Your Age Is No Secret: Inferring Microbloggers' Ages via Content and Interaction Analysis

Jinxue Zhang
Electric Engineering
Arizona St. University
650 E. Tyler Mall
Tempe, AZ 85281

Xia Hu
Computer Sci.&Engineering
Texas A&M University
330B H.R. Bright Building
College Station, TX 77843

Yanchao Zhang
Electric Engineering
Arizona St. University
650 E. Tyler Mall
Tempe, AZ 85281

Huan Liu
Computer Sci.&Engineering
Arizona St. University
699 S. Mill Ave
Tempe, AZ 85281

Abstract

Microblogging systems such as Twitter have seen explosive use in public and private sectors. The age information of microbloggers can be very useful for many applications such as viral marketing and social studies/surveys. Current microblogging systems, however, have very sparse age information. In this paper, we present MAIF, a novel framework that explores public content and interaction information in microblogging systems to explore the hidden ages of microbloggers. We thoroughly evaluate the accuracy of MAIF with a real-world dataset with 54,879 Twitter users. Our results show that MAIF can achieve up to 81.38% inference accuracy and outperforms the state of the art by 9.15%. We also discuss some countermeasures to alleviate the possible privacy concerns caused by MAIF.

1 Introduction

Microblogging systems have become important platforms for information sharing and social networking. As the end of 2015, Twitter—the most popular microblogging system in the world—has 320 million monthly active users. People have been using microblogging systems in social networking, massive information campaigns, public relationships, political campaigns, pandemic and crisis situations, business marketing, crowdsourcing, and many other public/private contexts (Zafarani, Abbasi, and Liu 2014).

Age information is much scarcer in microblogging systems than in traditional online social networks (OSNs). In a traditional OSN such as Facebook or LinkedIn, the users aim to maintain their personal identities and social connections with friends, so their personal profiles often contain true birthdate, school finishing/enrollment time, and other sensitive information, which can be directly used to infer accurate user ages (Dey et al. 2012). As more open social-networking platforms, however, microblogging systems are more informal than traditional OSNs in terms of maintaining social identities such that their user profiles often have no specific age-related information.

Age information in microblogging systems have important applications in both positive and negative ways. As an example for the positive aspect, the ability to select a group of users in the specific age range can enable numerous social and health studies such as investigating the diet habits of the college students between 18 and 22 years old and monitoring the workout habit of elderly or middle-aged people. The age information is also useful for cost-effective business marketing. For example, to launch a viral marketing campaign for a new wearable device via Twitter, a known strategy is for the marketer to seed the product with a few selected influential users from 30 to 50 years old who can potentially influence a disproportionately large number of others and also quickly trigger a cascade of influence (Zhang et al. 2015a). As an example for the negative aspect, being able to infer the age and other latent attributes based on the users' public information such as tweets can help the adversary better profile the users for planning more advanced attacks such as spam campaigns or phishing attacks aiming at the elderly.

Accurate age inference is still an open challenge in microblogging systems due to three reasons. First, as stated before, the age information in microblogging systems is scarce. The microblogging service provider such as Twitter offers no explicit channel for users to indicate their age information. Therefore, existing methods that inferring a user's personal attributes directly from his/her online social neighbors' (Mislove et al. 2010; Dey et al. 2012; Liao et al. 2014) are inapplicable because the neighbors' age information is also missing. Second, the microblogging messages (*microblogs* for short) posted by the users are highly unstructured, noisy, and massive. For example, each tweet in Twitter is composed of at most 140 characters, and hence microbloggers have created various slang and abbreviations to express their feelings and opinions, such as "wish4u a gr8 day" meaning "wish for you a great day." There are about 500M tweets per day, and each user is allowed to send up to 1000 tweets per day. These constrains make traditional text analysis for regular documents inapplicable in our context (Hu et al. 2009). Finally, the content information of the microbloggers is connected by online interactions such as following and retweeting. Traditional content analysis treating each user's content information independently fails to explore such rich online interactions.

In this paper, we propose a new framework to infer microbloggers' ages by seamlessly integrating the content and interaction information on microblogging systems. Our key idea is driven by the presence of *homophily*, which has been discovered in many social studies (Zamal, Liu, and Ruths

2012). In our context, homophily refers to the tendency of a microblogger to associate and bond with similar others. For example, two colleague alumni in the same age group would be more inclined to follow each other in microblogging systems. In addition, the microbloggers with more intensive online interactions are very likely to have more similar content information in their microblogs. To fully leverage the presence of homophily in microblogging systems, this paper aims to answer two critical questions. First, how can we model both the content and interaction information in microblogging systems? Second, how can we effectively combine the content and interaction information together to accurately infer a microblogger's age?

Our contributions are summarized as follows.

- We motivate and formally define the age inference problem in microblogging systems with both content and interaction information.
- We propose MAIF, a unified framework to model and seamlessly integrate both the content and interaction information by considering the homophily of the content information among connected microbloggers.
- We thoroughly evaluate the proposed framework on a real-world dataset with 54,879 Twitter users, the largest in the community. Our results show that MAIF can achieve up to 81.38% inference accuracy and outperforms the state of the art by 9.15%.
- We outline some countermeasures for those wishing to preserve age privacy if our system were in place.

The rest of this paper is organized as follows. Section 2 introduces the background and defines the problem. Section 3 details the age inference framework. Section 4 evaluates the proposed framework. Section 5 surveys the related work. Section 6 summarizes this paper and future work.

2 Background and Problem Statement

In this paper, we use Twitter as a representative microblogging system to illustrate our proposed framework. In what follows, we briefly introduce Twitter and then formally define the age inference problem.

After registering an account in Twitter, a user can post text-based microblogging messages of up to 140 characters, known as *tweet*s. S/he can also retweet, reply to, mark favorite any other public Twitter user's tweets. The user can also mention anyone else in the tweet by @someone. Unlike Facebook-like OSNs, the social relationships in Twitter are unidirectional by users *following* others. If user A follows user B, A is B's *follower*, and B is A's *followee*. In this paper, we call A and B are *friends* if and only if A and B follow each other.

In this paper, we are interested in classifying a user into one of c predefined age groups, which are further defined according to a widely-used adult development model (Levinson 1986) in Section 4.1. We do not want to infer the user's exact age for two main reasons. First, we observed that the majority of commercial advertisements and online surveys focus on the users of a specific age group. Hence our framework can well satisfy the requirements of such important applications. Second, there are no enough labeled users to infer exact ages. Nevertheless, our framework is flexible and extensible to infer the exact age by having one group per age as long as there are sufficient labeled users.

We assume that there is a set of *labeled* users in Twitter with explicit age information specified through tweets or other sources. As stated before, labeled users in microblogging systems are scarce. To tackle this challenge, we design a novel method to collect sufficient labeled users for building and evaluating our proposed framework. The details for labeled user collection are postponed to Section 3.1.

Problem Formulation. We formally model the microblogger's age inference problem as follows. Let $\mathcal{U}$ denote a set of n labeled users, $\mathcal{U}_{\mathcal{F}} \subseteq \mathcal{U}$ denote the union of each labeled user's friends in $\mathcal{U}$, $\mathcal{X}_u$ represent the microblogging messages of each user $u \in \mathcal{U}$ in the past year from the same given date, and $\mathbf{Y} \in \mathbb{R}^{n \times c}$ be an age-label matrix in which c is the number of classes, and $\mathbf{Y}_{i,j}$ is equal to 1 if user i is in age group j and 0 otherwise. We aim to build a classifier $\mathbf{W}$ to automatically assign the age labels for unknown users according to their microblogging messages. Here we leverage online interaction information (if there is) to train the classifier but do not need it for labeling unknown users, which is critical for the usability of the framework because the labeled users are scarce and so for the interactions between the unknown and labeled users.

3 Microbloggers' Age Inference Framework

As mentioned before, it is very challenging to infer the age information of Twitter users because of the tweets' unstructured, noisy, and massive nature as well as the scarcity of labeled users in Twitter. In this section, we first conduct an analysis of a dataset which is crawled via a novel method, and the analysis motivates the design of our microblogger's age inference framework (MAIF for short). Then we present a content metric to model each user u's tweet set $\mathcal{X}_u$ in Section 3.2. Next, we adopt a sparse representation method to model the content information for age inference in Section 3.3 and then use community structure to model the interaction information in Section 3.4. Finally, we integrate the content and interaction information to formulate the age inference problem as a convex optimization problem in Section 3.5 and then present our solution in Section 3.6.

3.1 Data Crawling and Analysis

We design a method to crawl the ground-truth labeled users. Inspired by (Zamal, Liu, and Ruths 2012; Liao et al. 2014), we found that many users like to send their birthday greetings to their friends by posting a tweet containing two parts: a phrase of "happy yth birthday" where y is the age of the friend, and a mentioned user who is likely to be the friend's Twitter name. For example, user A has posted a tweet "Happy 24th Birthday to my best friend @B." It is clear that user B is 24 years old now. We then use Twitter's Streaming API to record all the tweets which contain one of the keywords "happy yth birthday" with y ranging from 14 to 70. Since the tweets are noisy, we use the following tricks to refine the collected tweets. First, we only select the tweets which mention only one person because it is very difficult to determine

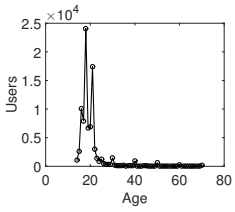
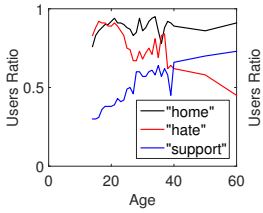
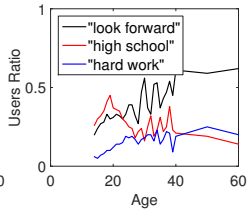


Figure 1: The age distribution in the ground-truth dataset.



(a) Example 1



(b) Example 2

Figure 2: The age-keyword usage pattern.

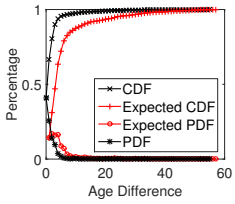


Figure 3: The distribution of the age gap on friend pairs.

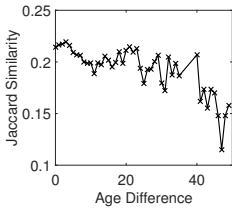


Figure 4: The Jaccard content similarity on friend pairs.

which user has the age information if more than one user have been mentioned. Moreover, if the tweet sender and the mentioned user are not friends, the tweet is excluded. This trick is to deal with the cases that the sender may just mention and require a celebrity to greet the sender's friend (e.g., an ordinary person, not mentioned). Since our framework relies on credible interactions among the users, such tweets and the corresponding users should not be considered. Finally, each user mentioned in the remaining tweets is assigned an age label y from the tweet, and we check the labeled users manually to exclude the users who are obviously not at the labeled age. The readers can check (Kumar, Morstatter, and Liu 2013) for more details on how to crawl and analyze the Twitter system.

Based on the above method, we crawled the largest age-based ground-truth Twitter dataset in the community which is composed of 54,879 labeled users, each user's labeled friends, and each user's tweets from June 1, 2014 to May 30, 2015. Fig. 1 shows the age distribution of our dataset, which is consistent with the result in (Zamal, Liu, and Ruths 2012; Liao et al. 2014). As we can see, 88.06% of labeled users are aged below 24, which is expected because young people are more likely to explicitly express their greetings using the social media. We notice that the dataset is biased toward young people, as Pew shows that 47% of Twitter users are older than 30 years old (Duggan et al. 2015). However, the datasets with the similar distribution have been used in many previous work (Zamal, Liu, and Ruths 2012; Liao et al. 2014), and it is still valuable and reliable for motivating our system design. We will also evaluate the impact of the biased dataset on system performance in Section 4.

Fig. 2 shows the generation gap (Giancola 2006) in terms of the word usage. Specifically, we selected six keywords, "home", "hate", "support", "look forward", "high school", "hard work", and check how many users at each specific age have used them in their tweet corpus. We can see that people with different ages have different keyword usage patterns. For example, users aged from 18 to 21 increase the usage of "home" because they might leave home for colleges; older people are less likely to use "hate" because they are more mature, but they are more likely to use "hard work" because they are highly engaged in the professional work; etc.

Fig. 3 and Fig. 4 demonstrate the social homophily in Twitter. Specifically, we first investigate the similarity in the ages of the users who have online interactions. For this purpose, we measure the age difference of each friend pair in the dataset and draw the distribution in Fig. 3. To evaluate the impact of dataset bias, we also calculate the expected distribution of the age difference. To that end, we let each user befriend with each of other 54,878 users, and then measure the number of user pairs with a specific age difference. As we can see, in the original dataset, 40.84% of friend pairs have the same age, and 93.64% of the pairs have the age difference within 5 years while only 14.21% and 63.32% of the pairs in the fully-connected network have the same age and the age difference within 5 years, respectively. To measure the corpus similarity of each friend pair, we treat each user's tweets as a set of words and compute a Jaccard metric as $\frac{|\mathcal{A}\cap\mathcal{B}|}{|\mathcal{A}\cup\mathcal{B}|}$, where $\mathcal{A}$ and $\mathcal{B}$ denote the word sets for the two users involved, respectively. For all the friend pairs with the same age difference, we average their Jaccard similarities. As shown in Fig. 4, the corpus similarity decreases as the age difference of a friend pair increases.

We can draw two observations from the above analysis. First, the users at different ages have different topics in their tweets due to the age gap. Second, because of the social homophily (Zamal, Liu, and Ruths 2012), a user is more likely to befriend with thoses of the similar age, and their tweet topics tend to have higher similarity than the friend pairs with large age difference. These two observations drive us to design a framework to well integrate the content and interaction information to infer a Twitter user's age.

3.2 Model Tweets by τ-gram

Given the $\mathcal{X}_i$ of tweets of any labeled user $i \in \{1, \cdots, n\}$ in the past year, we first need to construct a mathematical model to represent it. Here we use a feature matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ to model labeled users' tweets, where m refers to the dimension of a feature space $\mathcal{F}$ in the whole message space. In what follows, we describe how to construct the feature space $\mathcal{F}$ and then the feature matrix $\mathbf{X}$.

We first remove *stop words* in a stop-word list,[1] in which the words such as "the" and "those" are considered more general and meaningless. Then we conduct stemming (Porter 1997) to reduce inflected words to their stem forms such that the words with different forms can be related to the same word. For example, "watch", "watching", and "watched" are all reduced to "watch".

Next, we represent the feature space for the cleansed tweets using a τ-gram technique, which is widely used for

[1] http://www.lextek.com/manuals/onix/

statistical text analysis. The τ-gram technique splits a give message into sequences of τ contiguous words, each referred to as a τ-gram with τ ranging from 1 to the message length. For example, consider a tweet {"Playing basketball against those guys was a bad idea"}. After removing stop words and performing stemming, we have {"play basketball against guy bad idea"}. The corresponding 1-grams are {"play", "basketball", "against", "guy", "bad", "idea"}, and the corresponding 2-grams are {"play basketball", "basketball against", "against guy", "guy bad", "bad idea"}. We let $\mathcal{N}_i$ denote the τ-grams of $\mathcal{X}_i$ for each user $i \in \mathcal{U}$ for all possible values of τ. Then we choose the top m most frequent τ-grams in $\bigcup_{1\leq i\leq n} \mathcal{N}_i$ as the feature space $\mathcal{F}$.

Finally, we use the Term Frequency Inverse Document Frequency (TF-IDF) technique (Leskovec, Rajaraman, and Ullman 2014) to derive each element $\mathbf{X}_{i,j}$ in $\mathbf{X}$. Specifically, let $\Gamma(j)$ be the number of times a τ-gram j appears in the τ-gram list $\mathcal{N}_i$ of user i, $\Gamma_i^* = \max_{j\in\mathcal{N}_i} \Gamma(j)$, and $\Gamma'(j)$ denote the number of users in $\mathcal{U}$ whose τ-gram lists contain j. We define

$$\mathbf{X}_{i,j} = (0.5 + 0.5 * \frac{\Gamma(j)}{\Gamma_i^*}) * \log(\frac{n}{\Gamma'(j)}) \ . \tag{1}$$

The above normalization based on Γ_i^* is necessary because the users normally have very different tweet sets and thus different τ-gram lists. We refer interested readers to (Leskovec, Rajaraman, and Ullman 2014) for the details of the TF-IDF technique.

It is a common practice to use 1-grams and 2-grams only for high computational efficiency without significantly sacrificing the analysis accuracy. So the feature space and matrix can be constructed very quickly in practice.

3.3 Modeling Content Information

Given the feature matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ and the age-label matrix $\mathbf{Y} \in \mathbb{R}^{n\times c}$, a traditional method to build the classifier $\mathbf{W}$ is Least Square optimization(Lawson and Hanson 1974), which learns a weighted model to minimize the estimation and the labeled data by solving

$$\min_{\mathbf{W}} \ \frac{1}{2}\|\mathbf{XW} - \mathbf{Y}\|_F^2, \tag{2}$$

where $\|\mathbf{A}\|_F$ represent the Frobenius norm of matrix $\mathbf{A}$ which is defined as $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m \mathbf{A}_{i,j}^2}$.

The traditional Least Square method for a large feature set can lead to overfitting (Tibshirani 2011) in that the learned model may be too specific due to the limited training data and thus be inaccurate for inferring the ages of unknown users. Moreover, it has been observed in many domains that the underlying representations of many objects are sparse. For example, a signal could be efficiently reconstructed by far fewer samples in compressive sensing (Baraniuk 2007); when people speed-read documents, they may seek a sparse representation with key phrases or words instead of fully understanding every single word (Marinis 2003). These sparse features represent the given object more accurately and efficiently by capturing its underlying essence. In addition, by selecting a sparse and meaningful group of τ-grams rather than non-intuitive ones for each user, it could help sociologists, market planners and even the public to understand the behavior of the people in different age groups. To find and explore these sparse features in our feature space, we can improve the model defined in Eq. (2) by assigning higher weight to the most representative τ-grams. One widely-used method(Tibshirani 2011) is to introduce the ℓ_1-norm regularization for the weight matrix $\mathbf{W}$ as follows,

$$\min_{\mathbf{W}} \ \frac{1}{2}\|\mathbf{XW} - \mathbf{Y}\|_F^2 + \lambda_1\|\mathbf{W}\|_1, \tag{3}$$

where $\|\mathbf{W}\|_1 = \sum_{i=1}^n \sum_{j=1}^m |\mathbf{W}_{i,j}|$, and λ_1 is the parameter to control the sparse regularization. By adding this ℓ_1-norm constraint to the minimization problem, it enforces the coefficients of many non-representative features in $\mathbf{W}$ to be zero, thus making these features have no effect on the prediction model. With this strategy, we select relatively more "important" features (equivalently, τ-grams) to represent each age group.

3.4 Modelling Online Interaction Information

The content information in Twitter is networked. As shown in Section 3.1, people within the same age group have higher probability to share content similarity and also befriend with each other. For example, two college classmates follow each other on Twitter, often discuss final exam preparations for the same course, and/or cheer for the wins of their college sports teams. Given such observations, the content model in Eq. (3) should assign higher weights to similar τ-grams, such as "final" and "exam", so that the two users can be classified into the same age group with high probability. How to achieve this, however, is challenging because the two users very likely also tweet on different topics. Below we present how to model the online interactions among labeled users and then how to integrate the interaction information into the content model in Eq. (3).

We use the community concept to model the online interactions among the labeled users. For this purpose, it is worth noting that we can construct an undirected social graph from the labeled dataset, where each vertex corresponds to a labeled user, and an edge exists between two users if and only if they are friends (i.e., each other's follower and followee). It has been widely reported that the users with the similar attributes such as ages would connect with each other more than the users with different attributes, hence forming a local community (Mislove et al. 2010). The community structure can be inferred by maximizing the *modularity* (Newman and Girvan 2004), which is defined as follows.

Definition 1 (Modularity)**.** *Given an undirected graph $G = \langle \mathcal{U}, E\rangle$, where $|\mathcal{U}| = n$ is the user set, and $e_{ij} \in E$ equals 1 if users i and j are friends and equals 0 otherwise. Assume that G has been partitioned into k communities, and that each user belongs to one and only one community. The modularity of this partition is defined as*

$$Q = \frac{1}{2t}\sum_{i,j}(e_{ij} - \frac{d_i d_j}{2t})\delta(C_i, C_j), \tag{4}$$

where $t = \frac{1}{2}\sum_{i,j} e_{ij}$ is the number of edges in G, $d_i = \sum_j e_{ij}$ is the degree of user i, C_i is the community containing user i, and the $\delta-$function $\delta(C_i, C_j)$ is 1 if $C_i = C_j$ and 0 otherwise.

The intuition behind the modularity is as follows. $\frac{d_i d_j}{2m}$ represents the expectation that any two users with degree d_i and d_j could form an edge in the graph. If they are connected (i.e., $e_{ij} = 1$) and are in the same community (i.e., $C_i = C_j$), they will contribute to the whole modularity Q. If they are not connected (i.e., $e_{ij} = 0$) but are in the same community (i.e., $C_i = C_j$), they will reduce the modularity Q. Finally, if they are in different communities (i.e., $C_i \neq C_j$), they have no impact on Q. Hence, the more edges in the same community, the higher its modularity.

Next, we present how to infer the community structure by maximizing the modularity. Let matrix $\mathbf{G} \in \mathbb{R}^{n\times n}$ represent the adjacent matrix for graph G where $\mathbf{G}_{i,j}$ equals 1 if $e_{ij} = 1$ and 0 otherwise. Let matrix $\mathbf{C} \in \mathbb{R}^{n\times k}$ represent a community partition for G where $\mathbf{C}_{i,j}$ is 1 if user i is in community j, and 0 otherwise. Note that $\sum_j \mathbf{C}_{i,j} = 1$ since any user belongs to one and only one community. Then we could formulate the community partition problem as

$$\max_{\mathbf{C}} \quad Tr(\mathbf{CMC^T}), \quad s.t. \quad \mathbf{CC^T} = \mathbf{I} \tag{5}$$

where

$$\mathbf{M} = \mathbf{G} - \frac{\mathbf{d}\mathbf{d}^T}{2t} \tag{6}$$

where $\mathbf{d}$ is the degree vector for G, and $Tr(\mathbf{A}) = \sum_i \mathbf{A}_{i,i}$ represents the sum of the diagonal elements of $\mathbf{A}$. Since this problem is NP-hard (Newman and Girvan 2004), we resort to the widely used Louvain method (Blondel et al. 2008) to obtain the approximation result.

After the community structure $\mathbf{C}$ is obtained, we expect that the users from the same community are in the same age group. Therefore we can use the community structure to improve our model in Eq. (3). To that end, inspired by (Tang and Liu 2012), given $\hat{\mathbf{Y}}$ as the estimated age group labels for all the users in $\mathcal{U}$, we first compute the scatter of user pairs who are in the same community but have been estimated in either the same age group or two different age groups as:

$$\mathbf{S} \quad = \hat{\mathbf{Y}}^T\mathbf{F}\mathbf{F}^T\hat{\mathbf{Y}} \tag{7}$$

where $\mathbf{F}$ is the weighted community indicator matrix, which can be obtained from $\mathbf{C}$ as

$$\mathbf{F} = \mathbf{C}(\mathbf{C}\mathbf{C}^T)^{-\frac{1}{2}}, \tag{8}$$

where $\mathbf{F}_{ij}$ equals $\frac{1}{\sqrt{f_j}}$ if user i is in community C_j with f_j users and equals 0 otherwise.

It can be easily found that in Eq. (7), since $\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$ is a diagonal matrix with the (i,i)-th element equal to the number of users in the i-th age group, the (i,i)-th element of $\mathbf{S}$ measures how many user pairs in the i-th age group are in the same community, and the (i,j)-th $(i \neq j)$ element of $\mathbf{S}$ measures how many user pairs in the i-th age group and j-th age group are in the same community. Therefore, in order to classify the users in the same community into the same age range, we just need to maximize the sum of (i,i)-th element in $\mathbf{S}$, i.e.,

$$\max_{\mathbf{W}} \quad Tr(\mathbf{S}). \tag{9}$$

Note that we ignore the user pairs who are in the same community but in different age groups because they violate the community structure.

3.5 Integrating Content and Interaction Information

Many existing methods on age estimation use either content or interaction information independently by assuming that these two pieces of information are unrelated. This assumption is not valid according to the intuition and also our data analysis in Section 3.1. So we propose to integrate both the content and interaction information into a unified model.

Particularly, since $\hat{\mathbf{Y}} = \mathbf{XW}$, Eq. (9) can be re-written as

$$\max_{\mathbf{W}} \quad Tr(\mathbf{W}^T\mathbf{X}^T\mathbf{F}^T\mathbf{F}\mathbf{X}\mathbf{W}). \tag{10}$$

By considering both the content information and interaction information, the age estimation problem defined in Eq. (3) could be reformulated as follows,

$$\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{XW}-\mathbf{Y}\|_F^2+\lambda_1\|\mathbf{W}\|_1-\frac{\lambda_2}{2}Tr(\mathbf{W}^T\mathbf{X}^T\mathbf{F}^T\mathbf{F}\mathbf{X}\mathbf{W}), \tag{11}$$

where λ_1 and λ_2 are the parameters for sparse regularization (for content information) and integration of interaction information, respectively. By varying these two parameters, we could set the importance of sparse regularization and interaction integration on the original Least Square model.

3.6 An Optimization Algorithm

The problem defined in Eq. (11) is non-smooth because the ℓ_1 regularization $\|\mathbf{W}\|_1$ is not differentiable. Hence we transform it into its differentiable Lagrange dual function as:

$$\begin{aligned} \min_{\mathbf{W}} \quad & \frac{1}{2}\|\mathbf{XW} - \mathbf{Y}\|_F^2 - \frac{\lambda_2}{2}Tr(\mathbf{W}^T\mathbf{X}^T\mathbf{F}^T\mathbf{F}\mathbf{X}\mathbf{W}), \\ \text{s.t.} \quad & \|\mathbf{W}\|_1 \leq z, \end{aligned} \tag{12}$$

where $z \geq 0$ is the radius of the ℓ_1-ball and has a one-to-one correspondence with λ_1. Let

$$f(\mathbf{W}) = \frac{1}{2}\|\mathbf{XW} - \mathbf{Y}\|_F^2 - \frac{\lambda_2}{2}Tr(\mathbf{W}^T\mathbf{X}^T\mathbf{F}^T\mathbf{F}\mathbf{X}\mathbf{W}), \tag{13}$$

we can see that $f(\mathbf{W})$ is a smooth objective function, and the optimization problem is convex which can be solved by gradient descending methods. It is known (Bertsekas 1999) that the gradient step

$$\mathbf{W}^{(k)} = \mathbf{W}^{(k-1)} - \frac{1}{t^{(k)}} \bigtriangledown f(\mathbf{W}^{(k-1)}) \tag{14}$$

for solving the smooth optimization problem in Eq. (12) can be treated as finding the minimum Euclidean projection (Boyd and Vandenberghe 2004) of $\mathbf{W}^{(k)}$ defined above on the ℓ_1-ball $\|\mathbf{W}\|_1 \leq z$, which is

$$\mathbf{W}^{(k)} = \arg\min_{\mathbf{W}} M_{t^{(k)}}(\mathbf{W}, \mathbf{W}^{(k-1)}), \tag{15}$$

$$M_{t^{(k)}}(\mathbf{W}, \mathbf{W}^{(k-1)}) = f(\mathbf{W}) + \langle \mathbf{W} - \mathbf{W}^{(k-1)}, \bigtriangledown f(\mathbf{W}^{(k-1)}) \rangle + \frac{t^{(k)}}{2} \|\mathbf{W} - \mathbf{W}^{(k-1)}\|_F^2, \tag{16}$$

where $t^{(k)}$ is the step size, $\langle \mathbf{A}, \mathbf{B} \rangle = Tr(\mathbf{A}^T\mathbf{B})$ denotes the matrix inner product, and

$$\bigtriangledown f(\mathbf{W}^{(k-1)}) = \mathbf{X}^T\mathbf{X}\mathbf{W}^{(k-1)} - \mathbf{X}^T\mathbf{Y} - \lambda_2\mathbf{X}^T\mathbf{F}^T\mathbf{F}\mathbf{X}\mathbf{W}^{(k-1)}. \tag{17}$$

Let $\mathbf{U}^{(k-1)} = \mathbf{W}^{(k-1)} - \frac{1}{t^{(k)}} \bigtriangledown f(\mathbf{W}^{(k-1)})$. The Euclidean projection in Eq. (15) has a closed-form solution (Liu, Ji, and Ye 2009) as follows,

$$\mathbf{w}_j^{(k)} = \begin{cases} (1 - \frac{\lambda_1}{t^{(k)}\|\mathbf{u}_j^{(k-1)}\|})\mathbf{u}_j^{(k-1)} & \text{if } \|\mathbf{u}_j^{(k-1)}\| \geq \frac{\lambda_1}{t^{(k)}} \\ 0 & \text{o.w.} \end{cases} \tag{18}$$

where $\mathbf{w}_j^{(k)}$ and $\mathbf{u}_j^{(k-1)}$ are the j-th rows of $\mathbf{W}^{(k)}$ and $\mathbf{U}^{(k-1)}$, respectively.

Algorithm 1 details the algorithm which comprises an outer loop and an inner loop. The inter loop from Line 4 to 9 searches the step size $t^{(k)}$ to solve the gradient step defined in Eq. (15) according to Eq. (18). The outer loop then updates the $\mathbf{W}^{(k)}$. To accelerate the gradient descent in Eq. (15), we build a linear combination of $\mathbf{W}^{(k)}$ and $\mathbf{W}^{(k-1)}$ as $\mathbf{H}^{(k)}$ in line 3 (Ji and Ye 2009). The algorithm terminates when $|f(\mathbf{W}^{(k)}) - f(\mathbf{W}^{(k-1)})| \leq \epsilon |f(\mathbf{W}^{(k-1)})|$. Similar to the proof in (Liu, Ji, and Ye 2009), given the termination parameter ϵ, it is easy to verify that the convergence rate of our algorithm is $O(\frac{1}{\sqrt{\epsilon}})$.

Algorithm 1: Classier Training for Age Inference

input : $\mathbf{X}, \mathbf{Y}, \mathbf{F}, \lambda_1, \lambda_2, \epsilon$
output: $\mathbf{W}$, i.e., the feature-to-label matrix.

1 Initialize $\mathbf{W}^{(k)} \leftarrow \mathbf{0}$, $\eta^{(0)} \leftarrow 0$, $\eta^{(1)} \leftarrow 1$, $k \leftarrow 1$;
2 **while** $|f(\mathbf{W}^{(k)}) - f(\mathbf{W}^{(k-1)})| > \epsilon |f(\mathbf{W}^{(k-1)})|$ **do**
3 Set $\mathbf{H}^{(k)} \leftarrow \mathbf{W}^{(k)} + \frac{\eta^{(k-1)}-1}{\eta^{(k)}}(\mathbf{W}^{(k)} - \mathbf{W}^{(k-1)})$;
4 **while** *True* **do**
5 Set $\mathbf{U}^{(k-1)} \leftarrow \mathbf{H}^{(k-1)} - \frac{1}{t^{(k)}} \bigtriangledown f(\mathbf{W}^{(k-1)})$;
6 Compute $\mathbf{w}_j^{(k)}$ according to Eq. (18) ;
7 **if** $f(\mathbf{W}^{(k)}) \leq M_{t^{(k)}}(\mathbf{H}^{(k-1)}, \mathbf{W}^{(k)})$ **then**
8 break;
9 $t^{(k)} \leftarrow 2 \times t^{(k-1)}$;
10 $\mathbf{W} \leftarrow \mathbf{W}^{(k)}, \eta^{(k)} \leftarrow \frac{1+\sqrt{1+4(\eta^{(k-1)})^2}}{2}, k \leftarrow k+1$;
11 **return** $\mathbf{W}$.

3.7 Inferring Age Group of an Unknown User

After we build a classifier $\mathbf{W}$, we can estimate the age range of any unknown user u as follows. We crawl the tweets from u as $\mathcal{X}_u$ in the past year and then build the τ-gram list $\mathcal{N}_u$. Based on the feature space $\mathcal{F}$, we then construct the feature vector $\mathbf{x}_u \in \mathbb{R}^{1\times m}$ by calculating the TF-IDF of each τ-gram in $\mathcal{F}$ according to Eq. (1). The final step is to estimate the age group with the maximum likelihood as follows,

$$\arg\max_{i=\{1,2,\ldots,c\}} \mathbf{x}_u \mathbf{w}_i, \tag{19}$$

where c is the number of age groups, and $\mathbf{w}_i \in \mathbb{R}^{m\times 1}$ is the i-th column of the classifier matrix $\mathbf{W}$. Note that this step needs no interaction information from user u. This feature can be very useful because it makes our algorithm above directly applicable to an arbitrary unknown user with or without interactions with labeled users in the classifier $\mathbf{W}$.

Note that MAIF needs the labelled users and their content/network information to build the classifier $\mathbf{W}$, which can be crawled by the method presented in Section 3.1. Due to the scarcity of the age information, the network information between the labelled users might be limited. However, MAIF could work even with zero network information, and as shown in the evaluation below, the richer the network information, the better the performance.

4 Evaluation

In this section, we thoroughly evaluate the proposed framework. Specifically, we want to answer these four questions:

1. How accurate is the proposed framework in comparison with other age inference schemes?
2. What is the impact of dataset bias on the performance?
3. What is the benefit of integrating both the content and social interaction information?
4. What is the impact of key parameters in the framework?

In what follows, we first introduce the dataset as well as the evaluation methodology and metrics. Then we seek to answer the above questions. Finally, we briefly discuss possible countermeasures for sensitive users to preserve their age privacy if our framework were deployed.

4.1 Dataset, Methodology and Metrics

We first partition the Twitter users into five groups according to Levinson's adult development model (Levinson 1986):

- Group 1: 14-18. This group is for juvenile and adolescence users. Since Twitter only allows the users older than 13 years to access the service, we start this group from 14 years old.
- Group 2: 19-22. According to Levinson's model, this group is a transition phase from the pre-adulthood to the early adulthood. People in this age group are usually enrolled in the college.
- Group 3: 23-33. This group is the "time for building and maintaining an initial mode of adult living." People within this age group are beginning their professional career, building the family, or getting prepared for their career by further graduate study.
- Group 4: 34-45. This is the phase of early adulthood to define a new era which belongs to them.
- Group 5: $>$ 46. This group include people from 46 to 65 who are in their middle adulthood and people who are older than 65 in the phase of the late adulthood.

Table 1: The summary of the datasets.

Datasets	#Users	#Age Groups	Age Group Distribution	#Tweets	#Edges (Avg.)	#Communities
Original	54,879	Group 1-5	[0.505, 0.376, 0.076, 0.022, 0.021]	51,756,652	58,267 (1.06)	19,978
Sampled	8,958	Group 1-3	[0.333, 0.333, 0.333]	8,567,085	1,263 (0.141)	7,743

Table 2: The performance on the original dataset.

	Average Accuracy			F-score for each group				
	Precision	Recall	F-score	Group 1	Group 2	Group 3	Group 4	Group 5
Content-I	0.7397	0.7538	0.7456	0.8246	0.7300	0.2451	0.1022	0.0301
Content-II	0.7435	0.7585	0.7495	0.8284	0.7349	0.2481	0.0670	0.0465
Neighbor-I ($f = 10$)	0.6677	0.6816	0.6557	0.7737	0.5883	0.0694	0.3281	0.1429
Neighbor-I ($f = 20$)	0.6886	0.7038	0.6809	0.7879	0.6318	0.0952	0.2642	0
Neighbor-II	0.4861	0.5021	0.4899	0.5589	0.4729	0	0	0
MAIF	**0.8069**	**0.8349**	**0.8138**	**0.9022**	**0.8196**	0.1795	0.0122	0.0441

Table 3: The performance on the sampled dataset.

	Average Accuracy			F-score for each group		
	Precision	Recall	F-score	Group 1	Group 2	Group 3
Content-I	0.5828	0.5829	0.5823	0.5661	0.5253	0.6557
Content-II	0.6930	0.6946	0.6935	0.6857	0.6221	0.7726
Neighbor-I ($f = 10$)	0.5606	0.5626	0.5073	0.6547	0.2125	0.6548
Neighbor-I ($f = 20$)	0.5721	0.5663	0.5078	0.6824	0.1930	0.6481
Neighbor-II	0.2392	0.3460	0.2506	0.2808	0.4787	0
MAIF	**0.7587**	**0.7611**	**0.7582**	**0.7572**	**0.6828**	**0.8347**

We use two datasets to evaluate the proposed framework, as shown in Table 1. First, the original dataset crawled in Section 3.1 is partitioned to five age groups as described above. Fig. 1 shows that the age distribution is highly biased toward Group 1 and 2, which occupy 88.06% of all the users. This is because the young people are more active in posting their birthday greetings to their friends. We first evaluate MAIF on this original dataset. Moreover, to evaluate the impact of the dataset bias, we build a comparable and balanced dataset as follows. We keep all the users in Group 3, which have 2,986 users, and then randomly sample the same number of users from both Group 1 and 2. After sampling, the network is less connected. Specifically, in the original dataset, each user has on average 1.06 friends within the dataset in contrast to 0.141 friends in the sampled dataset.

We use cross validation to evaluate the proposed framework. Specifically, given a ground-truth dataset composed of users who have indicated their ages, we split it into five subsets and conducted the experiment by five rounds. In each round, we choose four different subsets to build the classifier $\mathbf{W}$, then apply it to the remaining subset to estimate the users' ages, and finally compare them with the ground truth.

Since we aim to classify a user into $c(c > 2)$ groups, we derive both the *separate accuracy* for each group and the *overall accuracy* for all the groups from the *confusion matrix*[2]. For each age group i, we denote the number of true positives, false positives, true negatives, and false negatives by $\#\mathtt{TP}_i, \#\mathtt{FP}_i, \#\mathtt{TN}_i$, and $\#\mathtt{FN}_i$, respectively. Then we define the $\mathtt{Precision}_i, \mathtt{Recall}_i, \mathtt{F-score}_i$ as the separate accuracy for age group i as follows:

$$\mathtt{Precision}_i = \frac{\#\mathtt{TP}_i}{\#\mathtt{TP}_i + \#\mathtt{FP}_i};\ \ \mathtt{Recall}_i = \frac{\#\mathtt{TP}_i}{\#\mathtt{TP}_i + \#\mathtt{TN}_i};$$
$$\mathtt{F-Score}_i = \frac{2 \times \mathtt{Precision}_i \times \mathtt{Recall}_i}{\mathtt{Precision}_i + \mathtt{Recall}_i}. \tag{20}$$

We then define the overall accuracy as $\mathtt{X} = \sum_{i=1}^{c} r_i \mathtt{X}_i$, where $\mathtt{X}$ represents $\mathtt{Precision}$, $\mathtt{Recall}$, or $\mathtt{F-score}$, and r_i is the ratio of users in age group i over the whole dataset, which is listed as the age distribution in Table 1.

4.2 Assessing Accuracy

We first evaluate the accuracy of the proposed framework and compare it with both the state-of-the-art methods and the baseline methods summarized as follows.

- Content-based methods I. The state-of-the-art content-based method is proposed by (Nguyen et al. 2013) to use the linear regression model with the ℓ_2 regularization, which is equivalent to adding $\|\mathbf{W}\|_F$ to the least square method defined in Eq. (2). We use the top-10000 1-gram and 2-gram as the features to infer the age information.
- Content-based methods II with sparse representation. We use the least square method with the ℓ_1 regularization in

[2]Here we didn't use the confusion matrix directly because it is not efficient to compare the MAIF with several baseline methods. However, the derived separate and overall accuracy can represent well the confusion matrix.

Eq. (3) to evaluate the inference performance by including the sparse representation for the content information.

- Neighbor-based method I. We infer the age information from neighbors' content information as used in (Zamal, Liu, and Ruths 2012; Chen et al. 2015). Specifically, for each user i, we use the least square method with the ℓ_1 regularization in Eq. (3) to estimate the age information of i's f friends who are not in the labeled user set, and then set the average value as i's age information. In the experiment, we set f be 10 and 20.
- Neighbor-based method II. We infer the age information from labeled neighbors' age information as used in (Dey et al. 2012; Liao et al. 2014). Specifically, we implement the more advanced method in (Liao et al. 2014) which assigns a weight between every friend pair in the labeled user set and then uses the label propagation to estimate the unknown users' ages. We use 80% of the users as the training set and the remaining as the testing set.
- The proposed framework. We set both λ_1 and λ_2 in the Eq. (11) to be 1 for the general experiment, and we will explore the effects of parameters later. Moreover, we set the size of the feature space $m = 10,000$ with 5,000 of 1-grams and 2-grams each and the termination condition $\epsilon = 10^{-4}$ in Alg. 1.

For each method, we compare the separate accuracy of each group and the overall accuracy, as shown in Table 2. We could draw three conclusions from the overall accuracy. First, the proposed MAIF is better than all other four methods, verifying that our framework can accurately integrate the content information and the interaction information, which are the essential behavior pattern of twitterers, to infer the age information. Second, the sparse representation in Content-II method outperforms the least square in Content-I, meaning that the content information in microblogging services is indeed sparse, and that the sparse features could represent age groups more accurately. Third, directly inferring the age information from labeled neighbors' age information as in Neighbor-II method is not effective for the dataset. The reason is that the age information in Twitter is so scarce that many users lack the neighbors who have specified their ages. As we can see from Table 1, the average friends in the original dataset is 1.06, meaning that every labeled user only has average one friend in the dataset. To overcome this issue, MAIF leverages the community structure which contains more users and integrates it with the content information.

As for the separate accuracy, we could see that although MAIF outperforms other methods in Group 1 and 2, all methods have low accuracy for the remaining groups. We conjectured that the low accuracy for Group 3 to 5 is caused by the bias of the dataset. To verify this conjecture, we applied these five methods on the sampled dataset described in Table 1 and obtained the results in Table 3. The results show that MAIF outperforms all other four methods in both average and separate accuracy. Moreover, the accuracy of MAIF for each age group on this dataset is significantly balanced than on the original dataset, which justifies our conjecture and also answers the second question stated in the beginning of this section.

4.3 Performance of the Content and Interaction information

Since the proposed MAIF framework explores content and also interaction information, we aim to investigate the contribution of each type and the benefit of the integration. Specifically, we consider the following methods.

- Content-only methods. We use both the widely-used Support Vector Machine (SVM) (Suykens and Vandewalle 1999) and the least square with the ℓ_1 regularization in Eq. (3) to evaluate the performance on the content feature matrix $\mathbf{X}$.
- Network-only methods. We use the adjacent matrix $\mathbf{G}$ as the feature matrix and then apply both SVM and the least square with the ℓ_1 regularization in Eq. (3) on them.

Fig. 5 shows the overall accuracy of content-only methods, network-only methods, and the proposed framework. As we can see, MAIF outperforms both content-only and network-only methods, meaning that the accuracy will increase if we integrate both content and interaction information instead of considering only one type of information. Moreover, content-only methods perform better than network-only methods, indicating that content information is more reliable and contributes more than the network information to infer the age information. Again, we conjecture that this is caused by the scarcity of the age information in Twitter and hence the scarcity of the network information between the labeled users.

4.4 Exploiting the Parameters

We show the impact of λ_1 and λ_2 in Fig. 6. As stated before, the parameter λ_1 indicates the weight of the sparse representation of the content information, and the parameter λ_2 indicates the weight of interaction information. As we can see, both parameters lead to smooth and meaningful results when they are between 0.01 and 100. Moreover, when λ_2 increases from 0.01 to 100, the overall accuracy first increases and then decreases; similar trend holds for λ_2. Hence we expect a local optimal parameter pair for λ_1 and λ_2 to be both 1. Moreover, the accuracy in this range is smooth and has limited variance, suggesting that in practice we could choose these two parameters from 0.01 to 100.

Fig. 7 show the impact of the dataset size. Given the original dataset with 54,879 users, we randomly sample 5000, 10000, 20000, 30000, 40000, and 50000 users, respectively, and use five-fold cross validation to yield the results. As we can see, increasing the dataset size improves the accuracy because of the richer content and network information. Specifically, the more users, the richer content and network information which could better represent the users in a specific age group, and hence the higher accuracy.

Our algorithm is stable in terms of other parameters. Fig. 8 demonstrate the impact of the training set size. 50% of the training set means that only 50% of the 80%, which equals 40%, of the users in the whole dataset are used as the

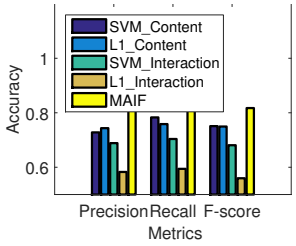


Figure 5: The performance of separate information.

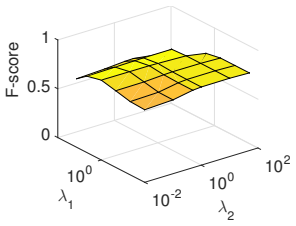


Figure 6: The impact of the parameters λ_1 and λ_2.

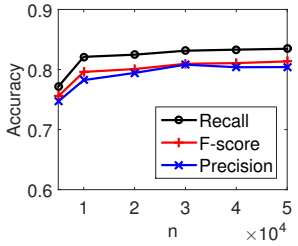


Figure 7: The accuracy under different dataset sizes.

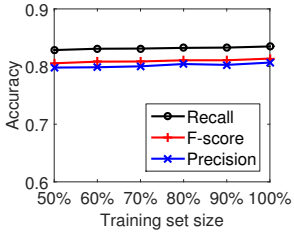


Figure 8: The accuracy under different training set sizes.

training set to predict the remaining 20% of users. The training set size varies from 50% to 100% (corresponds to 40% to 80% of all users). As expected, increasing the training set can slightly improve the accuracy. However, the improvement is less significant than increasing the dataset size, as shown in Fig. 7. The reason is that the training set, which has 40% of the 54,879 users, is still large enough to obtain the good results even if we only use 50% of the training set. Moreover, we have evaluated the impact of feature space size m (from 5000 to 20000), the different combinations of 1-gram and 2-gram in the feature space, and the variant definitions of TF-IDF in Eq. (1), and obtained similar results, which are omitted here due to space constraints.

4.5 Countermeasures

The above experiments demonstrate the efficacy of using the public content and interaction information to infer the age information, one type of highly-private personal attributes. Since the social homophily and generation gap always exist, it is very challenging for twitterers to evade the inference demonstrated in this paper. In order to preserve their privacy, one way is to set "protected"—a function provided by Twitter—the critical information such as followers, followees, and even all the tweets, such that only authorized users could visit while the unauthorized third party will fail to infer due to the absence of content and interaction information. Another way is to diversify the content and/or interaction information by posting with the style from other age groups and/or following people with different ages. As shown in Fig. 5, the absence of one type of information will lower the inference performance, so the age privacy can be protected to some extent. Nevertheless, this paper mainly aims to demonstrate a more effective method to infer the hidden age information from public content and interaction information in Twitter. More privacy implications of our framework and the thorough investigation of countermeasures are beyond the scope of this paper.

5 Related Work

In this section, we briefly present the existing work mostly related to this paper.

There has been some effort to infer hidden age information in microblogging systems. Nguyen *et al.* tried to classify the user ages from different angles such as age range, exact age, and life stage with the 1-grams constructed from the tweets (Nguyen et al. 2013). Oktay *et al.* proposed a method to infer users' age range by investigating their names. The idea is that different generations have different preferences on the baby naming (Oktay, Firat, and Ertem 2014). Liao *et al.* use the ages of online neighbors to infer the age of a given user (Liao et al. 2014). Dey *et al.* also used the similar method to infer the user age in Facebook (Dey et al. 2012). However, this method requires that some neighbors have specified their ages, which cannot be satisfied in microblogging systems where age information is scarce.

Other hidden attributes such as location (Li et al. 2012; Mahmud, Nichols, and Drews 2014; Compton, Jurgens, and Allen 2014; Zhang et al. 2015b), gender (Rao et al. 2010), political preference (Zamal, Liu, and Ruths 2012), and ethnicity (Chen et al. 2015) have also been inferred by either the content information and/or the interaction information. Mislove *et al.* used the local connections around the Facebook users to infer their major, college, and political view (Mislove et al. 2010). Location information has attracted many attentions recently. The content with geographical hints could be used to infer users' locations (Mahmud, Nichols, and Drews 2014). Since about 16% of Twitter users have specified their locations, inferring users' locations from their neighbors' locations can be more effective (Li et al. 2012; Compton, Jurgens, and Allen 2014) than inferring their ages from their neighbors' ages. Besides the privacy threats by inferring these sensitive attributes, there have been growing security issues against social network users (Boshmaf et al. 2011; Zhang et al. 2013; 2016).

6 Conclusion

In this paper, we propose MAIF, a novel framework which explores public content and interaction information in microblogging systems to infer the hidden ages of microbloggers. We thoroughly evaluate MAIF using a real-world dataset with 54,879 Twitter users. Our results show that MAIF can achieve up to 81.38% inference accuracy and outperforms the state of the art by 9.15%. In our future work, we seek to incorporate more interaction information such as retweets, replies, and mentions into MAIF. In addition, we will evaluate the performance of MAIF for other microblogging systems such as Tumblr. Finally, we plan to thoroughly investigate the privacy implications of MAIF and possible countermeasures for the microbloggers particularly wary of their age privacy.

Acknowledgments

We truly appreciate the anonymous reviewers for their constructive comments. This work is supported by ARO through W911NF-15-1-0328.

References

Baraniuk, R. 2007. Compressive sensing. *IEEE signal processing magazine* 24(4).

Bertsekas, D. 1999. *Nonlinear programming*. Athena scientific.

Blondel, V.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10):P10008.

Boshmaf, Y.; Muslukhov, I.; Beznosov, K.; and Ripeanu, M. 2011. The socialbot network: when bots socialize for fame and money. In *ACSAC'11*, 93–102.

Boyd, S., and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press. chapter 8, Geometric Problems, 411–413.

Chen, X.; Wang, Y.; Agichtein, E.; and Wang, F. 2015. A comparative study of demographic attribute inference in twitter. In *ICWSM'15*.

Compton, R.; Jurgens, D.; and Allen, D. 2014. Geotagging one hundred million twitter accounts with total variation minimization. In *IEEE Big Data'14*.

Dey, R.; Tang, C.; Ross, K.; and Saxena, N. 2012. Estimating age privacy leakage in online social networks. In *INFOCOM'12*.

Duggan, M.; Ellison, N.; Lampe, C.; Lenhart, A.; and Madden, M. 2015. Demographics of key social networking platforms.

Giancola, F. 2006. The generation gap: More myth than reality. *People and strategy* 29(4):32.

Hu, X.; Sun, N.; Zhang, C.; and Chua, T.-S. 2009. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *CIKM'09*.

Ji, S., and Ye, J. 2009. An accelerated gradient method for trace norm minimization. In *ICML'09*.

Kumar, S.; Morstatter, F.; and Liu, H. 2013. *Twitter Data Analytics*. New York, NY, USA: Springer.

Lawson, C., and Hanson, R. 1974. *Solving least squares problems*, volume 161. SIAM.

Leskovec, J.; Rajaraman, A.; and Ullman, J. 2014. *Mining Massive Datasets*. Cambridge University Press. chapter Data Mining, 7–9.

Levinson, D. 1986. A conception of adult development. *American psychologist* 41(1):3–13.

Li, R.; Wang, S.; Deng, H.; Wang, R.; and Chang, K. 2012. Towards social user profiling: Unified and discriminative influence model for inferring home locations. In *KDD'12*.

Liao, L.; Jiang, J.; Lim, E.-P.; and Huang, H. 2014. A study of age gaps between online friends. In *HT'14*.

Liu, J.; Ji, S.; and Ye, J. 2009. Multi-task feature learning via efficient l2,1-norm minimization. In *UAI'09*.

Mahmud, J.; Nichols, J.; and Drews, C. 2014. Home location identification of Twitter users. *ACM Trans. Intell. Syst. Technol.* 5(3):47:1–47:21.

Marinis, T. 2003. Psycholinguistic techniques in second language acquisition research. *Second Language Research* 19(2):144–161.

Mislove, A.; Viswanath, B.; Gummadi, K.; and Druschel, P. 2010. You are who you know: inferring user profiles in online social networks. In *WSDM'10*.

Newman, M., and Girvan, M. 2004. Finding and evaluating community structure in networks. *Physical review E* 69(2):026113.

Nguyen, D.; Gravel, R.; Trieschnigg, D.; and Meder, T. 2013. "how old do you think i am?"; a study of language and age in twitter. In *ICWSM'13*.

Oktay, H.; Firat, A.; and Ertem, Z. 2014. Demographic breakdown of twitter users: An analysis based on names. In *ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference*.

Porter, M. 1997. *Readings in information retrieval*. Morgan Kaufmann Publishers Inc. chapter An algorithm for suffix stripping, 313–316.

Rao, D.; Yarowsky, D.; Shreevats, A.; and Gupta, M. 2010. Classifying latent user attributes in twitter. In *SMUC'10*.

Suykens, J., and Vandewalle, J. 1999. Least squares support vector machine classifiers. *Neural processing letters* 9(3):293–300.

Tang, J., and Liu, H. 2012. Unsupervised feature selection for linked social media data. In *KDD'12*.

Tibshirani, R. 2011. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(3):273–282.

Zafarani, R.; Abbasi, M. A.; and Liu, H. 2014. *Social Media Mining: An Introduction*. Cambridge University Press.

Zamal, F.; Liu, W.; and Ruths, D. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM'12*.

Zhang, J.; Zhang, R.; Zhang, Y.; and Yan, G. 2013. On the impact of social botnets for spam distribution and digital-influence manipulation. In *IEEE CNS'13*, 46–54.

Zhang, J.; Zhang, R.; Sun, J.; Zhang, Y.; and Zhang, C. 2015a. Truetop: A sybil-resilient system for user influence measurement on twitter. *IEEE/ACM Transactions on Networking* PP(99):1–1.

Zhang, J.; Sun, J.; Zhang, R.; and Zhang, Y. 2015b. Your actions tell where you are: Uncovering twitter users in a metropolitan area. In *IEEE CNS'15*, 424–432.

Zhang, J.; Zhang, R.; Zhang, Y.; and Yan, G. 2016. The rise of social botnets: Attacks and countermeasures. *ArXiv*.