Ineluctable Background Checking on Social Networks: Linking Job Seeker's Résumé and Posts

Tomotaka Okuno
University of Electro-Communications
Tokyo, Japan
okuno@edu.hc.uec.ac.jp

Masatsugu Ichino
University of Electro-Communications
Tokyo, Japan
ichino@inf.uec.ac.jp

Isao Echizen
National Institute of Informatics (NII)
Tokyo, Japan
iechizen@nii.ac.jp

Akira Utsumi
University of Electro-Communications
Tokyo, Japan
utsumi@se.uec.ac.jp

Hiroshi Yoshiura
University of Electro-Communications
Tokyo, Japan
yoshiura@hc.uec.ac.jp

Abstract— **A growing source of concern is that the privacy of individuals can be violated by linking information from multiple sources. For example, the linking of a person's anonymized information with other information about that person can lead to de-anonymization of the person. To investigate the social risks of such linking, we investigated the use of social networks for *background checking*, which is the process of evaluating the qualifications of job seekers, and evaluated the risk posed by the linking of information the employer already has with information on social networks. After clarifying the risk, we developed a system that links information from different sources: information extracted from a job seeker's résumé and anonymous posts on social networks. The system automatically calculates the *similarity* between information in the résumé and in the posts, and identifies the job seeker's social network accounts even though the profiles may have been anonymized. As a part of our system, we developed a novel method for quantifying the implications of terms in a résumé by using the posts on social networks. In an evaluation using the résumés of two job seekers and the tweets of 100 users, the system identified the accounts of both job seekers with reasonably good accuracy (true positive rate of 0.941 and true negative rate of 0.999). These findings reveal the real social threat of linking information from different sources. Our research should thus form the basis for further study of the relationship between privacy in social networks and the freedom to express opinions.**

Keywords-privacy, social networks, anonymity

I. INTRODUCTION

Various types of personal information are disclosed on networks, and they are accessible in a number of ways. To prevent the violation of privacy, service providers have introduced a number of countermeasures, including disclosure control and anonymization. For example, social network service providers have increased the number of privacy settings and some users anonymize their profile by themselves. More than half of the 25 million users of Japan's largest social network, *mixi*, do not use their actual name in their profile, reflecting the Japanese tendency of preferring anonymity [1].

Even with such protective measures, however, there is still the risk of one's privacy being violated by the collection of personal information from different sources. This has been demonstrated in a number of studies in which anonymized personal information was linked with other available information, resulting in de-anonymization. For instance, Narayanan et al. showed that large datasets containing anonymized movie ratings can be de-anonymized by using information from another dataset as background knowledge [2]. These works are theoretically interesting, yet only a few have shown the real social threat of linking information from multiple sources.

In the work reported here, we focused on the concern that the checking of job seeker qualifications by using social networks can violate the job seeker's privacy. The current trend in *background checking* is for the prospective employer to search for a job seekers' profile on social networks by using information from the person's CV or résumé. Once the profile has been found, the employer can investigate the behavior and friendships of the job seeker by reviewing his or her posts on the network and the résumé. The findings may cause the job seeker to be excluded from employment consideration [3]. Background checking on employees or prospective employees using social networks has become a societal concern. For example, California has enacted legislation that will soon make it illegal for employers to demand employees' social network usernames and passwords [4].

In this paper, we show that it is feasible to identify the social network accounts of job seekers by using information on their résumés even if the job seekers have anonymized their profiles on social networks. This work shows the real social threat of linking two types of information: information in résumés and anonymous posts on social networks.

Our technical and societal contributions

1. We focus on the process of *background checking* using social networks and show the social threat of

978-1-4673-5077-8/13/$31.00 ©2013 IEEE

potential linking of information from different sources. We present a method of attack in which the accounts of job seekers are identified even if they have taken countermeasures against attack, e.g., by anonymizing their profiles on social networks.

2. We propose a methodology for linking different types of sources: résumés (a list of personal information that can be considered to be a set of records in a database) and informally written posts on social networks (that can be considered to be text information). This differs from matching between text information sources [5][6] and between network topologies [7]. We also propose, as part of our methodology, a method for quantifying the implications of terms in a résumé by using posts on social networks.
3. This work should form the basis for the further study of the connection between privacy on social networks and freedom to express opinions. For instance, an employer could feasibly identify an employee who had anonymously posted comments criticizing the company on social networks. Our proposed method could be used to develop technology for protecting privacy in such cases.

The remainder of this paper is organized as follows. We review related work in Section II and discuss background checking in Section III. We present our model for identifying the accounts of job seekers in Section IV and describe our implementation and evaluation in Section V. The limitations of our methodology are addressed in Section VI. The key points are summarized and future work is mentioned in Section VII.

II. Related Work

A. Privacy Concerns Related to Social Networks

There have been many studies of the use of privacy settings in social networks. One of the earliest studies was the one conducted by Gross et al. in 2005 in which they surveyed 4000 Facebook users [8]. They found that 89% of them were using their actual name and that 54% were sharing their address with everyone. Things have changed, however. Social networkers are steadily becoming more and more careful about their privacy. A survey conducted by Liu et al. in 2011 showed that about half (49%) of Facebook contents were disclosed only to friends and that 36% were open to all users [9]. Dey et al. surveyed the profiles of 1.4 million Facebook users and discovered that 53% of the users hid their friends list in 2011 while only 17% had hid it a year earlier [10].

Some social networkers not only use privacy settings but also anonymize their profiles. More than 60% of the mixi users in their early 20s and 70% of those in their late 20s do not use their actual names in their profiles [1]. About half of them used a name containing something that their friends could recognize.

In 2010, Meeder et al. analyzed 2.7 billion posts and 80 million profiles on Twitter and reported that retweeting tweets could violate the original tweeter's privacy [11]. In 2011, Mao et al. analyzed tweets containing sensitive information and proposed a method for classifying such tweets [12]. They were able to classify the tweets containing disease and vacation information with 76% accuracy and ones containing drinking and driving information with 84% accuracy.

A variety of methods for attacking social networks have been reported. In 2008, Lam et al. analyzed users on *Wretch*, the largest social network in Taiwan, and reported that the given name of 72% of the users and the full name of 30% of the users could be identified by analyzing comments posted by their friends [13]. Kótyuk analyzed information on the users of iWiW, the largest social network in Hungary, and found that a user's age could be estimated to within 4.8 years even if the birth year was not given in their profile [14]. Likewise, gender and marital status could be estimated with accuracies of 74% and 68%, respectively.

B. Attacks Using Information from Different Sources

A wide variety of studies regarding de-anonymization using information from different sources have been undertaken in recent years. Narayanan et al. applied the method used in their previous study [2] to network topology and used it for identifying users who used two different social networks [7]. They were able to identify users who were using both Twitter and Flickr with an error rate of 12%. Goga et al. subsequently proposed a method for identifying users who used different social networks by analyzing and combining the features of geo-location, timestamp, and writing style from their posts [15]. They were able to identify users who used both Yelp and Twitter and both Flickr and Twitter and showed the potential risk of de-anonymization using cross-site correlation techniques.

Authorship identification techniques have been used for various attack methods. In an early study (2004), Novak et al. used authorship analysis to identify texts posted on the Web by the same person with an accuracy of better than 90% [5]. In a later study (2012), Narayanan et al. applied authorship identification to a large number of blogs [6] and identified the authors of anonymous posts by 100,000 blog authors with high accuracy.

Polakis et al. used a method for linking the names of social network users with their e-mail addresses [16] and used it to match 43% of the user profiles extracted from Facebook with the user e-mail addresses. Acquisti et al. were able to identify the users of dating sites by using face recognition techniques to match photos posted on dating sites with ones posted on Facebook [17].

III. Background Checking

Background checking is the process of checking criminal and other records and is often done by prospective employers for evaluating job seekers' qualifications and character. However, it has the potential of violating the job seeker's privacy, depending on the purpose and methods used [18].

Such investigations are being extended to the use of social networks as well. A number of studies have reported that employers are searching social networks for background information. According to a survey conducted by

Careerbuilder.com in 2008, 21% of employers reported that they checked social networks for background information [3], and 34% of them reported finding content that caused them to remove someone from consideration. Another survey conducted by Clark et al. in 2008 showed that 43% of employers check social networks [19]. The top two reasons for removing them were "candidate posted information about them drinking or using drugs" (41%) and "candidate posted provocative or inappropriate photos or information" (40%) [3].

Conceivable countermeasures against background checking using social networks can be divided into two types.

1. Controlling the privacy settings of the contents, i.e., using the *friends only* or *protected* setting, so that only a limited group of users can browse the contents.
2. Anonymizing personal information in profiles, such as one's name, e-mail address, and photo.

Application of the first type of countermeasure would prevent prospective employers from being able to directly browse the information posted by job seekers. However, prospective employers could still determine whether a job seeker is a social networker and, if so, make giving the employer access to the job seeker's posted information a requirement for being considered for employment. The company could also take the approach of having an employee who had attended the same school to send a friend request to the job seeker. Such spoofing would enable the new "friend" (or friend of a friend) to browse the job seeker's information and posts. This is similar to reported instances in real life in which a company employee posed as an acquaintance of a job seeker for the purpose of background checking [18].

Application of the second type of countermeasure would enable job seekers to avoid simple identification, such as by name retrieval. For example, they could change their personal information so that only friends would be able to identify them. Furthermore, prospective employers would be unable to determine whether a job seeker is a social networker. The employer would thus be unable to demand access to their social network accounts. However, even if anonymity is achieved by deleting personal information from one's profile, other contents such as posted texts remain. Therefore, there is the risk that a job seeker's social network accounts could still be identified by, for example, linking information gleaned from the person's résumé and interview.

We focused on the latter type of countermeasure and clarified the potential risk of job seekers' social network accounts being identified by linking the information posted on social networks with the information employers gather during the recruitment process. Our first goal was to develop a method for identifying a job seeker that matches texts posted on social networks with information from the job seeker's résumé even if the job seeker has his or her profile to avoid background checking.

IV. Model for Identifying Anonymous Job Seekers

In the system model we developed for identifying job seekers, we assume that the candidate social network accounts of the job seeker are given. For example, by searching the posts on those accounts using keywords job seekers would typically use after a job festival such as company name, the term "job festival" itself, and the name of the closest train station. The system calculates the similarity between information in the résumé and that in the posts of each social network user. The job seeker's account is assumed to be the one for which the similarity is the highest. By using this system, employers can link the information in social networks with the résumés of job candidates and thereby learn about the job seekers' behaviors and friendships.

All of the information used in the model is in Japanese. We use the short messages posted on Twitter, known as *tweets*, as the social network posts. Since tweets contain no more than 140 characters and spaces, we consider a set of 100 tweets as one document:

$$D_{xl} = \bigcup_{s=100(l-1)+1}^{100l} T_{xs} \left(1 \le l\right) \tag{1}$$

where T_{xs} is the s-th tweet posted by user x. For instance, D_{Adam1} is a document containing the 1st to 100th tweets posted by Adam. Résumés in Japan are standardized and usually contain the job seeker's name, age, gender, address, phone number, educational history, job history, qualifications/licenses, interest information, and self promotion. We used address, educational history, job history, and qualifications/licenses.

V. Implementation and Evaluation

A. Framework for Similarity Calculation

The information in a résumé is a list of personal information that can be considered to be a set of records in a database. However, in this system, we consider the résumé to be a document and calculate the similarity between the résumé and a set of tweets D_{xl}. The similarity is computed by using a *cosine similarity*, an algebraic model commonly used in information retrieval. In this model, documents are represented as vectors, and the similarity between two document vectors is computed.

We created vectors consisting of all nouns and compound words from the résumé. They were extracted using Mecab [20], a Japanese language analyzer. The similarity between the résumé and a set of tweets was computed using

$$sim\left(D_{xl}, R_u\right) = \cos\left(D_{xl}, R_u\right) = \frac{D_{xl} \cdot R_u}{\left\|D_{xl}\right\| \left\|R_u\right\|} \tag{2}$$

where R_u is the résumé of job seeker U.

B. Basic Method

We initially took a classic approach to calculating the vector term weights. That is, we used *TF-IDF*, where TF is the term frequency and IDF is the inverse document frequency. The TF-IDF weight w_{ij} of term j in document d_i is given by

$$w_{ij} = tf_{ij} \times idf_j = \frac{n_{ij}}{\sum_k n_{ik}} \times \log \frac{N}{N_j} \tag{3}$$

where n_{ij} is the number of appearances of term j in document d_i, N is the total number of documents in the document set, and N_j is the number of documents containing the term j. We used the total number of web pages (10^{12}) and the total number of web pages containing term j, as determined by using a web search engine, for N and N_j, respectively. Using the term weights, we can modify expression (2) as follows.

$$sim(D_{xl}, R_u) = \cos(D_{xl}, R_u) = \frac{D_{xl} \cdot R_u}{\|D_{xl}\| \|R_u\|} = \frac{\sum_{j=1}^{m} w_{D_{xl},j} w_{R_u,j}}{\sqrt{\sum_{j=1}^{m} w_{D_{xl},j}^2} \sqrt{\sum_{j=1}^{m} w_{R_u,j}^2}} \quad (4)$$

C. Preliminary Evaluation of Basic Method

We evaluated the term weights of the vectors to analyze the effectiveness of the method and visualized them as shown in figure 1. The horizontal axis represents the terms extracted from the résumé of Adam, a job seeker. The vertical axis represents the term weights (TF-IDF) of the terms in Adam's set of tweets. In fact, the figure shows the term weights of Adam's tweets, where the terms were extracted from his résumé.

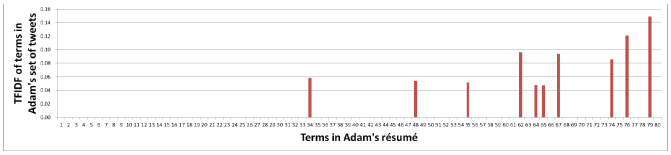

Figure 1. Term weights (TF-IDF) in Adam's set of tweets.

The weights of the commonly used terms, such as "computer" and "communication," are non-zero. In contrast, the weights of the terms that are more likely to represent the attributes of the job seeker, such as university name, are zero, meaning that these terms cannot be found in tweets. In this way, even though the weights of the job seeker's tweet set are used in the evaluation, 70 out of 80 terms extracted from job seeker's résumé are non-zero. At this level, the similarity between a person's résumé and that person's tweet set would be unduly low.

The main reason for this is that the terms used by someone in a résumé are not used exactly the same way as in tweets. For instance, students at the University of Electro-Communications usually abbreviate the name to UEC in tweets that mention the university. (Japanese abbreviations are usually more complex than this one; they are usually not initialized.) Moreover, the terms used in a résumé that contain personal or private information are often reworded into a phrase that only a subset of readers (e.g., friends or friends of friends) could understand, such as "university in Chofu city," instead of "UEC." These unwritten terms are not reflected in the measure of similarity because their TF-IDF would be zero since they cannot be string-matched to the terms in the résumé.

D. Improved method

We thus improved the algorithm to enable it to detect the unwritten terms and to calculate their correlation to the terms in the résumé. In our algorithm, the correlation between term j in the résumé and a set of tweets D_{xl} is computed in the range 0.0–1.0 as $HTF(D_{xl}, j)$, where HTF means hidden term frequency:

I. Extract the b-th tweet ($1 \le b \le 100$) from the set of tweets D_{xl}.

II. Morphologically extract the nouns and compound words from the tweets. Let n be the number of extracted terms.

III. From the n extracted terms, generate a query that contains a term or combination of terms and add the query to the query list. The number of queries in the query list is computed using

$$N = \sum_{a=1}^{m} \binom{n}{a} \quad (5)$$

($m = 3$ in this evaluation.)

IV. For query $q_i (1 \le i \le N)$ in the query list, execute the following process.

i. Use a search engine with q_i and retrieve the page title and the text summary of the top k results. (k = 20 in this evaluation.)

ii. For the d-th retrieved result, determine whether the page title and summary contain term j in the résumé and return f_d:

$$f_d = \begin{cases} 1 & (if\ j\ is\ found) \\ 0 & (if\ not) \end{cases} \quad (6)$$

iii. Using f_d, calculate the score of q_i:

$$score(q_i) = \frac{1}{C} \sum_{d=1}^{k} (k - d + 1) f_d \quad (7)$$

The score is normalized in the range 0.0–1.0 by normalizing coefficient C.

$$C = \sum_{d=1}^{k} (k - d + 1) \quad (8)$$

V. From all the $score(q_i)$ for query q_i, select the highest score as $h(t, j)$, which shows the correlation of term j with the b-th extracted tweet:

$$h(t, j) = \max_i (score(q_i)) \quad (9)$$

VI. Repeat steps I to V for each tweet in the set of tweets, and calculate $HTF(D_{xl}, j)$, which is the average of $h(t, j)$.

$$HTF(D_{xl}, j) = \frac{1}{100} \sum_{t=1}^{100} h(t, j) \quad (10)$$

The score computed in step V is the quantified value of the reachability of a term in the résumé from one in the tweet. For instance, the score is 1 if all the retrieved results contain the term j, and the score is 0.014 if only the 19th and 20th results contain j. Using this score, we quantified the reachability of "University of Electro-Communications" in a résumé from the tweet "Open campus and briefing session

on entrance exam for graduate school will be held on May 22." In this example, the score $h(t, j)$ is 0.12.

After calculating the HTFs, we use them for $W_{D_{xlj}}$ in expression (4) and use (4) to calculate the similarity.

E. Preliminary Evaluation of Improved Method

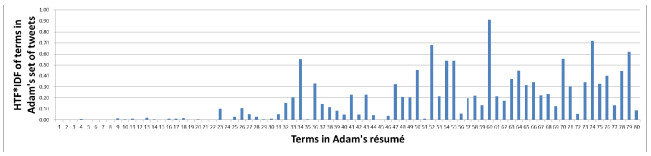

Figure 2. Term weights (HTF × IDF) in Adam's set of tweets.

Figure 2 shows the term weights of the vectors obtained using the improved method. The vertical axis represents the product of HTF and IDF, instead of that of TF and IDF, for each term. With the basic method, most terms in the résumé do not appear in the tweets, meaning that most of the term weights are zero. With the improved method, 66 out of the 80 were weighted as non-zero. In fact, the implications of terms in the résumé that were not mentioned in tweet, such as the name of the university and the name of the conference the job seeker attended, were quantified as non-zero.

F. Evaluation

We evaluated the improved method using the résumés of two job seekers (Adam and Bob) and 100 randomly selected Twitter users in September 2012. We extracted 1000 tweets from each of these 100 users and Adam, and 700 tweets from Bob, since his max number of tweets was 700.

Figure 3 shows the similarity between Adam's résumé and each user's tweets. The vertical axis represents the similarity between Adam's résumé and each user's set of tweets. The horizontal axis represents the number of tweet sets for each user. For instance, the 1 on the horizontal axis represents D_{x1}, which is the 1st set of tweets (tweets 1–100) of user x (x includes 100 users and Adam). The blue triangles represent the similarity between Adam's set of tweets and Adam's résumé. The Box plots in figure 3 represent the distribution of similarity of the 100 users' sets of tweets to Adam's résumé. The black box represents 50% of the users between the lower and upper quartiles of the distribution. The similarity between the tweets and the résumé of job seeker is the highest all the time because all of the blue triangles are above box plots.

Figure 4 shows the detailed distribution of similarity of the 100 users in D_{x1} in figure 3. Table 1 shows the evaluation results for Adam's résumé. The true positive rate (TPR) was 1.0 because the similarity between tweets and his résumé was highest for all of the tweets sets, and the true negative rate (TNR) was 1.0 because none of the similarity between his résumé and the other users' tweets became the highest.

Table II and Figures 5 and 6 show the results of similarity estimation between Bob's résumé and each user's tweets. The TPR was 0.857 and the TNR was 0.998 because D_{Bob1}, the 1st set of Bob's tweets, did not have the highest similarity. However, the similarity between D_{Bob1} and Bob's résumé was the second highest and close to the most similar set (within 0.002).

Combining the results for the two job seekers, we get a TPR of 0.941 and a TNR of 0.999. Considering the challenging nature of this problem, it is fair to say that our methodology can be used to identify a job seeker from a number of twitter users with reasonably good accuracy.

TABLE I. EVALUATION RESULTS FOR ADAM

	Actual +	Actual −
Predicted +	10	0
Predicted −	0	1000

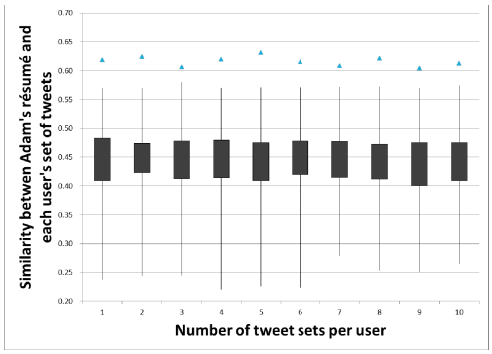

Figure 3. Similarity between Adam's résumé and each user's tweets.

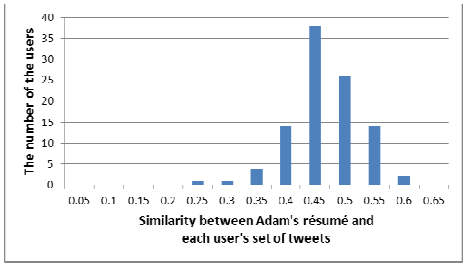

Figure 4. Distribution of similarity between Adam's résumé and 100 users in D_{x1}.

TABLE II. EVALUATION RESULTS FOR BOB

	Actual +	Actual −
Predicted +	6	1
Predicted −	1	699

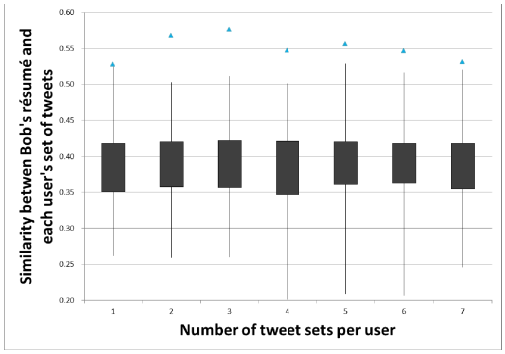

Figure 5. Similarity between Bob's résumé and each user's tweets.

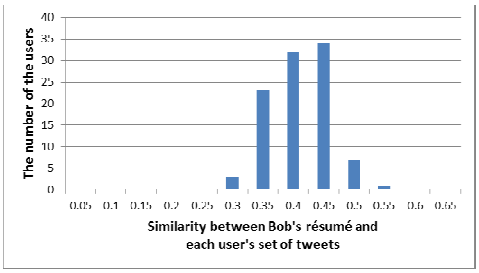

Figure 6. Distribution of similarity between Bob's résumé and 100 users in D_{x1}.

VI. LIMITATIONS

Our study does have two limitations. First, we were only able to evaluate with two job seekers since résumés contain personal information and are thus hard to obtain. Second, as mentioned in section IV, we need to identify a set of social network users that includes the job seeker.

We have three strategies for overcoming these limitations. We are currently recruiting participants who can provide us with at least part of the information on their résumés with the aim of conducting a larger evaluation with them as participants. Second, we are extending our system to handle a larger number of social network users. Finally, we are investigating practical methods for narrowing down the number of candidate job seekers before applying the system.

VII. CONCLUSION

We have addressed the increasing concern about the privacy of individuals being violated by linking information from multiple sources. We focused on *background checking* and investigated the risk of linking information employers would have with information on social networks.

We developed a system for identifying the social network accounts of job seekers even if they have anonymized their profiles in social networks. The system identifies the accounts by linking two types of sources: information from the job seeker's résumé (which can be considered to be a record in a database) and posts on social networks (which can be considered to be text information).

The information is linked by quantifying the implications of terms in the résumé from posts. In an evaluation with two job seekers and 100 Twitter users, we were able to identify the accounts of the job seekers with reasonably good accuracy (TPR of 0.941 and TNR of 0.999). Our results demonstrated the social risk of linking information from different sources. Our research should thus inspire new research on social networks and the freedom to express opinions. For instance, our method could be used to enhance privacy protection when employers try to identify the employees who criticized the company on social networks.

Our future work is aimed at two challenges. The first is to extend the evaluation to a larger number of participants who could provide us with information on their résumés. The second is to develop a technique for protecting social networks from attacks like ours that use information from different sources.

REFERENCES

[1] Mixi, "Infographics for finding out the newest data of mixi," http://pr.mixi.co.jp/2011/06/01/infographics.html (in Japanese).

[2] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," in Proceedings of the 29th IEEE Symposium on Security and Privacy, 2008, pp.111–125.

[3] CareerBuilder.com, "One-in-Five Employers Use Social Networking Sites to Reasearch Job Candidates, CareerBuilder.com Survey Finds", http://www.careerbuilder.com/share/aboutus/pressreleasesdetail.aspx?id=pr459&sd=9%2F10%2F2008&ed=12%2F31%2F2008.

[4] Reuters, "California schools, employers banned from social media snooping", http://www.reuters.com/article/2012/09/27/us-usacalifornia-privacy-idUSBRE88Q1UI20120927.

[5] J. Novak, P. Raghavan, and A. Tomkins, "Anti-Aliasing on the Web," in Proceedings of the 13th International Conference on World Wide Web, 2004, pp.30–39.

[6] A. Narayanan, et al., "On the Feasibility of Internet-Scale Author Identification," in Proceedings of the 33rd IEEE Symposium on Security and Privacy, 2012, pp.300–314.

[7] A. Narayanan and V. Shmatikov, "De-anonymizing Social Networks," in Proceedings of the 30th IEEE Symposium on Security and Privacy, 2009, pp.173–187.

[8] R. Gross, and A. Acquisti, "Information revelation and privacy in online social networks," in Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society, 2005, pp.71–80.

[9] Y. Liu, K.P. Gummadi, B. Krishnamurthy, and A. Misolve, "Analyzing facebook privacy settings: user expectations vs. reality," in Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, 2011, pp.61–70.

[10] R. Dey, Z. Jelveh, and K. Ross, "Facebook Users Have Become Much More Private: A Large-Scale Study," in Proceedings of IEEE 4th International Workshop on Security and Social Networking, 2012, pp.346–352.

[11] B. Meeder, J. Tam, P. Kelly, and L.F. Cranor, "RT@ IWantPrivacy: widespread violation of privacy settings in the Twitter social network," in Proceedings of the Web 2.0 Privacy and Security Workshop, 2010.

[12] H. Mao, X. Shuai, and A. Kapadia, "Loose Tweets: An Analysis of Privacy Leaks on Twitter," in Proceedings of the 10th ACM Workshop on Privacy in Electonic Society, 2011, pp.1–12.

[13] I-F. Lam, K-T. Chen, and L-J. Chen, "Involuntary Information Leakage in Social Network Services," in Proceeding of the 3rd International Workshop on Security: Advances in Information and Computer Security, 2008, pp. 167–183.

[14] G. Kótyuk and L. Buttyan, "A Machine Learning Based Approach for Predicting Undisclosed Attributes in Social Networks," in Proceedings of IEEE 4th International Workshop on Security and Social Networking, 2012, pp.361–366.

[15] O. Goga, et al., "On Exploiting Innocuous User Activity for Correlating Accounts Across Social Network Sites," ICSI Technical Reports - University of Berkeley, 2012.

[16] I. Polakis, et al., "Using Social Networks to Harvest Email Addresses," in Proceedings of the 9th ACM Workshop on Privacy in Electonic Society, 2010, pp.11–20.

[17] A. Acquisti, R. Gross, and F. Stutzman, "Faces of Facebook: Privacy in the Age of Augmented Reality," in BlackHat USA, 2011.

[18] Johnson v. K Mart Corp., No. 1-98-2172, http://www.state.il.us/court/opinions/appellatecourt/2000/1stdistrict/january/html/1982172.htm.

[19] Roberts, S. and Clark, L., "Myspace, Facebook, and Other Social Networking Siters: How Are They Used by Human Resource Personnel?," in Delta Pi Epsilon National Conference, 2008, pp.35-43.

[20] T. Kudo, "MeCab. Yet Another Part-of-Speech and Morphological Aanalyzer," http://mecab.sourceforge.net/ (in Japanese).