Breaking Anonymity of Social Network Accounts by Using Coordinated and Extensible Classifiers Based on Machine Learning

Eina Hashimoto, Masatsugu Ichino, Tetsuji Kuboyama, Isao Echizen, Hiroshi Yoshiura

Breaking Anonymity of Social Network Accounts by Using Coordinated and Extensible Classifiers based on Machine Learning

Eina Hashimoto[1], Masatsugu Ichino[1], Tetsuji Kuboyama[2], Isao Echizen[3], Hiroshi Yoshiura[1]

[1] University of Electro-Communications, Tokyo, Japan
{e-hashimoto, yoshiura}@uec.ac.jp
[2] Gakushuin University, Tokyo, Japan
[3] National Institute of Informatics, Tokyo Japan

Abstract. A method for de-anonymizing social network accounts is presented to clarify the privacy risks of such accounts as well as to deter their misuse such as by posting copyrighted, offensive, or bullying contents. In contrast to previous de-anonymization methods, which link accounts to other accounts, the presented method links accounts to resumes, which directly represent identities. The difficulty in using machine learning for de-anonymization, i.e. preparing positive examples of training data, is overcome by decomposing the learning problem into subproblems for which training data can be harvested from the Internet. Evaluation using 3 learning algorithms, 2 kinds of sentence features, 238 learned classifiers, 2 methods for fusing scores from the classifiers, and 30 volunteers' accounts and resumes demonstrated that the proposed method is effective. Because the training data are harvested from the Internet, the more information that is available on the Internet, the greater the effectiveness of the presented method.

Keywords: social network, privacy, de-anonymization, re-identification

1 Introduction

Online social networks enrich human communication. They are used not only for communication among friends and family members but also for job hunting, marketing, branding, and political communication such as among political activists. On the other hand, they can reveal personal information and cause privacy problems. They can also reveal confidential information and enable posting of copyrighted, offensive, or bullying contents.

To mitigate the privacy problems, social network services provide mechanisms that enable users to limit the disclosure of posted content (text, photos, etc.) to friends, followers, etc. However, because defining an appropriate disclosure range for each post is cumbersome [1], users tend to use the same range for all their posts, resulting in too much disclosure for sensitive content and/or unnecessarily limited disclosure for less sensitive content. Furthermore, disclosure by friends and followers, such as retweets, is a big loophole in disclosure control. Another approach to privacy protec-

tion is anonymizing social network accounts. Users omit, change, or obscure identifying and pseudo-identifying information, such as name, age, address, affiliation, face, in their posts and profiles so that only friends can recognize the poster. Such anonymization is widely used in Japanese social networks for example [2].

The anonymization approach can be compromised, however, by linking an anonymized account to an account in another social network. For example, Narayanan and Shmatikov showed that accounts in two social networks used by the same person can be identified by finding similar social graphs in the two networks [3]. Goga et al. also identified accounts used by the same person by comparing location information and time stamps attached to posts and writing styles [4]. Almishari et al. and Narayanan et al. also pursued the same objective by using machine learning to compare writing styles [5] [6]. Their methods, however, are indirect because they simply link accounts and/or blogs—knowing that account-1 and account-2 are used by the same person does not directly reveal the person's identity.

In contrast, we have developed a method that links a social network account to a resume, which directly represents a person. Given social network accounts and resumes, the method matches accounts to resumes. Because most organizations, e.g. companies, universities, and public institutions, have resumes or resume-like information for their members, and governments have similar information on residents, the proposed method has generality.

Our research thus clarifies a serious privacy risk; that is, persons of concern to organizations and governments can be identified and their freedom of speech can be suppressed. Besides clarifying a privacy risk, the proposed method can be used for protective purposes. It can be used to identify a person in an organization who misuses a social network (e.g. by revealing confidential information or posting copyrighted contents). It does this by linking the misused account to a candidate resume.

Although our method uses machine learning as did previous research, we encountered a difficulty in preparing training data that did not arise in the previous research. Almishari's method, for example, uses a naïve Bayes classifier to learn writing styles of texts posted from one account [5]. It then identifies texts posted from another account that has similar writing styles, and that account is considered to probably be used by the same person. The training data for Almishari's method are texts posted from the first account, which are not difficult to obtain. The training data for Narayanan's method, which identifies blogs posted by the same person, are not difficult to obtain either [6]. Preparing training data for these methods is not difficult because the linking falls into a particular pattern, i.e. learning features of texts and identifying other texts having similar features. However, preparing training data is not easy if the linking falls outside this pattern.

Our problem of linking an account to a resume does not fall into the pattern. Our method could learn writing styles of texts posted from the account but cannot identify a resume by using the learned writing styles. This is because a resume is not conventional text consisting of sentences but a list of keywords that represents characteristics of the person.

To overcome this difficulty, we use machine learning to implement a component classifier for each characteristic described in the resume, e.g. a component classifier

for whether a social network account is used by a person whose hobby is dancing and one for whether the account is used by a person who is a computer engineer. We then compose a classifier for the resume itself by combining these component classifiers and use this classifier to determine whether an account is used by the resume owner. We can search the Internet for social network accounts that have specific characteristics (hobby of dancing) and use the text from them as training data for learning the component classifiers.

This work makes three contributions to social network privacy.

(1) In contrast to previous methods, our proposed method links social network accounts directly to identities by linking them to resumes, which are held by most organizations and governments. It revealed a privacy risk more serious than that revealed in previous research and can be widely used to deter misuse of social networks.

(2) We overcome a difficulty in preparing training data, which most previous research did not encounter, by decomposing the learning problem into subproblems for which we can harvest training data from the Internet.

(3) The greater the amount of information available from the Internet, the greater the amount of training data, which makes our proposed method more effective.

2 Related work

Much work has been done on extracting personal information from social networks. Earlier work mainly focused on estimating users' sensitive information by using keyword and graph matching with heuristic algorithms. In 2007, Backstrom et al. de-anonymized anonymous social network accounts by searching the social network for subgraphs of known human relationships and identifying the subgraphs' nodes that represented users and friends [7]. In 2008, Lam et al. correctly estimated the first names of 72% of the users of a social network and the full names of 30% of the users by keyword-matching analysis of comments from friends [8]. In 2011, Mao et al. identified tweets containing sensitive information about travel and medical conditions with 76% precision and tweets posted under drinking with 84% precision by using learning algorithms of naïve Bayes and support vector machine (SVM) [9]. The training data were tweets that had been labelled by hand as either sensitive or non-sensitive. In 2012, Kótyuk and Buttyan estimated age, gender, and marital status, which were not disclosed in the user profiles, from disclosed parts of the profiles, friend information, and user group memberships by using learning algorithms of neural networks [10]. In 2014, Caliskan-Islam et al. used naïve Bayes and AdaBoost to classify users into three levels of revealing private information [13].

Recent related work has generally focused on linking a target account or post with another account or post. In 2009, Narayanan et al. reported a linking method based on subgraph matching that had been used to link the Twitter and Flicker accounts of the same users with an error rate of 12% [3]. In 2010, Polakis et al. reported a method for linking the names of social network users to their e-mail addresses [11] and used it to match 43% of the user profiles extracted from Facebook to the user e-mail addresses.

In 2012, Goga et al. proposed a method for identifying users who used different social networks (Yelp, Twitter, Flickr, and Twitter) by analysing and combining the features of geo-location, timestamp, and writing styles from their posts [4]. In the same year, Narayanan used several machine learning algorithms including SVM and linear discriminant analysis to identify blogs posted by the same person [6]. In 2014, as mentioned above, Almishari et al. used a naïve Bayes classifier to identify Twitter accounts used by the same person [5].

3 Linking Social Network Account to Resume

3.1 Representative Application

Given social network accounts and resumes, our method identifies pairs of matching accounts and resumes, thus linking accounts to resumes, which represent identities. A representative application of our method is use by a company that finds that posts from an anonymous social network account include objectionable content such as content criticizing the company or exposing company wrongdoing. The company determines whether the account belongs to an employee by calculating the linkability between the account and each resume it holds and assuming the most linkable resume probably represents the target person, whom the company may punish.

Note that a company obtains a person's resume when the person joins the company and maintains it. Additional information about salary, promotions, changes in job, family members, addresses, etc. are collected over time. Here we refer to all this information simply as "resume".

3.2 Difficulty in Using Machine Learning

One of the biggest challenges in using machine learning is preparing the training data because the effectiveness of the learning critically depends on that data. As mentioned in Section 1, previous methods, which link accounts and/or blogs, learn writing styles of texts (posted from an account or included in a set of blogs) and identify texts posted from another account or included in another set of blogs that have similar writing styles [5][6]. Training data for these methods are text at hand and are not difficult to prepare. In the method proposed by Kótyuk and Buttyan, learned correlations between disclosed attributes (age, gender, marital status, number of friends, language used, etc.) are used to infer undisclosed attributes [10]. The training data for this method are attributes disclosed in profiles and texts on social networks and are not difficult to obtain.

However preparing training data is not always that easy. Mao used known sensitive and non-sensitive tweets as positive and negative examples of training data. Because these training data are manually labelled "sensitive" or "non-sensitive" [9], preparing the training data is time consuming. Mao's method therefore does not work on a large scale and requires manual preparation of training data whenever it is used for new kinds of sensitive tweets (ones related to income, addresses, drug use, etc.). Caliskan-

Islam et al. mitigated this problem by socially outsourcing the labelling task [13]. They did not solve the problem, however, because the time and effort needed were not reduced but simply shifted from the researchers to outsourced workers. Hart et al. used corpora for training data [14], but time and effort are needed to prepare such corpora.

Preparing training data is much more difficult for our problem in which an anonymized social network account is linked to a resume. Our method could learn writing styles of texts posted from the account but cannot identify a resume by using the learned writing styles because a resume is not a conventional text consisting of sentences. The use of outsourced workers is not an option because the training data could not be labelled by such workers. We overcome this difficulty as described in the next section.

3.3 Our Method Using Machine Learning

A resume consists of pairs of attributes and attribute values, for example, gender = female, current address = "Chofu city, Tokyo", hometown address = Osaka, affiliation = Company A, educational history = "Ph.D. from Tokyo Univ. in 2000, Master's degree from Kyoto Univ. in 1997, etc.", and hobbies = "dancing, painting". The attribute values represent the characteristics of the owner of the resume. We use machine learning to implement a component classifier for each attribute value. For example, we implement a classifier for determining whether a social network account is used by a woman[1] , one for determining whether the account is used by a person from Osaka (based on Osaka dialect), and one for a person with dancing as a hobby.

We then compose a classifier for the resume itself by combining the component classifiers for all the attribute values on the resume. This classifier is used to determine whether an account is used by the owner of the resume, i.e. a person having all attribute values on the resume. The number of component classifiers used for the resume classifier depends on the number of attribute values on the resume. The score for the resume classifier is the aggregation of the scores of the component classifiers.

Given social network accounts and resumes, our method identifies matching accounts and resumes as follows (Fig. 1).

(1) Implement component classifiers for all attribute values on resumes.

(2) Compose a classifier for each resume.

(3) Obtain text posted on each social network account.

(4) Input the text from each account into the classifier for each resume. Then output the classifier score for each resume for each account. Each score represents how likely the account is used by the resume owner.

(5) For each account, the resume with the highest classifier score is selected. The selected resume is assumed to represent the account owner.

Effective component classifiers can be implemented for gender and address attributes as shown in [15] [16]. It may also be possible to implement component classifiers

[1] More precisely, the classifier determines whether text posted on a social network account were written by a woman.

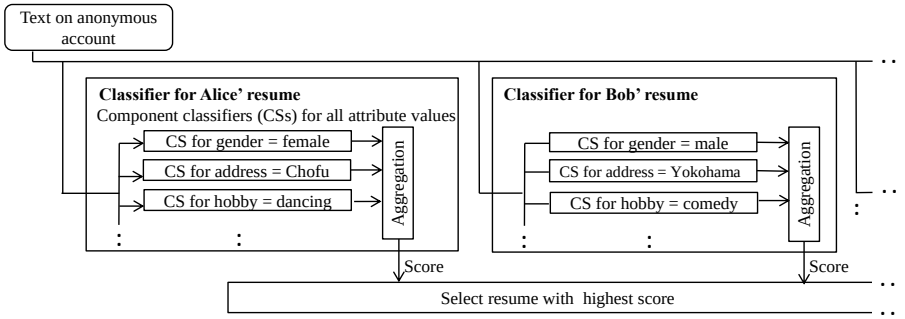

Fig. 1. Proposed model for de-anonymization

for other attributes as did Pennacchiotti et al. for political affiliation, ethnicity, and coffee brand preference [17]. Collecting positive examples of training data is automatized by using a tool such as TwiPro [12], which searches the Internet for social network accounts for which the user profile includes a given attribute value (e.g. hobby = dancing). This search works for most attribute values though the tool cannot collect a sufficient number of accounts for unusual attribute values such as "hobby = cooking eel". Collecting negative examples of training data is easier—the same tool is used to search for social network accounts for which the user profile does not include a given attribute value.

4 Data Description

4.1 Sample Data from Volunteers

Hereafter we abbreviate "social network account" as "account". We obtained Twitter accounts and resumes from 30 volunteers attending our university. Table 1 shows their demographics. The tweets and resumes were originally written in Japanese and are translated into English here.

The volunteer resumes included 12 attributes such as name, birthdate, gender, current address, hometown address, educational history, and qualifications. These attributes were selected in accordance with the Japanese standard for resumes of students' seeking jobs. They do not include job history or family structure (marital status, children, etc.) because students in Japan usually do not have job histories and are not married.

Of these 12 attributes, we used 7 in our experiment: (1) gender, (2) current address, (3) hometown address, (4) educational history, (5) favourite subject, (6) hobbies, and (7) qualifications. Because educational history is generally complex, we simply used the departments in which the volunteers were studying as representative information.

We also obtained access to the Twitter accounts of the 30 volunteers and to their tweets. The number of tweets obtained from each account ranged from 2167 to 3000 (2771 on average). All of the account profiles and tweets were anonymized by the

Table 1. Volunteer demographics

(a) Gender		(b) Age				(c) School year			(d) Department		
Male	Female	20	21	22	23	2nd	3rd	4th	Informatics	Electronics	Machanical engineering
20	10	10	14	5	1	4	21	5	24	2	4

volunteers themselves, who omitted, changed, or obscured identifying and pseudo-identifying information. In the evaluation described in Section 6, the number of tweets for the test data was 2771 (all tweets), 1000, 250, 60, or 15 per account. Among the previous methods mentioned in Section 2, the method of Almishari et al. [5] is most similar to ours because it uses tweets for linking but is different in that it matches writing styles while our method matches attribute values. Almishari et al. used 100, 50, 20, 10, or 5 tweets per account for their test data. We used more test data because attribute values (e.g. dancing as a hobby) do not often appear in tweets while writing style can be observed in a few tweets.

For these data, the sample problem in our experiment was to match the 30 Twitter accounts to the 30 resumes. Although this is a small problem, it was sufficiently difficult to evaluate our method. We therefore used it for an initial evaluation. The problem is difficult because the resumes were very similar, so the classifiers were provided with little information. For example, all the volunteers were undergraduate students at the same university and were in one of three departments (informatics, electronics, or mechanical engineering), which are in neighbouring buildings. Their current addresses are close to the university and close to each other. The Informatics and Electronics Departments share many subjects such as computer architecture, programming, and signal processing. The Electronics and Mechanical Engineering Departments also share many subjects, and the Informatics and Mechanical Engineering Departments share some basic subjects such as physics. The volunteers therefore had similar educational experiences. Their daily schedules were also similar. They were similar in age and school year as well, and none of them were married or had job histories. .

The problem derived from the representative application described in Section 3.1, i.e. a company is to identify an employee of concern, is larger in scale but is probably easier to solve. The resumes of employees include much more information because employees are different in terms of job history, position, salary, and family structure while the resumes of the student volunteers did not include such information at all. Employees have different daily schedule depending on their job and more qualifications than students. Their ages have a wider range, and their addresses vary greatly if they work in different parts of the company that are in different geographic areas.

4.2 Training Data

We obtained training data by using TwiPro [12], as mentioned in Section 3.3. For positive examples of training data, we collected tweets from 30 random Twitter accounts, each having more than 1000 posted tweets, and used up to the latest 3000 tweets from each account. For the attribute values on a resume for which we could not

collect data from the 30 accounts, we used data from as many accounts as possible as long as we could collect data from at least ten accounts. Otherwise, we did not implement a classifier for that attribute value. Negative examples of training data were similarly prepared.

5 Preliminary Experiment

We carried out a preliminary experiment to identify the attributes in the resumes most effective for the linking and the sentence features that should be extracted from tweets as well as to test the machine learning algorithms and methods for aggregating the component classifier scores. We evaluated all attributes on the resumes and evaluated bag-of-words (frequency of words appearing in tweets) and binary (appearance or non-appearance of words) models for feature extraction. Random Forest, linear SVM, and logistic regression were tested as the learning algorithm for component classifiers.

When texts from M accounts are input into N component classifiers, M vectors consisting of N scores are output, each of which represents an account. Machine learning could also be used to generate resume classifiers that classify these N-dimensional vectors in accordance with the resumes. However the implementation of such learning needs more research because training data are sparse in a high dimensional learning space[2]. We therefore used simple score fusion methods to generate resume classifiers for the experiments described here. That is, we used the score average and score product from the component classifiers to clarify the viability of our approach. The use of machine learning algorithms (e.g. SVM, Random Forest, and boosting) will be studied in future work.

5.1 Sample Data

In the preliminary experiment, we used sample data for three of the female and three of the male volunteers. Table 2 shows the attributes and attribute values extracted from their resumes (city names have been anonymized for privacy). There were 7 attributes and 46 unique attribute values. We implemented only 40 component classifiers as we could not obtain a sufficient number of positive examples of training data for 6 of them (the underlined values). We used all tweets of the 6 volunteers for the test data.

5.2 Calibration

The scores for the component classifiers were calibrated using the following formula before fusion by averaging. We do not explain the rationale for using this

[2] For example, there are 119 attribute values in the 30 sample resumes mentioned in Section 4.1, so we have only 30 samples in 119-dimensional space.

Table 2. Attributes and values used for preliminary experiment

Volunteer	Gender	Current address	Hometown address	Department	Favourite subjects	Hobbies	Qualifications
1	F	City A, Kanagawa	City E, Saitama	Informatics	Programming	Comedian, Audrey, eye glasses	Driving license
2	F	City A, Kanagawa	County F, Nagano	Informatics	German	Playing piano, piano circle	Driving license
3	F	City B, Tokyo	City B, Tokyo	Electronics	Physical exercise, music, mathematics	Martial arts of Aikido, basketball, music, reading, baking sweets	Driving license
4	M	City C, Kanagawa	City C, Kanagawa	Mechanical engineering	Plastic & cutting processing	Robot mechatronics, engineering circle	Driving license
5	M	City D, Tokyo	City G, Hokkaido	Informatics	Art	Tennis, futsal, watching TV, football	Driving license, sales representative, financial planner
6	M	City D, Tokyo	City H, Aomori	Electronics	Electronic circuits, electromagnetics, web design, programming	Baseball, baseball circle, watching social network	–

formula because it is standard in data analysis and other researchers of de-anonymization (e.g. Narayanan [6]) have used it.

$$\alpha_{ij} = \frac{x_{ij} - \overline{x_j}}{\sigma_j} \tag{1}$$

where M and N are the number of accounts and number of component classifiers, respectively. They were set to 6 and 40 for the preliminary experiment. The x_{ij} is the original score of the j-th component classifier calculated for the i-th account, where 1<=i<=M and 1<=j<=N. Note that the j-the component classifier was implemented with respect to the j-th attribute value. The α_{ij} is the calibrated value of x_{ij}, and $\overline{x_j}$ and σ_j are, respectively, the average and standard deviation of x_{ij} over 1<=i<=M.

5.3 Results

Figure 2 shows the distribution of scores for the component classifiers with the bag-of-words model used for the sentence features and Random Forest used as the learning algorithm. The horizontal axis represents the component classifier for each attribute value. The vertical axis represents the value of the calibrated scores. The distribution of the M scores calculated using a classifier is represented by a box, lines above and below the box, and dots. The left most ones, for example, represent the score distribution of the classifier for "current address = City A in Kanagawa". The box represents the scores between the lower and upper quartiles of the distribution (i.e. 50% of the scores). The two lines above and below the box represent the top and bottom 25% of the scores, and the two dots represent the scores for the two accounts belonging to the two volunteers who actually live at this address. Thus, the higher the dots, the more correct the classifier.

Table 3 shows the rankings of the accounts that actually had the corresponding attributes (shown by dots in Fig. 2). The rankings were averaged over each sentence feature, algorithm, and attribute. For example, the average of the six classifier scores represented by dots for the current address attribute in Fig. 2 was 3.83, which is shown in the corresponding (upper-left) cell in Table 3. The smaller the value of the

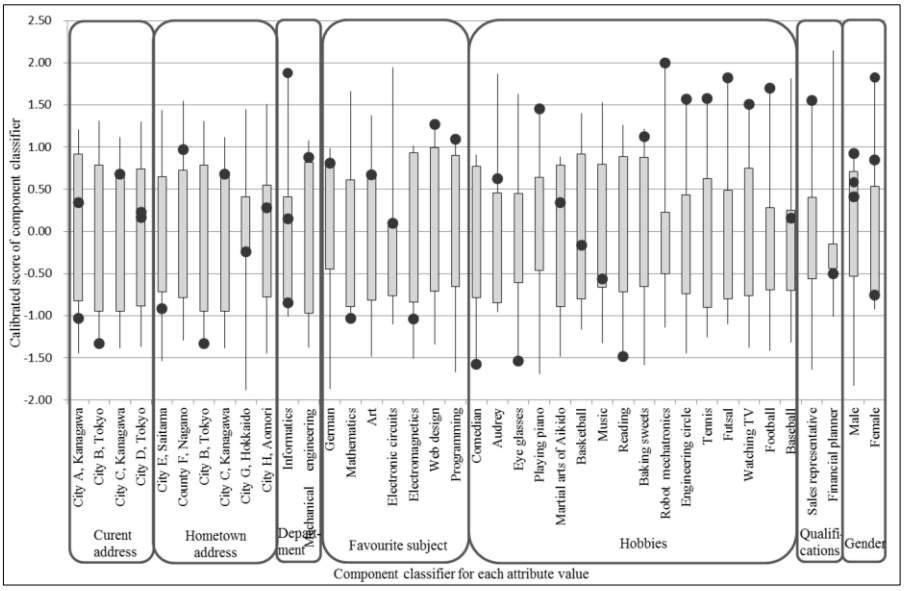

Fig. 2. Score distribution of component classifiers with bag-of-words and Random Forest

Table 3. Ranking of accounts that actually had corresponding attributes

Feature	Learning algorithm	Current address	Hometown address	Department	Favourite subjects	Hobbies	Qualifications	Gender	Average over all attributes	Average over 4 attributes
Bag-of-words	Random Forest	3.83	3.83	2.75	2.86	2.75	3.00	2.67	3.17	2.76
	Linear SVM	3.50	2.83	3.00	1.86	2.88	2.00	2.00	2.68	2.43
	Logistic regression	3.33	3.00	3.00	2.14	2.69	2.00	2.00	2.69	2.46
Binary	Random Forest	3.50	2.67	4.00	3.43	3.13	4.50	3.50	3.54	3.51
	Linear SVM	3.17	3.67	2.75	3.14	4.06	2.00	4.00	3.13	3.49
	Logistic regression	3.83	3.50	4.00	3.86	3.13	5.50	2.83	3.97	3.45

average ranking, the more correct the score of the component classifier. Note that the expected value for ranking is 3.5 because there are six possible rankings (1 through 6). From Table 3, we can see that bag-of-word was a better model than binary and, when we focus on rankings in bag-of-word model, we can see that the attributes most effective for de-anonymization were department, favourite subject, hobbies, and gender.

We therefore considered 12 cases: one of the three learning algorithms (Random Forest, linear SVM, or logistic regression), all attributes or the four most effective attributes (department, favourite subject, hobbies, and gender), and fusion by average or by product with the bag-of-words model used for the sentence features. Table 4 shows the results for the first case (Random Forest, all attributes, and average). The classifier scores in Table 4 were calibrated again using the method described in Section 5.2. The highest score in each row is shown in bold italic and, positioning on the diagonal (shaded cells) indicates that the account was correctly linked to the resume of the account owner. Four accounts were correctly linked here.

Table 4. Resume classifier scores for case of Random Forest, all attributes, and average

		Resume no.					
		1	2	3	4	5	6
Account no.	1	-1.8772	-1.8091	-1.7393	-0.6563	-0.5510	***-0.2724***
	2	0.5327	***1.4159***	1.2572	0.4543	0.0738	-0.2863
	3	-0.7332	***0.0897***	-0.5668	-1.3735	-1.6809	-1.5570
	4	0.7635	0.4048	0.9986	***1.8512***	0.7651	1.7652
	5	0.9798	0.4715	0.2185	-0.0616	***1.5058***	-0.1729
	6	0.3344	-0.5728	-0.1682	-0.2141	-0.1128	***0.5234***

Learning algorithm	Attributes	Fusion method	First	Second
Random Forest	all	product	2	1
		average	4	1
	Department, favourite subject, hobbies, gender	product	4	1
		average	3	1
Linear SVM	all	product	3	1
		average	3	1
	Department, favourite subject, hobbies, gender	product	3	1
		average	3	1
Logistic regression	all	product	4	0
		average	4	1
	Department, favourite subject, hobbies, gender	product	4	0
		average	4	1

Table 5. Number of times correct resume was ranked first or second

Table 5 shows the number of times the correct resume (i.e. the resume of the account owner) was ranked top or second for each case. The best cases were (a) Random Forest - all attributes – average, (b) Random Forest – four effective attributes – product, (c) Logistic regression – all attributes – average, and (d) Logistic regression – four effective attributes –average, which we will evaluate in detail in the next section.

6 Evaluation

6.1 Results

We evaluated the four cases ((a), (b). (c) and (d)) described in Section 5.3 for the accounts and resumes of the 30 volunteers described in Section 4.1. We implemented component classifiers for 119 attribute values on 30 resumes using Random Forest and logistic regression, and thus implemented 119×2 component classifiers.

Figure 3 shows the results for case (b) (Random Forest - four effective attributes – product). The horizontal axis represents each of the 30 accounts, and the vertical axis represents the resume classifier scores calculated for the corresponding accounts. The

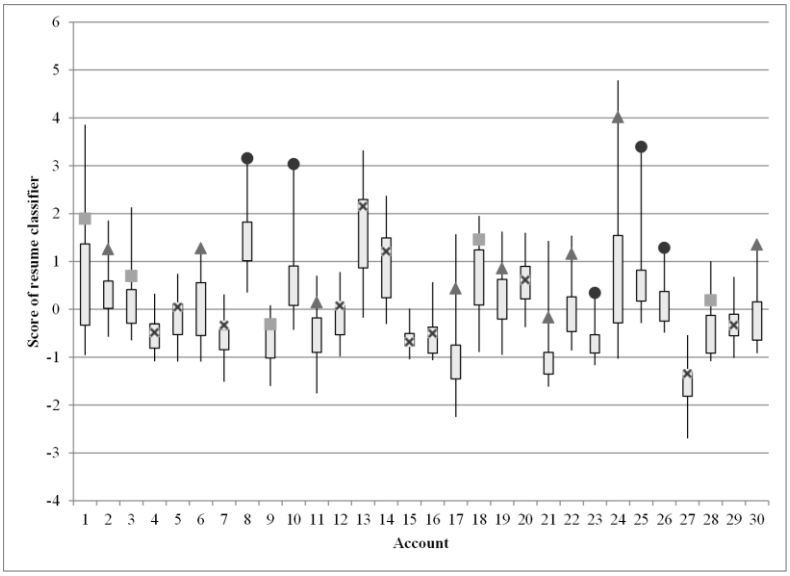

Fig. 3. Distribution of resume classifier scores with Random Forest, four attributes, and product

Table 6. Number of correct resumes being on top, in top 10%, and in top 20%

Case	Learning algorithm	Attribute	Fusion method	Top	Top 10%	Top 20%
(a)	Random Forest	All	Average	3	10	15
(b)	Random Forest	Department, favourite subject, hobbies, gender	Product	5	14	19
(c)	Logistic regression	All	Average	6	12	16
(d)	Logistic regression	Department, favourite subject, hobbies, gender	Average	2	12	18

distribution of each score calculated by 30 classifiers is represented by a box, lines above and below the box, and dots. Symbols •, ▲, □, and × represent the score of the account owner's resume. The • indicates that the account owner's resume was the top resume, meaning that the resume (i.e. the person) was correctly identified. The ▲ and □ indicate that the account owner's resume was in the top 10% (top 3) and 20% (top 6), respectively, and the × indicates otherwise. For example, the resume of account 1's owner was in the top 20%.

Table 6 shows the numbers of correct resumes being on top, in the top 10%, and in the top 20% for the four cases. The two best cases were case (b), in which 5 resumes were correctly identified, 14 resumes (including the 5 resumes) were in the top 10%, and 19 were in the top 20%, and case (c), in which 6 resumes were correctly identified, 12 resumes were in the top 10%, and 16 were in the top 20%.

Figures 4 (b) and (c) show the performance of the proposed method with less data in the best cases. The horizontal axis represents the number of tweets per account for the test data. The vertical axis represents the number of correct resumes being on top,

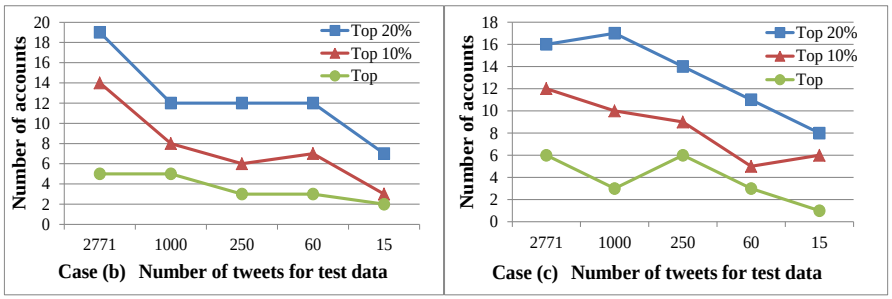

Fig. 4. Performance of proposed method with less data

in the top 10%, and in the top 20%. Basically, the less data, the less precisely the method performs though there are some fluctuations. The number of correct resumes approached the expected value with random choice (i.e. 1 for the top, 3 for the top 10%, and 6 for the top 20%) with 15 tweets.

6.2 Analysis

The results for accounts 10 and 25 were good for all cases. This was because the tweets posted from those accounts contained words related to attribute values in the corresponding resumes, especially those related to favourite subjects and hobbies. Resume 10 included, for example, "favourite subject = differential and integral calculus" while tweets from account 10 included phrases related to this subject such as "Let's practice on partial differential equations". The account owners' resumes were ranked 10, 12, and 3 for the accounts 13, 16, and 22 for case (b) but they were ranked 5, 5 and top for case (c). We may be able to improve these results by fusing scores in both cases, i.e. combining the scores for Random Forest and Logistic regression.

The results for accounts 7 and 14 were bad in all cases. Tweets from account 7 mostly contained words such as "Good morning" and "Sleepy", which were not related to the corresponding resume. Our de-anonymization method using resumes cannot work well for this kind of account. Resume 14 included "hobby = music", and tweets from account 14 mentioned music pieces and singers. However, because those music pieces and singers are not well known, the words in those tweets did not overlap words in the positive training data, i.e. the tweets of 30 music lovers. To handle this case, we need some abstraction, e.g. to learn using music and singer categories instead of words that directly appear in tweets.

While the number of attribute values described on the 30 resumes was 169, we implemented and used component classifiers for only 119 attribute values because we could not obtain sufficient numbers of positive training data for the other 50 attribute values from the Internet. This means that we could implement component classifiers for more attribute values and could improve the precision of the de-anonymization if a larger number and a wider variety of accounts were available on the Internet.

7 Summary of Our Contribution

7.1 Theoretical contribution

Previous methods that use machine learning for social network de-anonymization can be classified into two types. Methods in the first type learn general rules that are used for identifying texts meeting certain conditions (e.g. tweets revealing travel plans) [9] [13] and for inferring attributes of users (e.g. inferring marital status from the number of female friends) [10] [17]. The training data are texts and profiles from ordinary people in social networks. Methods in the second type learn person-specific rules (e.g. person's writing style) that are used for linking an account or text to another account or text that belongs to the same person [5] [6]. The training data are texts written by that person.

Our proposed method does not belong to either type. Though its purpose is similar to that of the second type, i.e. linking two objects belonging to the same person, one of the objects (i.e. resume) is not a conventional text while the other is a conventional text (tweet). Writing styles learned from the conventional text cannot be used for resume identification. Our method therefore learns general rules (e.g. those for identifying texts written by females) as do methods of the first type. It then composes person-specific rules (e.g. those for identifying texts written by the owner of a resume) from the learned general rules. The training data for our method are texts written by ordinary people. Thus, we have enabled linking different kinds of objects that belong to the same person by composing person-specific rules though learning general rules.

7.2 Implications to Stakeholders

There are four main stakeholders for our proposed method, the attacker who uses the method to identify the poster of content, the victim who is identified, the potential victim who would be identified if content was posted, and the system developer who implements the method into a real system. Our theoretical contribution most helps the system developer. Because the training data are texts from ordinary people (e.g. texts disclosed in Twitter), the system developer can obtain them without permission from a specific person. He or she can thus harvest a huge amount of training data through social networks, and the more text available in networks, the more effective the method.

8 Conclusion

8.1 Summary

We have presented a method that uses machine learning to link social network accounts to resumes, which directly represent identities. In this method, a classifier is implemented for each resume that quantifies how likely the owner of a social network account is the owner of the resume. The difficulty in using machine learning for de-

anonymization, i.e. preparing training data, is overcome by decomposing the classifier for a resume into component classifiers for characteristics (such as having dancing as a hobby and being a computer engineer) described on the resume so that training data for the component classifiers can be obtained from the Internet. Because the training data are harvested from the Internet, the more information available on the Internet, the more effective the method. It can be used widely because most organizations and governments have resume or resume-like information.

Our research clarifies a serious privacy risk: persons of concern to organizations and governments can be identified and their freedom of speech can be suppressed. The proposed method can also be used to identify a person who misuses a social network (e.g. revealing confidential information, posting copyrighted contents) by linking the misused account to a candidate resume.

8.2 Future Research Directions

(1) For the component classifiers, we will test other learning algorithms such as basic ones like naïve Bayes and more sophisticated ones like non-linear SVM and deep learning, and their combinations.

(2) For the resume classifiers, we will test learning algorithms instead of simple average and product methods. Resume classifiers need to cope with hundreds or more scores from component classifiers to precisely identify the corresponding resumes. Boosting, which adaptively optimizes the weights of scores by focusing on erroneously classified data at each stage, is therefore a promising algorithm for resume classifiers.

References

1. Gurses, S., Rizk, R., Gunther, O.: Privacy design in online social networks: learning from privacy breaches and community feedback. In: Proceedings of 29th International Conference on Information Systems, pp.1-10, Paris (2008)
2. Mixi: Infographics for finding out the newest data of mixi. http://pr.mixi.co.jp/entry/2011/06/01/infographics.html (in Japanese)
3. Narayanan, A., Shmatikov, V.: De-anonymizing social networks, In: Proceedings of 30th IEEE Security & Privacy, pp.173-187, Oakland (2009)
4. Goga, O., Lei, H. et al.: On exploiting innocuous user activity for correlating accounts across social network sites. ICSI Technical Reports - University of Berkeley (2012)
5. Almishari M, Kaafar, M., et al.: Stylometric Linkability of Tweets. In: Proceedings of 13th Workshop on Privacy in the Electronic Society, pp.205-208, Scottsdale (2014)
6. Narayanan A., Paskov, H. et al.: On the feasibility of Internet-scale author identification. In: Proceedings of 33rd IEEE Symposium on Security and Privacy, pp.300–314, San Francisco (2012)
7. Backstrom, R., Dwork, C., Kleinberg, J.: Wherefore art thou R3579X? anonymized social networks, hidden patterns, and structural steganography. In: Proceedings of 16th International World Wide Web Conference, pp. 181-190, Banff (2007)

8. Lam, I., Chen, K., Chen, L.: Involuntary information leakage in social network services. In: Proceedings of the 3rd International Workshop on Security, LNCS 5312, pp.167--183, Takamatsu (2008)
9. Mao, H., Shuai, X., Kapadia, A.: Loose Tweets: An analysis of privacy leaks on Twitter. In: Proceedings of 10th ACM Workshop on Privacy in the Electronic Society, Denver (2011)
10. Kótyuk, G., Buttyan, L.: A Machine learning based approach for predicting undisclosed attributes in social networks. In: Proceedings of IEEE 4th International Work-shop on Security and Social Networking, pp.361–366, Budapest (2012)
11. Polakis, I., Kontaxis, G., et al.: Using social networks to harvest email addresses. In: Proceedings of 9th ACM Workshop on Privacy in Electronic Society, pp.11–20, Chicago (2010)
12. TwiPro: Searching profiles of Twitter users. http://twpro.jp/ (In Japanese)
13. Caliskan-Islam, A., Walsh, J., Greenstadt, R.: Privacy detective: detecting private information and collective privacy behavior in a large social network. In: Proceedings of 13th Workshop on Privacy in the Electronic Society, pp. 35–46, Scottsdale (2014)
14. Hart, M., Manadhata, P. Johnson, R.: Text classification for data loss prevention. In Proceedings of 11th Privacy Enhancing Technologies Symposium, pp.18–37. Waterloo, 2011.
15. Burger, J., Henderson J., et al.: Discriminating Gender on Twitter. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, pp.1301–1309, Edinburgh (2011)
16. Cheng, Z., Caverlee, J., Lee, K.: You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In: Proceedings of the 19th ACM international conference on Information and knowledge management, pp.759–768, Toronto (2010)
17. Pennacchiotti, M., Popescu, A-M: A Machine Learning Approach to Twitter User Classification. In: Proceedings of 5th International AAAI Conference on Weblogs and Social Media, pp. 281–288, Barcelona (2011)