the authors of [12] investigate the origins of homophily in a large university community, using network data in which interactions, attributes and affiliations are all recorded over time. In [8] the authors try to find the role of homophily in online dating choices made by users. Given attributes of some fraction of the users in an online social network, [19] infers the attributes of the remaining users. In [1] the authors leverage the principle of homophily to the inference of three attributes: gender, political orientation and age.

2.2 Inferring User Attributes

Inferring online users' attributes such as location and age has been studied extensively in recent years. Online social networks like Twitter have provided abundant resources. By learning distinguishing attributes of certain classes of users through third-person text, [5] aims to classify users in the analysis of first-person communication. The attribute extraction method is based on [2]. Using the networks and cities of US LiveJournal members, [15] finds that the likelihood of friendship is almost inversely proportional to distance of location. Based on the assumption that people tend to make friends with those having similar geographical location attributes, [4] observes and measures the relationship between geography proximity and friendship on Facebook. In [22] the authors solve two intimately related tasks for online social networks: link and location prediction.

2.3 Age Prediction

There has long been interests concerned with how various morphological, phonological and stylistic aspects of language can vary with a person' age. Early work has an emphasis on predicting author properties based on the usage of function words, parts-of-speech, punctuation and some spelling/grammatical errors [11]. Recently, researchers have focused less on the sociolinguistic implications and more on the tasks themselves, which leads to classifiers with feature representations capturing content in addition to style. These features include function/content words, word classes [13], content word classes [3] and unigrams [20]. There are also applications of simple classifiers to map a sequence of queries into the gender and age of the user issuing the queries [9]. In [21], stacked-SVM-based classification algorithms over a rich set of features are applied to classify several user attributes, such as gender, age, regional origin and political orientation.

However, none of these studies has leveraged information of associations between social network users to predict user age. In fact, social network information has been widely used in tasks like location inference in social media platforms. Sadilek et al. estimate a user's future location through the locations of users in his or her ego network [22]. Their approach requires both users' locations to be known in order to estimate the social relationship. Davis Jr et al. [7] use a user's Twitter follower network to do this task. Although their approach is based on location information in an individual's ego network, it uses location names only and their approach is non-iterative. Backstrom et al. [4] propose a location inference method for the Facebook social network using probabilistic inference to select the location from a user's friends. Our algorithm also predicts users' ages through ages of others in their ego network. However, we incorporate clique heuristics into our proposed propagation algorithm and the algorithm is iterative. Our algorithm is efficient which can run in seconds, and only a small number of initial ages are needed.

3. DATA SETS

In order to study the age gaps between online friends, we need online social networks with users' age information. Our data come from two sources: (1) MG[1], a social network platform for mobile users, and (2) Twitter.

MG is a social network platform developed by an Internet advertising company targeted at mobile users. The platform has attracted millions of users from over 100 countries. The platform allows users to establish friendship links with other users in the network and engage in online as well as offline conversations using mobile phones, in a way similar to the well-known service WhatsApp Messenger. When users sign up for the service, they self-report various personal information such as gender, age and country.

Country	USA	Australia	India	Singapore
all users	173,845	19,116	1,721,295	57,186
users with age	91,094	10,223	805,738	29,279
friend links	107,234	9,604	6,099,372	156,346

Table 1: Statistics of the MG data.

We collected the data from MG before April 2012. This subset of data contains over 6 million users, among which 43.7% specified their date of birth in their profiles. We observe that most users befriend with people from the same country, which is not surprising. To simplify our analysis, we further chose only users from the following four countries to form four subsets of data: USA, Australia, India and Singapore. These countries are among the top-ranked countries in terms of number of subscribed users.

As links in the MG network are mutually established—a friendship invitation has to be verified by the other party first before a link is established—we can directly treat the existing links in the data as friendship links. As for the age information, because age has been self-reported by a large proportion of users, we use the age information of these users as ground truth. We still keep the other users in the network as they can be used for age propagation in our age prediction algorithm later. Some statistics of the data are given in Table 1. We can see that around half of the users chose to input their age information. To get an idea of the age distribution of these users, in Figure 2 we plot the number of users for each age ranging from 10 to 60. We ignore users who are below 10 or above 60. The number of users is plotted in log scale as there is much difference between the numbers of users in different countries and with different ages. From the figure we can see that most MG users are in their 20s or early 30s. The distributions of the different countries are similar.

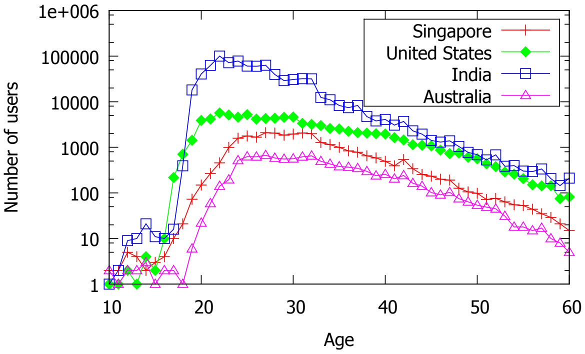

Figure 2: Age distribution of four representative countries in the MG data set.

Twitter is one of the most popular social media platforms for users to post short messages in real time, which are known as "tweets." Twitter users can "follow" other users, i.e. to automatically receive all the tweets published by those users. These following relations are one-directional and usually do not indicate friendships. However, when two users mutually follow each other, there is a strong indication that they are interested in each other's posts and we can loosely regard them as friends [10]. When signing up for the service, Twitter users have the option to reveal their age in their profiles, but most users do not explicitly specify their age or date of birth, making it hard to directly obtain the age information of users. Additionally, Twitter also provides a mechanism for users to specifically reference other users by their usernames (`@username`) in tweets, which we call "at-mentions." Below we will show how we exploit at-mentions to infer users' age.

We collected a Twitter data set as follows. Starting from a set of 59 seed users in Singapore, we first crawled these users' direct followers and followees and then crawled their followers/followees' followers and followees, i.e. we crawled all users who are either one or two hop(s) away from the seed users. Using features derived from [14] on Twitter bot detection, we filtered out potential spammers, promoters and other automated Twitter accounts so that the remaining data consist primarily of "regular" Twitter users. After these preprocessing steps, we were left with 25,703 users.

We created a friendship link between two users if they mutually follow each other. For age information, we employed the following two strategies to obtain the ground truth. (1) For those users who mentioned either their age or date of birth in their short profile biographies, we used a set of patterns to extract such information. In this way, we obtained

[1] The social network is anonymized due to an NDA with the company.

more than 700 users with age information after manual correction. (2) Inspired by [24], we observe that many tweets contain the pattern "happy X-th birthday" (where X is a number) together with an at-mention. By extracting these expressions, we can infer the age of the users who were mentioned. With this strategy applied to tweets within a one year span, we were able to obtain the age of a little more than 3000 users, and we found the accuracy of this strategy to be around 87% based on a manual inspection on 200 sample users. For those 87% sample users, the age extracted indeed indicates an age after the human annotator read the corresponding tweet. While there is no way for us to verify whether this age is true because people can always lie in social media, we consider it to be correct. Figure 3 shows the age distribution of all those users with age information. We can see that Twitter has a younger user base with many users between 15 and 20. The figure peaks at age 18. This may come from two possible reasons: (1) Twitter has a younger user base in general. (2) Eighteen is probably a special age that indicates that the person has become an adult, and therefore we see more 18th birthday greetings on Twitter. We do not know which reason dominates, and we are aware that the age labels obtained this way contain bias.

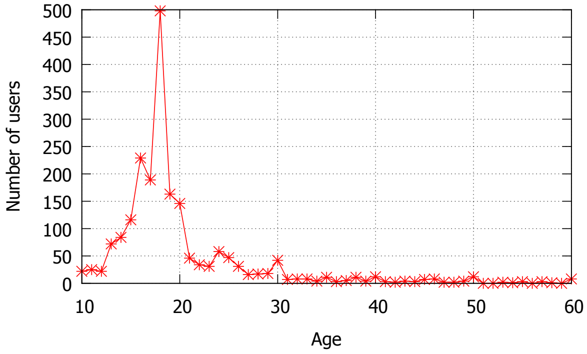

Figure 3: Age distribution of the Twitter data set.

4. AGE GAP BETWEEN ONLINE FRIENDS

In this section we empirically validate the hypothesis that online friends tend to have similar age. Since the age information in the Twitter data set is very sparse, we only plot the number of friendship links versus age gap for the MG data set. It is shown in Figure 4. Here the x-axis is the absolute value of the age difference between a pair of linked users (i.e. friends) and the y-axis is the number of linked user pairs (i.e. friendship links) with that age gap. We can see that most links have a relatively small gap, which demonstrates that users of similar age are more likely to become friends than users with a larger age gap. Later in our experiments we will see that age prediction based on a randomly selected friend's age can achieve a reasonable performance already. However, for a randomly picked age, we could not be sure about its certainty. Furthermore, from Figure 4, we can see that there is still a significant percentage of friendship links with a relatively large age gap. For example, 32.2% of the friendship links have an age gap of 5 or above. If we perform propagation-based age prediction, these links will likely deteriorate the performance.

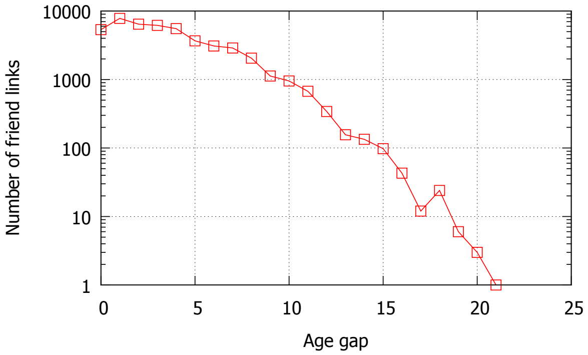

Figure 4: Number of friendship links vs. age gap in the MG data set.

5. AGE GAP IN CLIQUES

As demonstrated in the previous section, there are more friendship links with smaller age gaps than larger age gaps, but the number of friendship links with relatively large age gaps is still significant and should not be ignored totally. However, we hypothesize that chances that several people with very different ages forming a friendship clique are small. Before turning to the validation of our hypothesis, we first revisit a few key concepts in graph theory. Figure 5 gives an simple example. The node in the middle is the ego and the other nodes are her friends. In the right hand side of Figure 5 we have highlighted 3 maximum cliques with a clique size of at least 3. There are 3 other maximum cliques of size 2 that are not highlighted.

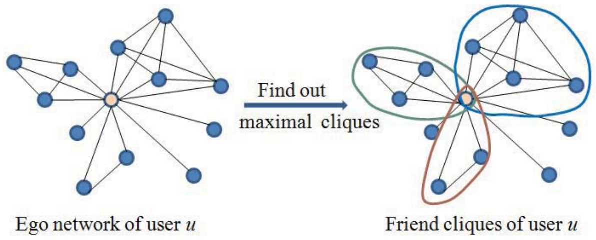

Figure 5: An example of an ego network and the maximum cliques in it.

Definition 1 (Clique). *In an undirected graph, a clique $\mathcal{C}$ is a subset of the nodes in the graph such that for every two nodes in $\mathcal{C}$ there exists an edge in the graph that connects the two nodes. The subgraph induced by $\mathcal{C}$ is complete.*

Definition 2 (Maximal Clique). *In an undirected graph, a maximal clique is a clique that is not contained in any other clique in the graph.*

Definition 3 (Maximal Clique Size). *The clique size of the maximal clique $\mathcal{C}$, denoted as $|\mathcal{C}|$, is the number of nodes in this clique.*

Our hypothesis can be stated as follows:

> *When several users in a social network form a clique, they tend to have a small age gap. The larger the clique size is, the smaller the age gap is between users in the clique.*

The hypothesis stated above is intuitive. When users form a clique in a social network, they have stronger ties among themselves and are more likely to be similar to each other. For example, a group of classmates are likely to form a clique in a social network, and classmates usually have the same age. For the typical example shown in Figure 1, there are 2 large cliques with a size of 6. The left one is a clique of schoolmates in high school. Apparently they are about the same age. The right one is a clique of college friends. There are some age differences in it but overall the ages are still very close. We also observe that there are quite a lot of smaller cliques in user's ego network. The age gaps within those small cliques are more randomly ranged. Most large gaps occur in those small maximal cliques with a size of 2 or 3. This makes sense, since user might become friends with some random people online while the chances of many random people become pair-wise friends with each other are relatively low.

This is a stronger hypothesis than the one we tested in the previous section. Essentially in the previous section we only looked at all cliques (not necessarily maximal cliques) of size 2. With this new hypothesis, we expect to see the age gap to decrease when clique size increases.

To validate this stronger hypothesis, we processed our data in the following way to obtain some useful statistics. Our main idea is to check for each user whether her age difference from friends in a large maximal clique is generally smaller than her age difference from friends in a smaller maximal clique. To do so, first, we found all maximal cliques from our data sets. Figure 6 shows the numbers of maximal cliques of different sizes in the MG data set. We can see that as expected the number of maximal cliques decreases as the clique size goes up.

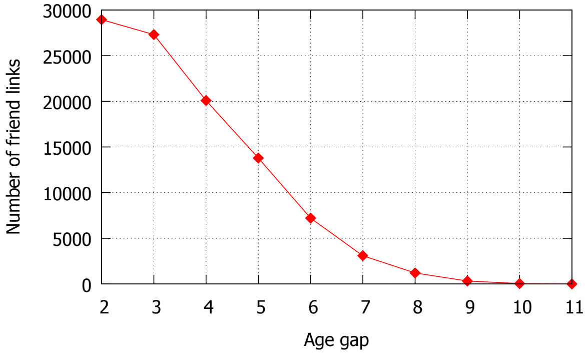

Figure 6: Maximal clique frequency distribution.

We then calculated a measure which we call MAG (mean age gap) for each user with respect to each maximal clique that contains this user. MAG is defined as follows:

$$MAG(u,\mathcal{C}) = \frac{\sum_{u'\in\mathcal{C}\setminus u}|age(u)-age(u')|}{|\mathcal{C}|-1}, \qquad (1)$$

where $\mathcal{C}\setminus u$ is the set of users from $\mathcal{C}$ excluding u, $age(u)$ is the age of user u, and $|\mathcal{C}|$ is the size of $\mathcal{C}$. Essentially $MAG(u,\mathcal{C})$ is the average age gap between u and all other users in $\mathcal{C}$. Our hypothesis is that if $\mathcal{C}$ is large, then this average age gap is small.

Since a user may be inside more than one maximal clique in the social network, we further define $MAG^{(n)}(u)$ as follows:

$$MAG^{(n)}(u) = \frac{1}{|\mathcal{S}_u^{(n)}|}\sum_{\mathcal{C}\in\mathcal{S}_u^{(n)}} MAG(u,\mathcal{C}), \qquad (2)$$

where $\mathcal{S}_u^{(n)} = \{\mathcal{C} : |\mathcal{C}| = n \text{ and } u \in \mathcal{C}\}$. Basically $\mathcal{S}_u^{(n)}$ is the set of maximal cliques of size n which contain u, and $MAG^{(n)}(u)$ is the average of $MAG(u,\mathcal{C})$ over all cliques $\mathcal{C}$ of size n which contain u.

Finally, we define $MAG^{(n)}$ to be the average of $MAG^{(n)}(u)$ over all users who are inside at least one maximal clique of size n. We expect that $MAG^{(n)}$ becomes smaller when n becomes larger, i.e. the average age gap in larger maximal cliques tends to be smaller.

We observe that sometimes the extreme age values in a maximal clique may be outliers. To alleviate the impact of these extreme values, we follow the practice of trimmed estimators in statistics and consider three trimmed versions of $MAG^{(n)}$. Specifically, $MAG^{(n)}_{\neg\text{min}}$ is the version where when we compute $MAG(u,\mathcal{C})$ we exclude the friend of u in $\mathcal{C}$ with the minimum age. $MAG^{(n)}_{\neg\text{max}}$ is defined similarly. $MAG^{(n)}_{\neg\text{min, max}}$ is the version where both the minimum and the maximum ages are excluded when computing $MAG(u,\mathcal{C})$.

Given the definition of $MAG^{(n)}$ above, we can plot the values of $MAG^{(n)}$, $MAG^{(n)}_{\neg\text{min}}$, $MAG^{(n)}_{\neg\text{max}}$ and $MAG^{(n)}_{\neg\text{min, max}}$ against n. We use the Singapore users from the MG data set to plot these curves. Specifically, we use only maximal cliques in which all users' age values are known. The plots are shown in Figure 7. We can see that indeed as the clique size goes up, the mean age gap decreases. When the clique size is 7 and above, the mean age gap is below 5. This empirical analysis gives us the basis for the age prediction algorithm that we will present in the next section.

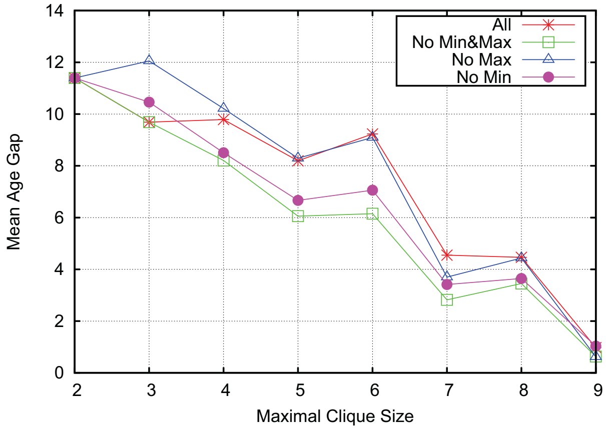

Figure 7: Relationship between mean age gap and maximal clique size.

6. AGE PREDICTION

In the previous section we empirically showed that in large cliques in online social networks, we observe users with small age gaps. This observation inspired our idea of using online cliques to help predict users' age with higher confidence. The assumption is that if two users are friends and they are inside a large clique in the online social network, we can use one user's age to infer the other's age.

First of all, we need a scalable algorithm to find cliques in a large undirected graph. Second, verified age information is still sparse in many social networks such as Twitter, which means prediction based on immediate friends' age would have a low coverage. Third, a user may be inside multiple different maximal cliques, and how to make use of these multiple cliques together to infer this user's age is not clear.

We address the three concerns above in the following way. To find maximal cliques, we make use of the *Bron-Kerbosch* algorithm [6], which is a recursive backtracking algorithm. To tackle the data sparseness problem, we allow predictions using multiple hops of friendship links. And finally since a user can be inside multiple different maximal cliques, we propose an edge weighting function related to clique size to make use of larger maximal cliques.

6.1 Problem Statement and Solution Overview

Let us use $\mathcal{L} = \{(u_1, a_1), \ldots, (u_M, a_M)\}$ to denote a set of labeled users, i.e. users with known age. Here u_i is a user and a_i is her age. We use $\mathcal{U} = \{u_{M+1}, \ldots, u_{M+N}\}$ to denote the unlabeled users. Our goal is to predict the age of the users in $\mathcal{U}$ using $\mathcal{L}$.

We now present our maximal clique based age propagation (MCAP) algorithm. An overview of our algorithm is shown in Figure 8.

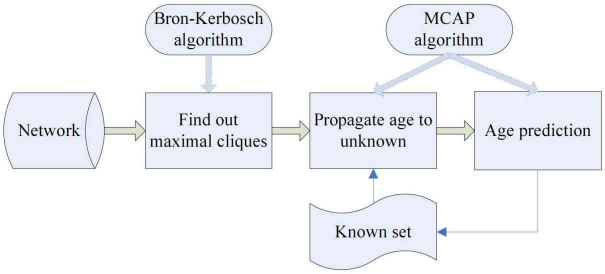

Figure 8: The framework of our age prediction algorithm.

6.2 The MCAP Algorithm

In this section we formally present the maximal clique based age propagation algorithm. We assume that we are able to find all the maximal cliques in a social network. The details of how to find maximal cliques will be given in the next subsection.

The MCAP algorithm is a label propagation algorithm. Label propagation is a type of semi-supervised and iterative algorithms designed to infer labels for items connected in a network [25]. Usually, only a small number of items in the network has known labels, which serve as a source of ground truth information for the estimation of other nodes' labels. Label propagation algorithms usually proceed iteratively where in each iteration some items with unknown labels receive predicted labels based on their neighbors.

Based on our validated hypothesis, we want users who are located in large cliques to have small age gaps. With this goal in mind, in the social network we have, we set the weight between two nodes based on the maximal cliques we have found in the network. Specifically, for two connected users u_i and u_j in the network, we define $w_{i,j}$, the weight for the edge between these two users, to be the size of the largest maximal clique that contains u_i and u_j. Let $\mathcal{N}(u)$ denote the set of neighbors of user u. We now define a propagation probability from user u_i to user u_j as follows:

$$p(i \to j) = \frac{w_{ij}}{\sum_{u_k \in \mathcal{N}(u_i)} w_{ik}} \quad (3)$$

Let $\boldsymbol{p}$ denote the propagation probabilities as defined above for all pairs of connected nodes. We will use these probabilities to propagate the age information.

Algorithm 1 The MCAP algorithm.

```
MCAP($\mathcal{L}$,$\mathcal{U}$,$\boldsymbol{p}$)
Input:
$\mathcal{L}$: A set of users with known age
$\mathcal{U}$: A set of users with unknown age
$\boldsymbol{p}$: The propagation probabilities, where $p(i \to j)$ is the probability to propagate $u_i$'s age to $u_j$
Method:
while $\mathcal{U}$ is not $\emptyset$ do
  for each user $u_j \in \mathcal{U}$ do
    define $\mathcal{A}_j = \emptyset$ for $u_j$
  end for
  for each user $u_i \in \mathcal{L}$ do
    for each user $u_j \in \mathcal{N}(u_i)$ and $u_j \notin \mathcal{L}$ do
      add the pair $(a_i, p(i \to j))$ to $\mathcal{A}_j$
    end for
  end for
  for each user $u_j \in \mathcal{U}$ do
    if $\mathcal{A}_j \neq \emptyset$ then
      $a_j \leftarrow$ age in $\mathcal{A}_j$ with the maximum probability
      $\mathcal{L} \leftarrow \mathcal{L} \bigcup \{u_j\}$
      $\mathcal{U} \leftarrow \mathcal{U} \setminus \{u_j\}$
    end if
  end for
end while
```

We outline the MCAP algorithm in Algorithm 1. The algorithm iteratively propagates the age of labeled users to the unlabeled users. In each iteration, an unlabeled user stores the propagated age from her neighbors together with a probability score, which is based on the weights of the edges between her and her neighbors. At the end of each iteration, if an unlabeled user has received some propagated age values, we set the age with the maximum probability to be the age of this user. The user is then added to the labeled set and removed from the unlabeled set. The algorithm continues until all users in the unlabeled set have been labeled and moved to the labeled set.

While the MCAP algorithm may appear very simple, it is also very efficient. In our experiments, we find that for a large social network with around 25% labeled users, we can predict the age of all the unlabeled users within 3 or 4 iterations.

6.3 Finding Maximal Cliques

In this subsection we discuss how we find all maximal cliques inside a social network. As defined, a maximal clique cannot be extended by including one more adjacent node, that is, a clique which does not exist exclusively within the node set of a larger clique. We make use of the *Bron-Kerbosch* algorithm [6] for finding maximal cliques. The basic form of the *Bron-Kerbosch* algorithm is a recursive backtracking algorithm that searches for all maximal cliques in a given graph G. As the *Bron-Kerbosch* algorithm is a well-known existing algorithm, we do not give the details here.

7. EMPIRICAL EVALUATION

In this section, we carry out a set of experiments to evaluate our algorithm. We find that our propagation algorithm is quite efficient. After finding the maximal cliques in a social network, running the MCAP algorithm on our data sets can be finished in seconds when running on a regular laptop machine with a double core 1.80GHz processor and 4GB of memory.

7.1 Evaluation Metrics

We introduce the following metrics to help us evaluate the performance of our proposed algorithm. We compare the predicted age of a user versus her actual age. The first metric we consider is the **Error Gap** which quantifies the gap in years between the actual age of the user $a_{\text{act}}(u)$ and the predicted age $a_{\text{pre}}(u)$. The **Error Gap** for user u is defined as:

$$\text{ErrGap}(u) = |a_{\text{act}}(u) - a_{\text{pre}}(u)|. \tag{4}$$

In order to give a strong insight into the distribution of age prediction errors, the next metric **Accuracy** considers the percentage of users with their **Error Gap** capped within d years:

$$\text{Accuracy}(d, \mathcal{U}) = \frac{|\{u : u \in \mathcal{U} \text{ and } \text{ErrGap}(u) \leqslant d\}|}{|\mathcal{U}|}. \tag{5}$$

7.2 Experiments on MG

7.2.1 *Leave-One-Out Evaluation*

In the first set of experiments, we want to check when only the test user's age is unknown but all other users' ages are known, how our algorithm can perform. Due to the richness of gold standard data in the MG data set, it can easily meet the requirement of all friends' ages in an ego network being known. As age information is quite sparse in Twitter, this requirement seems too rigorous which will leave us with not enough data. Thus we only carry our this experiment on the MG data. Since the Twitter data we have is restricted in Singapore, this experiment was carried out only on the Singapore MG data to keep consistency with later experiments. In this Leave-One-Out experiment, for each user with known age, we hide her age and run the MCAP algorithm to predict her age. We also consider the following baselines: (1) Random Guess: We randomly assign an age to the user based on a uniform distribution over all the age values. (2) Friend Random: We randomly select a friend of the user and use the friend's age for prediction. (3) Friend Average: We use the average age of the friends to predict the user's age. (4) Friend Median: We use the median age of the friends to predict the user's age. For our own MCAP algorithm, we consider two variations. The first is the standard algorithm and the second, which we refer to as "No Min&Max," is the version where we ignore the minimum and maximum age in a clique. The results are shown in Figure 9.

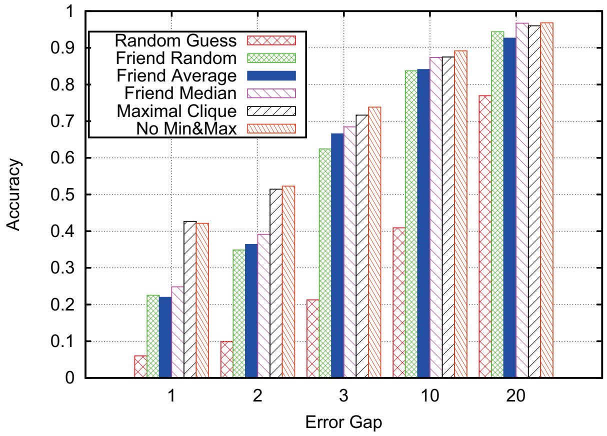

Figure 9: Results for Leave-One-Out for MG.

The figure shows that maximal clique age propagation algorithm as well as its optimized version performs better than the four baselines. With **ErrGap** no bigger than 3, the accuracy of the optimized MCAP can reach 73.84%. At the same time, it is worth mentioning that when **ErrGap** is less than 2, our algorithm performs substantially better than the baselines.

7.2.2 *Leave-Many-Out Evaluation*

The Leave-One-Out experiment results are promising but the setting might not be close to real life cases. So we carry out another evaluation method, attempting to recover the ages of many individuals simultaneously. To do this, we first remove the age information from 75% of the individuals who have provided it. We then attempt to recover the age of all other users. Here we keep users with at least one friend remaining in the set with known age. The performance is shown in Figure 10. We can see that overall the performance is worse than Leave-One-Out results, as we now have much less information about the age of a user's friends. Predicting in this way correctly predicts 54.62% of users within 3 years of gap from their actual age after 3 iterations.

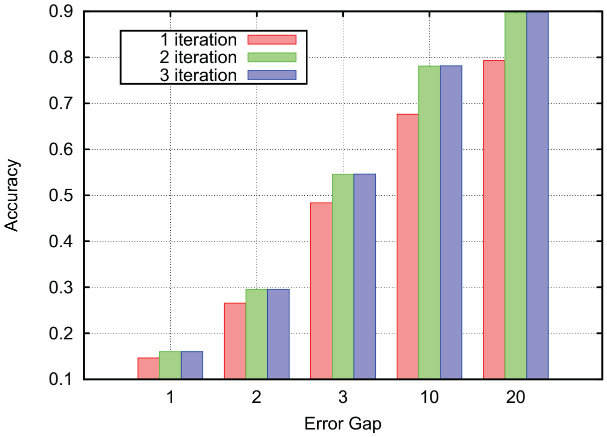

Figure 10: Results for Leave-Many-Out for MG.

In our experiment, we can run our prediction algorithm iteratively, using the newly guessed ages as input as well as the ages provided by the 25% users. Figure 10 shows the performance of such iterative approach. We can see that the

ErrGap	**3**	**5**	**10**	**20**
MCAP	0.6829	0.7902	0.9258	0.9830
Friend Median	0.4907	0.6830	0.8381	0.9350
Average	0.2573	0.5040	0.8554	0.9496

Table 2: Results for age prediction on Twitter

second iteration is substantially better than the first iteration. For error gap being no bigger than 3, the performance rise to 54.60% from 48.35%. For error gap being no bigger than 10, the performance rises to 78.09% from 67.64%. That is to say, when we are predicting the age of many individuals at once, we can perform better by using the information contained in the links between the individuals whose ages we are trying to predict. In the first pass, we make our prediction based only on the known ages. In subsequent passes, we can use the predicted ages as part of the input, to improve performance. As shown, the results converges quickly.

7.3 Experiments on Twitter

To better understand the portability of the age predictor, we next conduct experiments on the Twitter data set we collected. As mentioned above, gold standard data in Twitter is really rare. For about 25K Twitter users we crawled, after filtering and manual correction, there are only about 2.94% users who have directly stated their age information. Even after the heuristic process of extracting birthday mentions, the rate of gold standard data still remains pretty small (only 13.21%). When doing propagation, users with age information and users without age information are both kept for building graph. When computing the performance, we only look at users with age information. We compare our method with a baseline using the median age of friends (referred to as “Friend Median”). In the case when a user has no friend with age information, we resort to using the average age of all the training users. We also compare with another baseline (referred to as “Average”) by assigning the average age of all the training users to our test users.

To address the problem of insufficient data, we divide the ground truth data (manually corrected data) into 10 subsets. Then, we perform 10-fold cross validation. For each round, we hold out one subset, using other subsets as initial labeled data to run the maximal clique based age propagation algorithm. After prediction, we compute the performance based on the held-out subset. The final results are averaged over the 10 rounds. The results are shown in Table 2.

The results in Table 2 are surprisingly good, considering that only 13.21% of users' age information is leveraged and 10% of it has been held out for evaluation. Even in such circumstances, the results are still much better than Leave-Many-Out results on MG. This suggests that the Twitter data set might have a much more concentrated age distribution on cliques, making it easier to do age prediction using maximal cliques.

8. CONCLUSIONS

The social relationships in social media platforms provide strong evidence of an individuals' age information. We validated our hypothesis about the relationship between average age gap and clique size on the MG data set. Based on this hypothesis, we presented a new algorithm, maximal clique-based age propagation (MCAP), that leverages the age distribution of a user's ego network to predict the user's age. We first find out all maximal cliques in the undirected network built on user friendships, and then weight the network based on clique size. We iteratively propagate the ages of the age-known users to infer the age of age-unknown users and add them to the set of age-known users. With a small number of initial age-known users, the age predictor efficiently infers 54.62% of MG users within an error gap of 3 and 79.02% of Twitter users within an error gap of 5 from their actual age.

As a purely social network based approach, we anticipate continued refinement of this approach through incorporating more information. For example, there are not only friendship relations in MG, but also best-friend and blacklist relations. We are also interested in combining this purely relationship-driven approach with some linguistic features to develop more robust predictors.

9. ACKNOWLEDGMENTS

This research is supported in part by the PhD Joint Education grants from Beijing Institute of Technology, Chinese National Program on Key Basic Research Project (Grant No. 2013CB329605). This research is also partially supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA). We thank the anonymous reviewers for providing their valuable comments on the pervious version of this paper.

10. REFERENCES

[1] F. Al Zamal, W. Liu, and D. Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*, 2012.

[2] E. Alfonseca, M. Pasca, and E. Robledo-Arnuncio. Acquisition of instance attributes via labeled and related instances. In *SIGIR*, pages 58–65, 2010.

[3] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9), 2007.

[4] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[5] S. Bergsma and B. Van Durme. Using conceptual class attributes to characterize social media users. pages 710–720, 2013.

[6] C. Bron and J. Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577, 1973.

[7] C. A. Davis Jr, G. L. Pappa, D. R. R. de Oliveira, and F. de L Arcanjo. Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, 15(6):735–751, 2011.

[8] A. T. Fiore and J. S. Donath. Homophily in online dating: when do you like someone like yourself? In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, pages 1371–1374. ACM, 2005.

[9] R. Jones, R. Kumar, B. Pang, and A. Tomkins. I know what you did last summer: query logs and user

privacy. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 909–914. ACM, 2007.

[10] D. Jurgens. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[11] M. Koppel, J. Schler, and K. Zigdon. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 624–628. ACM, 2005.

[12] G. Kossinets and D. J. Watts. Origins of homophily in an evolving social network1. *American Journal of Sociology*, 115(2):405–450, 2009.

[13] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205, 2006.

[14] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots+ machine learning. In *SIGIR*, pages 435–442, 2010.

[15] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[16] J. M. McPherson and L. Smith-Lovin. Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American sociological review*, pages 370–379, 1987.

[17] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[18] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: Inferring user profiles in online social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 251–260, 2010.

[19] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260. ACM, 2010.

[20] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder. "How old do you think i am?" A study of language and age in twitter. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[21] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in Twitter. In *Proceedings of the Second International Workshop on Search and Mining User-generated Contents*, pages 37–44, 2010.

[22] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 723–732. ACM, 2012.

[23] S. B. Kurth. Friendships and friend relations. *Social Relationships*, pages 136–170, 1970.

[24] F. A. Zamal, W. Liu, and D. Ruths. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *ICWSM*, 2012.

[25] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.