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 $\#\texttt{TP}_i, \#\texttt{FP}_i, \#\texttt{TN}_i$, and $\#\texttt{FN}_i$, respectively. Then we define the $\texttt{Precision}_i, \texttt{Recall}_i, \texttt{F}-\texttt{score}_i$ as the separate accuracy for age group i as follows:

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

We then define the overall accuracy as $\texttt{X} = \sum_{i=1}^{c} r_i \texttt{X}_i$, where $\texttt{X}$ represents $\texttt{Precision}, \texttt{Recall}$, or $\texttt{F}-\texttt{score}$, 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.