Public Information Exposure Detection: Helping Users Understand Their Web Footprints

Lisa Singh, Grace Hui Yang, Micah Sherr,
Andrew Hian-Cheong, Kevin Tian, Janet Zhu, Sicong Zhang
Georgetown University
Department of Computer Science
Washington, DC, 20057 USA

***Abstract*—To help users better understand the potential risks associated with publishing data publicly, as well as the quantity and sensitivity of information that can be obtained by combining data from various online sources, we introduce a novel information exposure detection framework that generates and analyzes the *web footprints* users leave across the social web. Web footprints are the traces of one's online social activities represented by a set of attributes that are known or can be inferred with a high probability by an adversary who has basic information about a user from his/her public profiles. Our framework employs new probabilistic operators, novel pattern-based attribute extraction from text, and a population-based inference engine to generate web footprints. Using a web footprint, the framework then quantifies a user's level of information exposure relative to others with similar traits, as well as with regard to others in the population. Evaluation over public profiles from multiple sites (Google+, LinkeIn, FourSquare, and Twitter) shows that the proposed framework effectively detects and quantifies information exposure using a small amount of initial knowledge.**

I. Introduction

The popularity of digital communication has led to the sharing of enormous amounts of personal information on social media sites. While it is likely that many users of these online services understand that they are sharing personal information with strangers, they may not understand the potential risks and implications of doing so. High levels of exposed information can (and do) lead to severe consequences such as stalking [17], identity theft [19], and job loss [20]. Given these possible life changing risks, it is important for users of online social networks (OSNs) and other online services to understand (1) the increase in vulnerability that occurs when they choose to share different golden nuggets of information online, and (2) how their privacy risks are further exacerbated when they share personal information on *multiple* independent OSNs. We posit that the lack of intuitive privacy metrics and a comprehensive exposure detection framework makes it difficult for users to assess their privacy risks due to online sources of information. This paper examines this problem of quantifiably measuring online privacy risks.

Our methods and algorithms can be grouped into a more general problem that we refer to as *public information exposure detection* (PIE detection). The objective of PIE detection is to develop effective and efficient algorithms for identifying and quantifying components of a user's public profile that reduce the user's privacy and potentially expose the user to adversarial behaviors. Our techniques construct users' individual public profiles or *web footprints*. Web footprints are traces of one's social activities represented by a set of attributes that are known or can be inferred with a high probability by combining data across sites on the Internet. The adversary's task is to digitally "stalk" a victim (which we refer to as the *user* of our framework) and discover as much information as possible about him or her, either through direct observation of posted information or by gaining knowledge using inference logic.

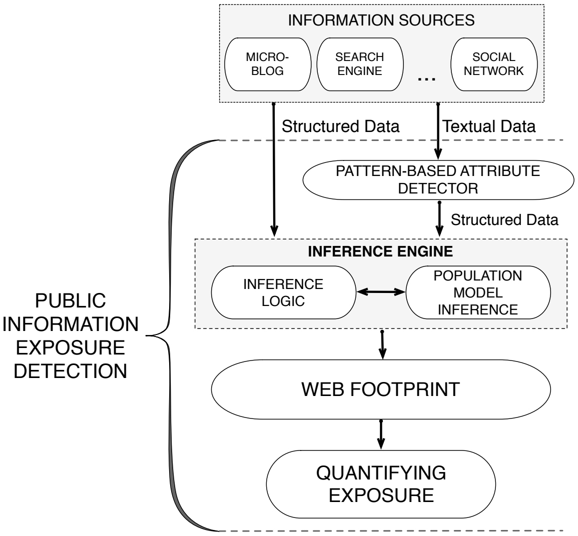

Fig. 1: The PIE framework.

The proposed framework constructs web footprints (as would an adversary intent on stalking a user) and reports to the user his or her particular level of vulnerability based on the user's publicly shared information. The framework (shown in Figure 1) first creates the user's web footprint by combining publicly accessible information from various online services such as social media sites, micro-blogging sites, data aggregation sites, and search engines. Since much web data is unstructured, we also introduce a *pattern-based attribute extractor* that uses bootstrapped patterns found in a corpus to extract structured attribute values from the unstructured text, thereby increasing the amount of usable information for web footprint construction. In addition to observable data, probabilistic inference logic is applied to supplement web footprints with probable attribute value pairs learned using algebraic dependencies between attribute values in user profiles on different sites. Finally, we use site-level population data to further infer the user's attribute values. To allow for population level comparison, our framework also quantifies a user's level of public information exposure relative to others with similar traits as well as with regard to others in the population. We demonstrate the added value of each component within the framework and show that the combined knowledge from all three complementary components is greater than any subset.

In summary, the contributions of this paper are as follows: (1) we formally define public information exposure detection; (2) we develop a new holistic methodology for analyzing an individual's web footprints that incorporates new operators for probabilistic closure and join, pattern-based attribute extraction from text, and rules generated from a population-based inference engine; (3) we propose different metrics for quantifying information exposure and identifying *risky* attributes; and (4) we evaluate our framework on thousands of public profiles gathered from four social networking sites and show that an adversary can generate a cross-site web footprint with high accuracy when using our framework.

II. Related Work

There has been an emerging interest in linking individuals across online social networks (OSNs) [4, 7–11]. Goga et al. [4] introduce a correlation attack that attempts to find accounts on different social networking sites belonging to the same person. The three features used in the attack are post location (using geo-location information), post timestamp, and the writing style of the user. Iofciu et al. [8] have a similar study that uses username across social tagging systems (Flickr, Delicious, and StumbleUpon) to map users across sites. Both of these works differ from ours because they make the assumption that the attacker is a 'pseudo-friend'—i.e., having certain knowledge about one (or more) of the social network account ids on one of the sites. In our adversarial model, we assume knowledge of common, publicly available attributes, but not of user ids.

Similar in spirit to our work, Irani et al. [9] introduce the idea of a social trace that maps data across social networks. They show that using one social network username across websites increases the amount of leakage considerably. Jain and Kumaraguru [10] propose the *Finding Nemo* system that uses user profile information, a user's network, and user generated content to map users across Twitter and Facebook. Unlike our work, the authors do not consider text data or inference using online population data. Finally, Moore et al. [12] consider different population-based inference algorithms (Naive Bayes, LDA, and Association Mining) for determining public attribute values of approximately 100,000 users on LinkedIn and Google+. While that work focuses on understanding the accuracy of different algorithms for population-based inference of publicly available attributes (not individuals), our work uses population inference in conjunction with probabilistic inference logic and pattern-based attribute extraction to conduct a more comprehensive, larger inference process for individuals.

Re-identification is a closely related problem to information exposure detection [2, 5, 13, 14, 18]. In re-identification, the goal is to match anonymized personal data with their true owners. In this paper, rather than evaluating how well data anonymization techniques can protect privacy, we measure the amount of privacy loss due to publicly exposed *non-anonymized data* (for example, the attribute values that appear in public profiles on OSNs). We also note that our techniques are related in their goals to work that aims to build more privacy-preserving social networking services [1, 3]. In contrast to these approaches, our research addresses the problem of quantifying and reducing exposure due to *existing* social networking services. Our approach further differs from previous work in its ability to quantify information exposure at multiple resolutions, focusing both at the site-level (*micro-analysis*) and at Internet-scale (*macro-analysis*).

Finally, a body of existing work explores information leakage due to side channels in OSNs. Song et al. [16] observe that URL shortening services (e.g., bit.ly) offer click analytics that reveal user agent (browser) information about users who clicked on links; an adversary can leverage this information to infer which Twitter users clicked on shortened URLs. Earlier work by Humbert et al. [7] demonstrates that private profiles on OSNs can be discovered by targeting crawls based on the profile owner's geographic region. While the above techniques focus on exploiting information leakage—i.e., data that are exposed through accidental channels—we address a more basic question: how much information is revealed by directly publishing data on the web.

III. Problem Definition

We consider a person P about whom an adversary is attempting to learn information. For clarity, we omit explicitly denoting P in our notation. We assume that P has publicly revealed certain features or attributes (e.g., name and age) or that such information is otherwise publicly available, perhaps from a data aggregation site. Some of the revealed information may be sensitive, e.g. birthday or income. If a revealed attribute cannot be matched to P, it is considered *hidden*. Let $A = \{A_1, A_2, \ldots, A_m\}$ be the set of P's attributes. For each attribute A_i, P has one or more values $\alpha_i = \{\alpha_i^1, \alpha_i^2, \ldots, \alpha_i^h\}$. As shorthand, we denote P's attribute value(s) for an attribute A_i as α_i, and all attribute values belonging to P (across all attributes) as $\overline{\alpha}$. When referring to both the attribute (A_i) and its value (α_i) as a pair, we will use the notation $\langle A_i, \alpha_i \rangle$.

A. Adversary Model

We frame the public information exposure detection problem in the context of an adversary who wishes to (1) gather publicly available information about a target individual P, and (2) infer additional attribute values about P by applying inference techniques to the publicly available information.

We assume that the adversary uses publicly available information about P to form *beliefs* about P that are not originally known to the adversary. We model a belief $\mathcal{B}_j = \langle A_i, \alpha_i^j, \mathrm{conf} \rangle$ as a triple consisting of an attribute, a single attribute value, and a confidence. Conceptually, the confidence $\mathrm{conf} \in (0,1]$ is an assessment as to the likelihood that a true value for A_i is α_i, where a confidence of 1 denotes certainty. To avoid maintaining beliefs that are weak, we maintain only beliefs that have a confidence above a threshold.

To learn information about P, we allow the adversary to query a set of sites $S = \{s_1, s_2, \ldots, s_q\}$. For example, these sites could consist of microblogging sites, online social networks, search engines, and data aggregation sites. Importantly, we restrict the adversary to access only publicly available information from the sites. We also assume that the adversary is not connected to P in a social network or otherwise has special access to P's information that is not available to the general public Although the adversary does not have to obey sites' acceptable use policies, he cannot directly access the sites' backend databases and must learn information about P by using sites' exposed APIs or by screen scraping.

We assume that the adversary has some background knowledge about P (e.g., P's name). The set of attributes and associated attribute values known a priori to the adversary is referred to as the initial *attribute value core*, $\mathcal{B}_{\text{core}}$, where $\mathcal{B}_{\text{core}} = \{\mathcal{B}_1, \ldots, \mathcal{B}_n\}$ and $\forall \mathcal{B}_i \in \mathcal{B}_{\text{core}}$, $\mathcal{B}_i = \langle A_i, \alpha_i^j, 1\rangle$ and $\alpha_i^j \in \overline{\alpha}$. In other words, this attribute value core is a set of *correct beliefs* about the targeted user known a priori by the adversary. Because these beliefs are given and assumed to be true, they have a confidence of 1. The adversary's goal, therefore, is to determine hidden attributes (initially unknown to him), $\overline{\alpha} \setminus \mathcal{B}_{\text{core}}$, of person P.

B. Web Footprint

The web footprint of P, which we denote as $\mathcal{W}$, is composed of a set of beliefs held by the adversary: $\mathcal{W} = \{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_w\}$. We define $\mathcal{B}_{\text{true}}$ to be the subset of beliefs in $\mathcal{W}$ that are correct. That is, $\mathcal{B}_{\text{true}} = \{\langle A_i, \alpha_i^j, \text{conf}\rangle \in \mathcal{W} : \alpha_i^j \in \overline{\alpha}\}$. Note that, $\mathcal{B}_{\text{core}} \subseteq \mathcal{B}_{\text{true}} \subseteq \mathcal{W}$. The adversary discovers the attributes in $\mathcal{W}$ that are not in $\mathcal{B}_{\text{core}}$ by applying different matching and inference rules to $\mathcal{B}_{\text{core}}$, as well as obtaining information about the service population by querying online services. It is the adversary's goal to determine as many attribute values as possible in $\overline{\alpha}$ with high confidence (that is, to correctly learn the attribute values belonging to P that are not present in $\mathcal{B}_{\text{core}}$).

C. Information Exposure and Information Accessibility

We let $\sigma_X(s_j) = p_j = \{p_j^1, p_j^2, \ldots, p_j^t\}$ represent the set of *profiles* returned by a query against site s_j using the attribute values in set X. Each profile is a set of attribute/attribute value pairs, with the possibility that a given attribute may have multiple values belonging to a user of the site. For multi-valued attributes, e.g., school-attended, each attribute value is represented as a separate belief. Note that since many users may have the attribute values specified in X, multiple profiles may be returned for a given query. For simplicity, we assume that if $X \subseteq \mathcal{B}_{\text{core}}$, then exactly one of the profiles in $\sigma_X(s_j)$ belongs to the user P (i.e., the user has a single profile on the queried site). Let p_j^{true} be P's profile on s_j.

Information exposure occurs if the adversary identifies attribute values in profile p_j^{true} that (1) are not in $\mathcal{B}_{\text{core}}$, and (2) have confidences (conf) above a threshold., θ, where θ is significantly higher than a random guess.

To reflect the relative sensitivity of attributes, we define a weight function $\textsc{weight} : A \rightarrow [0,1]$ that quantifies an attribute's importance relative to the other attributes. Hence, $\sum_{A_i \in A} \textsc{weight}(A_i) = 1$. For ease of exposition, we assume that WEIGHT is fixed for all users of the system. (We note, however, that adjusting WEIGHT for each person requires only a minimal modification to our model.)

We define the **information accessibility score** χ due to the web footprint as the weighted sum of the learned beliefs and confidence values: $\chi = \sum_{\langle A_i, \alpha_i^j, \text{conf}\rangle \in \mathcal{W} \setminus \mathcal{B}_{\text{core}}} \big(\textsc{weight}(A_i) \cdot \text{conf}\big)$. If $\mathcal{W} \setminus \mathcal{B}_{\text{core}} = \varnothing$, we set $\chi = 0$.

We remark that $\chi \geq 0$, with larger values reflecting higher *accessibility*. Notice that χ does not incorporate accuracy. Instead, it represents the ease of determining information and the adversary's confidence in the non-core attribute values discovered. To measure the accuracy of these beliefs, we define the **information exposure score** Sc to be the fraction of beliefs in $\mathcal{W}$ that are accurate, weighted by the attributes' importance:

$$Sc(\mathcal{W}) = \frac{\sum_{\langle A_i, \alpha_i^j, \text{conf}\rangle \in \mathcal{B}_{\text{true}}} \textsc{weight}(A_i)}{\sum_{\langle A_i, \alpha_i^j, \text{conf}\rangle \in \mathcal{W}} \textsc{weight}(A_i)}$$

As will be shown in Section V, these two scores provide a clear assessment of the adversary's discovered knowledge.

IV. Public Information Exposure Detection

We now formally define this specific data mining privacy problem related to publicly available data and refer to it as *public information exposure detection* or PIE detection.

Definition 1: **Public Information Exposure Detection:** Using a set of prior knowledge ($\mathcal{B}_{\text{core}}$) for a given person P, identify any attribute values ($\overline{\alpha} \setminus \mathcal{B}_{\text{core}}$) belonging to P across different sites S with confidence greater than a threshold θ.

In other words, the goal of public information exposure detection is to determine correct attribute values with high confidence (i.e., having a confidence above θ) using prior knowledge $\mathcal{B}_{\text{core}}$ and publicly accessible data obtained from different websites S.

A. Algorithm Overview

Our approach augments traditional structured attribute inference with three complementary methods: pattern-based inference (Section IV-C), distributed probabilistic-join inference (Section IV-D), and population-based inference (Section IV-E). Our high level algorithm for PIE detection (shown in Algorithm 1) collects information about a person from different public websites. The input to our algorithm is the set of core attributes, $\mathcal{B}_{\text{core}}$, the minimum confidence thresholds[1] for probabilistic joins ($\theta_{\text{cross-site}}$) and population inferences (θ_{site}), and the set of public websites to search, S. The output of the algorithm for a person P is the web footprint $\mathcal{W}$.

The algorithm begins by assigning the initial set of beliefs based on $\mathcal{B}_{\text{core}}$ to the web footprint $\mathcal{W}$ (line 4). Each of these beliefs has a confidence of 1. We also initialize a candidate set of beliefs $\mathcal{B}_{\text{cand}}$ (line 5) and a set p of profiles (line 6). In lines 7 and 8, the algorithm queries each site to find profiles that contain the attribute values in $\mathcal{B}_{\text{core}}$, adding the resulting profiles to set p. Next, in lines 9–11, we iterate through all the unstructured (text) attributes in any of the returned profiles and

[1] Algorithm 1 permits different thresholds for beliefs derived using probabilistic joins and population inferences. For clarity, our definition of information exposure detection (Section III-C) assumed a single threshold θ.

Algorithm 1 Information Exposure Detection Algorithm

```
1: Input: $\mathcal{B}_{\text{core}}$, $\theta_{\text{cross-site}}$, $\theta_{\text{site}}$, $S$
2: Output: $\mathcal{W}$
3:
4: $\mathcal{W} \leftarrow \mathcal{B}_{\text{core}}$
5: $\mathcal{B}_{\text{cand}} \leftarrow \varnothing$
6: $p \leftarrow \varnothing$    ▷ set of profiles to consider
7: for all $s_i$ in $\mathcal{S}$ do    ▷ find profiles on site $s_i$ that match $\mathcal{B}_{\text{core}}$
8:     $p \leftarrow p \cup$ GATHER_PROFILES($\mathcal{B}_{\text{core}}$, $s_i$)
9: for all $p_i$ in $p$ do    ▷ infer values from unstructured text
10:     for all $\langle A_j, \alpha_j \rangle$ in $p_i$ s.t. $A_j$ is an unstructured attribute do
11:         EXTRACT_STRUCTURED_VALUES($\alpha_j, p_i$)
12: repeat
13:     for all $\alpha_j^i$ in $p$ do    ▷ iterate over values in all profiles
14:         $\mathcal{B}_{\text{cand}} \leftarrow$ DETERMINE_DEPENDENCIES($\alpha_j^i, p$)
15:         $\mathcal{W} \leftarrow$ UPDATE_WEBFOOTPRINT($\mathcal{B}_{\text{cand}}$, $\theta_{\text{cross-site}}$)
16:         $\mathcal{B}_{\text{cand}} \leftarrow \mathcal{B}_{\text{cand}} - \mathcal{W}$    ▷ remove beliefs where conf $\geq \theta_{\text{cross-site}}$
17:     for all $b_j$ in $\mathcal{B}_{\text{cand}}$ do    ▷ iterate over low confidence beliefs
18:         $\mathcal{B}_{\text{cand}} \leftarrow$ COMPUTE_POPULATION_INFERENCE($b_j$)
19:         $\mathcal{W} \leftarrow$ UPDATE_WEBFOOTPRINT($\mathcal{W}, \mathcal{B}_{\text{cand}}$, $\theta_{\text{site}}$)
20: until $\mathcal{W}$ does not change
21: return $\mathcal{W}$
```

Algorithm 2 Pattern-Based Attribute Detection

```
Input: $\mathcal{C}$, $A_U$
Output: $\mathcal{P}$

$\mathcal{C}_{\text{seed}} \leftarrow$ SELECT_RELEVANT_TEXTS($\mathcal{C}, A_U$)
repeat
    $\mathcal{P} \leftarrow$ EXTRACT_RELEVANT_PATTERNS($\mathcal{C}_{\text{seed}}$)
    $(\mathcal{P}, \mathcal{Q}) \leftarrow$ EVALUATE_PATTERNS($\mathcal{P}$)
    if $\mathcal{Q} > \mathcal{Q}_{control}$ then    ▷ find more seeds and patterns
        $\mathcal{C}_{\text{seed}} \leftarrow$ BOOTSTRAP_MORE_SEEDS($\mathcal{P}$)
    else
        return $\mathcal{P}$
until $\mathcal{C}_{\text{seed}} = \varnothing$
return $\mathcal{P}$
```

use our pattern-based attribute detection algorithm (explained in Section IV-C) to identify and extract missing structured attribute values. Learned structured values are "inserted" into the corresponding profiles. At this stage, we have our possible set of values for the final web footprint.

In lines 13–16, the algorithm applies probabilistic operators (explained in Section IV-D) to all attribute values in the collected profile to determine dependencies between attribute values. Conceptually, the probabilistic operators use site-level and cross-site inference techniques to (1) infer additional attribute values and (2) assign confidences to those values. The resulting set of beliefs are stored in $\mathcal{B}_{\text{cand}}$ (line 14) and added to the web footprint iff the belief's confidence is at least $\theta_{\text{cross-site}}$ (line 15).

Then, for the set of beliefs that still have lower confidence, we use the population inference engine to see if we can improve our confidence in these different beliefs or learn other new ones (lines 17–19). After using the population inference engine (line 18; explained in Section IV-E), the set of beliefs is rechecked to determine if any additional beliefs should be added to the web footprint. The above process (lines 12–20) repeats until no new information can be added to the web footprint, in which case, the algorithm returns the web footprint. In what follows, we will use the running example shown in Table I to explain our approach more clearly.

B. Profile Gathering

This subtask uses attributes in $\mathcal{B}_{\text{core}}$ to (1) query the sites in S using a public API and/or other techniques (e.g., screen scraping), and (2) collect the set of matching profiles p from each site. If an attribute in A contains unstructured data, the Pattern-Based Attribute Detector attempts to identify structured data values using patterns. Otherwise, the next step is to begin the inference process. In our Table I example, the user's first and last name are the attributes in $\mathcal{B}_{\text{core}}$— that is, they are assumed to be known a priori by the adversary. The other attributes – city, favorite color, age, state and sports team – are the ones the adversary is trying to determine. From querying the known first and last name, three, four, and two records (profiles) are respectively returned from Sites 1, 2 and 3.

C. Pattern-Based Inference

Much personally identifiable information exists in public text data (for example, Tweets and blogs). However, these fields are in their natural language form and are not readily usable for inference. We use a pattern-based attribute detection algorithm to extract structured values and represent these extracted values as attribute-attribute value pairs, $\langle A_j, \alpha_j \rangle$. The extraction is done from attributes containing plain text found on different sites, S. Our approach is a bootstrapping approach [6, 21]: Given a free text corpus $\mathcal{C}$ and a set of seed attribute-attribute value pairs $\langle A_j, \alpha_j \rangle$, which we call *instances*, the algorithm outputs an expanded set of new instances $\langle A_j, \alpha_j \rangle$ and a set of lexico-syntactic patterns ($\mathcal{P}$) having a high recognition precision. More specifically, from a few seed instances of labeled attribute-attribute value pairs in $\mathcal{C}$, the bootstrapping algorithm first learns new lexico-syntactic patterns around the seed instances and then uses these new patterns to identify more instances, i.e., more new attribute-attribute value pairs, that map to these newly discovered patterns. The process alternates between using instances to get more patterns and using newly learned patterns to get more instances until high quality patterns and instances are no longer findable.

To clarify our approach, we will focus on a particular pattern-based attribute detector that extracts birthdays from Twitter tweets. A pattern is a regular expression consisting of place holders for instances, and other lexico-syntactic elements for the connecting terms between the attribute type and the attribute value. The place holders in a pattern can be used to mark and expose the instances of interest. Birthday extraction was previously used in the research field of question answering (QA) to answer questions such as "when was Barack Obama born?" Surface level lexical syntactic patterns have been hand-crafted to extract birthday from newswire articles [15]. For instance, two lexicon-syntactic patterns

TABLE I: Example data for web footprint creation.

	User ID	*First Name*	*Last Name*	*City*	*State*	*Favorite Color*	*Age*	*Sports Team*
CORE ($\mathcal{B}_{\text{core}}$)		Mary	Smith					
SITE 1	S1-1	Mary	Smith	Springfield		red	25	
	S1-2	Mary	Smith	Springfield		red		
	S1-3	Mary	Smith	Springfield		red	45	
SITE 2	S2-1	Mary	Smith	Seattle	WA	blue	25	
	S2-2	Mary	Smith		MA	red	45	
	S2-3	Mary	Smith		MA	red	45	
	S2-4	Mary	Smith	Austin	TX	green	25	
SITE 3	S3-1	Mary	Smith		MA			Patriots
	S3-2	Mary	Smith		WA			Seahawks

for birthday question answers are: *<NAME> was born in <BIRTHDATE>* ... and *<NAME> (... born <BIRTHDATE> ...)* for the following text *"Barack Obama was born on August 4, 1961"* and *"Barack Obama (...; born August 4, 1961)"*.

Since tweets are shorter text than traditional news articles, we expect most tweets to lack complete sentence structure. However, as will be demonstrated in Section V, using patterns based on sentences still leads to effective attribute value extraction. Therefore, given a corpus of tweets $\mathcal{C}$, we begin by splitting each tweet into sentences. Each sentence is then parsed using the Stanford NLP Parser(http://nlp.stanford.edu/software/) to obtain the part-of-speech (POS) tags and the named entity (NE) tags. Both of these tags will be used to determine patterns. We then use a bootstrapping-based approach to find birthdays from tweets.

A high level description of our approach is shown in Algorithm 2. Given a corpus $\mathcal{C}$ and attribute $A_U = \{\text{birthday}\}$, we begin by extracting all the tweets of a particular user that are relevant to a particular attribute of interest, A_U (line 4). These tweets become our initial seed tweets ($\mathcal{C}_{\text{seed}}$). In the case of birthday, one can select the tweets that have the term 'birthday' or 'b-day' in them. Then the algorithm looks for relevant patterns in these text instances (line 6). Specifically, we accomplish this by finding commonalities among a subset of the instances in $\mathcal{C}_{\text{seed}}$ and creating expressions/patterns based on the common structures. While the initital patterns use instance values, the patterns are generalized using POS tags in conjunction with instance values. Table II contains examples of high precision lexicon-syntactic patterns that are identified using this approach. For the example tweet – @usera RT @userb: happy birthday @dindoos !! best wishes for youuuu ;D – the structured attributes extracted are $\langle \text{birthday\_person}, \text{dindoos} \rangle$ and $\langle \text{friend\_birthday\_date}, \text{2009-10-01} \rangle$. These patterns are then evaluated in the corpus to see if expected attribute values are returned (line 7). A low quality pattern can easily pollute the instances we would like to extract and the pattern set. Therefore it is important to control the quality of a new pattern.

Our approach tests the quality of newly learned patterns using two statistical measures: information gain and point-wise mutual information. Only patterns that pass a strict pattern selection threshold $\mathcal{Q}_{\text{control}}$ can proceed to the next stage. If the quality of the pattern, $\mathcal{Q}$, is above the minimum threshold, $\mathcal{Q}_{\text{control}}$, then the pattern is retained. This high quality pattern can then be used to generate more patterns (lines 7 and 8). We iteratively keep generating patterns, evaluating them, and then using them to seed more patterns until no additional high quality patterns can be extracted.

D. Inference Using Distributed Probabilistic Operators

We now add to the core set of beliefs using attribute values from queried sites (i.e., the profiles in σ_X) by introducing different dependencies and operators. Intuitively, we use these dependencies and operators to infer attribute values using the following rules: (1) if attribute values for a particular attribute are common for a large fraction of profiles in p, those attribute values can be added to the web footprint (site-level inference), and (2) using common attribute values across sites, we can identify additional beliefs (cross-site inference).

Site-level inference. In relational theory, a functional dependency $A \rightarrow B$ is a mapping between two sets of attributes, A and B, where the values of B are uniquely determined by the values of A. Similarly, we define a *value matching dependency* $a_I \overset{\theta}{\Longrightarrow} a_J$ to be between two sets of values a_I and a_J such that when the values in a_I are all considered true, the values in a_J are determined to be true with a confidence of at least θ. For example, if $\{\text{Joe}, \text{Smith}\} \overset{0.85}{\Longrightarrow} \{\text{male}\}$, then when "Joe" and "Smith" are the respective values for attributes first and last name, then the gender attribute value is "male" with confidence of at least 0.85.[2]

Our approach finds the value matching dependencies for person P on a single site s_k using values in $\mathcal{B}_{\text{core}}$ as the determinant, thereby using the set of true beliefs in $\mathcal{W}$ to inform us about other possible beliefs. Using these dependencies, we propose a new operator – *probabilistic value closure* – for finding other possible beliefs. Similar to the traditional relational closure operator, probabilistic value closure begins with the set $\mathcal{B}_{\text{core}}$ and iteratively adds attribute values with value matching dependencies to the set. The basic algorithm proceeds as follows. The probabilistic value closure starts with $M = \mathcal{B}_{\text{core}}$. For every value matching dependency $a \overset{\theta}{\Longrightarrow} b$ such that a is a subset of M and b is not, b is added to M. The process repeats until no new attribute values can be added to M. We denote the probabilistic value closure of $\mathcal{B}_{\text{core}}$ as $\{\mathcal{B}_{\text{core}}\}^+_\theta$. Once $\{\mathcal{B}_{\text{core}}\}^+_\theta$ is computed, the non-core attribute value in $\{\mathcal{B}_{\text{core}}\}^+_\theta$ can be added as beliefs to $\mathcal{W}$.

[2] We pause to mention that a parallel mapping can exist between multi-valued dependencies and value matching dependencies.

TABLE II: Lexicon-syntactic patterns for BIRTHDAY.

Pattern-1: Someone elses bday w/ mention	$\{happy\vert Happy\vert HAPPY\}\{birthday\vert Birthday\vert BIRTHDAY\}@SOMEONE$
Pattern-2: Someone elses bday in retweet w/ mention	$\{@SOMEONE\}^n * \{happy\vert Happy\vert HAPPY\}$ $\{birthday\vert Birthday\vert BIRTHDAY\} * \{@SOMEONE\}^m$
Pattern-3: Person's own birthday	my birthday is $\{in\vert on\vert \epsilon\}[TimeExpression]$

For our example in Table I, on site 1 for $\mathcal{B}_{\text{core}} = \{\text{Mary}, \text{Smith}\}$, $\{\mathcal{B}_{\text{core}}\}^+_\theta = \{\text{Mary}, \text{Smith}, \text{Springfield}, \text{red}\}$. From this, the following beliefs are added to the web footprint: $\langle\text{city}, \text{Springfield}, 1\rangle$ and $\langle\text{favorite color}, \text{red}, 1\rangle$.

Cross-site inference. Inorder to infer attribute values that are shared across sites in p, we also introduce a *web footprint join* operator. Let each site s_k be viewed as a virtual relation containing a tuple for each profile p^i_k. The web footprint join operator pairs tuples from site s_k with tuples from site s_j using an equity predicate on common attributes. A resulting relation R contains tuples having the same attribute values for common attributes. The probabilistic value closure can then be computed on the attribute values in $\mathcal{W}$ using relation R to identify new beliefs that have high confidence and should be added to $\mathcal{W}$. Conceptually, this approach repeatedly creates temporary relations based on common attribute, attribute values pairs found across different sites.

More specifically, using the web footprint, incompatible profiles (those with conflicting attribute values) are initially removed from p. The remaining profiles on each site represent a virtual relation containing tuples with useful data to infer. These virtual relations are joined to each other using the beliefs in $\mathcal{W}$. Then a probabilistic value closure is computed on the common attribute, attribute value pairs. If the confidence in any of the identified pairs is above $\theta_{\text{cross-site}}$, a new belief is added to $\mathcal{W}$. This process continues until all the pairs of sites have attempted to generate rules and no additional rules are possible to generate.

Returning to our example, after the closure operation is completed at each site, we generate a virtual relation for each site is created. The virtual relations for sites 1, 2, and 3 contain 3, 2, and 2 tuples in them, respectively. A web footprint join for the virtual relations on site 1 and site 2 results in adding $\langle\text{age}, 45\rangle$ and $\langle\text{state}, \text{MA}\rangle$ to $\mathcal{W}$. Joining sites 2 and 3 results in the Patriots being added to $\mathcal{W}$, while joining sites 1 and 3 results in the empty set. Using single-site and cross-site inference, the adversary increases the target's web footprint.

E. Population-Based (Macro-Analysis) Inference

In this section, we propose using frequent patterns of attribute values of a site's broader population to infer some common attributes of P. We accomplish this by creating an inference engine that samples subpopulations from different sites and uses the attribute values of these subpopulations to infer additional attribute values for P. Algorithm 3 describes steps the population inference engine takes to generate rules that can be used as beliefs in $\mathcal{W}$. For each site specific population database D_S, the inference engine generates rules by using the core beliefs as input parameters for identifying rules that apply to P and have strong confidence across a large, random population sample. The initial construction of the population inference engine uses Latent Dirichlet Allocation (LDA) and Association Rule mining to generate rules that represent the sample population.

The LDA inference considers each individuals profile as a document and each attribute value as a concept or word. In this way, the LDA model is built so that each person is viewed as a mixture of attribute based *topics*. When beliefs are input, this method searches for users with a 'similar' distribution over the topics. Attributes are inferred from the k users that are most similar using a majority vote. The association rule inference uses population profiles to generate association rules. Each individual in the population is viewed as a transaction and all the attributes for the individual are viewed as items. Large itemsets are found using all the transactions in the database. Then association rules are generated and stored with their support and confidence. When beliefs are input, rules containing antecedents matching the beliefs are identified. If the identified rules have a high support and confidence, the consequent of the rule is returned as a belief.

For our example, suppose the population database stores the following association rule inference - $\langle\text{firstname}, \text{Mary}\rangle$ implies $\langle\text{gender}, \text{female}\rangle$ 90% of the time. This rule can be translated to a belief for $\mathcal{W}$ since Mary is in our core set of beliefs. The population inference engine is constructed offline and is updated only periodically. The adversary thus can apply the population-based inference engine at low amortized cost.

V. Evaluation

We now evaluate our approach to public information exposure detection. We first measure the information accessibility and information exposure of our constructed web footprints using public data sources. Then we assess our pattern-based attribute extraction and our population inference engine.

Information sources. We collected public profile data for our web footprint construction from Google+, LinkedIn, Twitter, and FourSquare. Our ground truth data set maps actual accounts on different sites for specific individuals. To construct the ground truth, we used the about.me API[3]; about.me

Algorithm 3 Population Inference Computation

```
Input: $\mathcal{B}_{\text{core}}$
Output: $\mathcal{B}_{\text{pop}}$

$\mathcal{B}_{\text{pop}} \leftarrow \varnothing$
for all $d_i$ in $D_S$ do
    $R \leftarrow \text{LDA}(\mathcal{B}_{\text{core}}, d_i)$ ▷ Construct rules from LDA, etc.
    $R \leftarrow R \bigcup \text{ASSOCIATION}(\mathcal{B}_{\text{core}}, d_i)$
    $\mathcal{B}_{\text{pop}} \leftarrow \mathcal{B}_{\text{pop}} \bigcup \text{TRANSLATE\_RULES}(\mathcal{B}_{\text{core}}, R)$
return $\mathcal{B}_{\text{pop}}$
```

[3] http://about.me/developer/api/

TABLE III: Ground truth statistics.

Site	# of Profiles	# of Ground Truth Profiles
Google+	264,266	12,964
LinkedIn	71,253	50,109
Twitter	73,439	3916
FourSquare	112,764	6352

TABLE V: Number of True Beliefs

Initial Beliefs ($\mathcal{B}_{\text{core}}$)	Gold	PIE
first name, last name	2	6
first name, last name, gender	3	7
first name, last name, location	3	10
first name, last name, education	4	11
first name, last name, city	4	27
first name, last name, relationship status	4	13
first name, last name, birthday	4	11
first name, last name, college	4	6

offers its users the ability to publicly advertise their unique identifiers for their social media accounts. Because users post this information, e.g. Twitter handle, themselves, we assume it to be accurate.

Table III summarizes the number of profiles collected for each site and the number of ground truth individuals for each site. Not all individuals have accounts on all four social media sites; 1543 ground truth individuals have accounts on all four sites. Unless otherwise indicated, the experiments in Section V-A use this subset of data We pause to mention that in the case of Twitter, the 1543 profiles results in extracting over 400,000 tweets.

Population inference engine data. Our population inference engine is based on 100,000 public profiles from Google+ and 49,823 public profiles from LinkedIn. This data set was collected by querying each service for random names. All returned profiles were added to our corpus for the population inference engine. We emphasize that the collection process for the inference engine does not include any ground truth data that we obtained from about.me.

A. Public Information Exposure and Accessibility

In this experiment we evaluate the accessibility and accuracy of the constructed web footprints. For all the PIE experiments, we assume that the attributes have equal weights and unless otherwise specified, $\theta_{\text{site}} = 0.9$ and $\theta_{\text{cross-site}} = 0.3$ and $\theta_{\text{pop}} = 0.6$. We experimented with different threshold values ranging from 0.5 to 1 for θ_{site}, 0.15 to 0.8 for $\theta_{\text{cross-site}}$, and 0.3 to 0.9 for θ_{pop}. While many of the thresholds led to similar accuracy results, the variability in the TP/FP and TP/TP-Max ratios was more significant and led to the specified thresholds. For brevity, we do not show this analysis.

We test information exposure breaches by considering different initial $\mathcal{B}_{\text{core}}$ sets. Here we focus on the ground truth users that are on all four sites. In Table IV we report three PIE scores for each attribute core averaged over all of the ground truth users: the number of true beliefs, information accessibility, and information exposure. Recall that information accessibility is the weighted sum of the learned beliefs and the confidence values, and the information exposure score is the fraction of beliefs in $\mathcal{W}$ that are accurate, weighted by attribute importance. We see that the exposure for this group of individuals is between 0.83 (when using only name as the initial core beliefs) and 0.96 (when using name, gender, city, location, and education). Adding data to the core that is not considered sensitive increases the information exposure by approximately 13%. This indicates that the sample of people with the same name on these social media sites have different common attribute values. In other words, there is enough variation in common attributes to uniquely identify people with high accuracy if the adversary knows a small number of these attributes. There are times when adding more attribute values to the initial core reduces the information exposure. This results because all of our initital 1543 ground truth profiles do not have all of the same attributes. For example, while 1543 profiles have first and last name, only 1079 have first name, last name, and education.

We also compare our approach to a *gold standard for accuracy* that uses exact-match record linkage (string matching) across profiles from different sites to find new beliefs. This gold standard adds an attribute, attribute value pair only if there is a matching attribute value across two sites for a particular attribute and there is no conflicting attribute value for that attribute. Otherwise, the attribute, attribute value pair is not added. This means that the accuracy will be close to one when we have at least one additional attribute with the name. Table V shows the comparison between the gold standard and PIE detection. While the accuracy of the gold standard is optimal, the number of true beliefs discovered is low, usually no more than one attribute more than the core. In contrast, our approach increases the number of true beliefs significantly, with an increase of between 4 and 24 more beliefs.

When using the population inference engine to infer values for two attributes, *high school attended* and *state*, the inference accuracy ranged from 0% to 50% depending upon the original attribute core. The predictive accuracy was similar for both attributes, but more predictions were made for state as compared to high school. As the attributes in the core increased, the accuracy of the inferred attributes also improved in most cases. In general, though, these results are harder to evaluate since they are only for the subset of predictions for which we had a ground truth value. The majority of predictions made could not be validated. This is highlighted in Figure 2. Here the blue represents true positives, the red represents false positives and the green represents predictions that could not be validated based on the data we had. We did consider predictions for other attributes, including birthday year, college, country, and occupation, but could not assess the quality since they were not in the ground truth when predictions were made.

Finally, we consider the contribution of each components of the framework. The site-level inference and cross-site inference account for the majority of beliefs discovered (77%), both pattern-based inference using Twitter data and population-inference augment the overall set of beliefs by over 20%. In other words, one fifth of the beliefs would not be discovered without the combined framework.

TABLE IV: Avg. public info. exposure and accessibility scores for various $\mathcal{B}_{\text{core}}$.

Initial Beliefs ($\mathcal{B}_{\text{core}}$)	Nbr of True Beliefs	Information Accessibility	Info. Exposure
First Name, Last Name	6	16	0.83
First Name, Last Name, Location	7	11	0.92
First Name, Last Name, Education	10	17	0.85
First Name, Last Name, City	11	16	0.87
First Name, Last Name, Relationship Status	27	38	0.88
First Name, Last Name, Birthday	13	20	0.86
First Name, Last Name, College	11	17	0.87
First Name, Last Name, Gender, Location	6	7	0.9
First Name, Last Name, Gender, Location, City	7	8	0.93
First Name, Last Name, Gender, Location, City, Education	10	11	0.96
F. Name, L. Name, Gender, Loc., City, Edu., Relationship Status	11	12	0.96

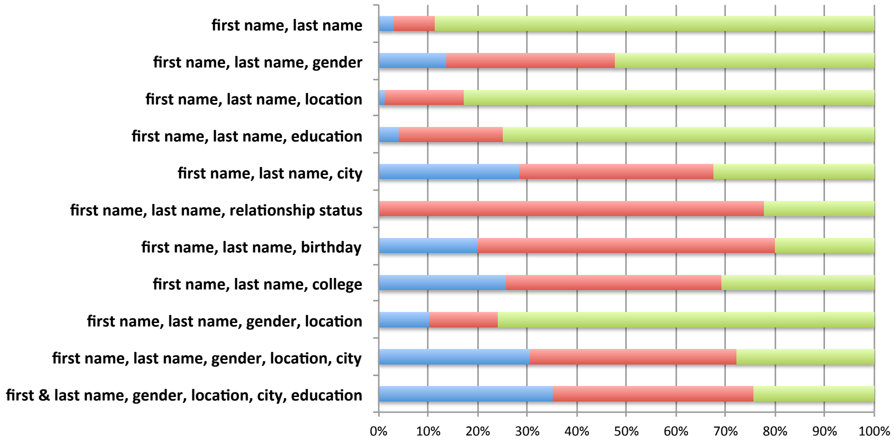

Fig. 2: Accuracy of population inference engine when predicting high school attended and state, for various initial beliefs ($\mathcal{B}_{\text{core}}$).

B. Pattern-Based Attribute Extraction

To evaluate the effectiveness of our pattern-based attribute extraction approach we use Twitter data. The data set[4] used for this analysis contains 467 million Twitter posts (tweets and retweets) from 20 million users over a 7 month period from June 1, 2009 to Dec. 31, 2009. Each post includes the following attributes: the author of the post, the time of the post, and the content of the post. For this analysis, we sampled a four hour period (12 am to 4 am) on Oct. 1 from the initial data set. During this period, there were 549,757 posts.

We extracted 5 attributes from these posts: birth date, birthday person, sports team, location, and brand. While others could also have been extracted, these attributes provided opportunities to connect to data on other sites. For all of these experiments, we set the pattern selection threshold $\mathcal{Q}_{\text{control}}$ to 0.5. As mentioned in Sec. IV-C for birthday attributes, we begin by searching for tweets containing different variations of the term birthday. For the location attribute, we began with 239 countries and 8609 city names (this list is based on locations of different airports around the world). For the brand attribute, Forbes' world's most valuable 100 brands (as of Nov. 2013) are used for pattern-based attribute detection. Finally, for the sports team attribute, we use 218 US and Canadian sports teams from 12 leagues for our initial attribute values.

For the tweets generated by this set of users, over 45,000 structured attributes were extracted using our pattern-based attribute detection algorithm. Our experiments show that the number of birthday tweets far out-paced the number of tweets with recognizable locations and brands (see Figure 3). This is surprising since birthday is considered a more sensitive attribute than location or brand.

In order to evaluate the precision of the extracted attributes (i.e., did we actually find a birthday), we manually examined and annotated a subset of the tweets. For those patterns with 100 or fewer tweets, we manually annotated all of them. For patterns with more than 100 tweets, we randomly selected 100 tweets to annotate and evaluated the precision based on that subsample. Table VI shows the attributes, the pattern, the number of tweets with the pattern in our subsample, and the extraction precision. Our precision is generally above 85%, with exceptions occurring for birthday pattern 2 and location.

False positives for the birthday attribute occur when the user mentioned with the @symbol shares a birthday with

TABLE VI: Coverage & Precision of Pattern-Based Extraction

Structured Attribute	Pattern Interpretation	# of Posts w Pattern	Precision
Birthday	Pattern 1	36	35/36 = 97%
Birthday	Pattern 2	207	84/100 = 84%
Birthday	Pattern 3	36	33/36 = 92%
Brand	Concern/interest in brand	11,095	87/100 = 87%
Sports team	Interest in sport's team	572	99/100 = 99%
Location	Visited location	34,296	51/100 = 51%

[4] http://snap.stanford.edu/data/twitter7.html

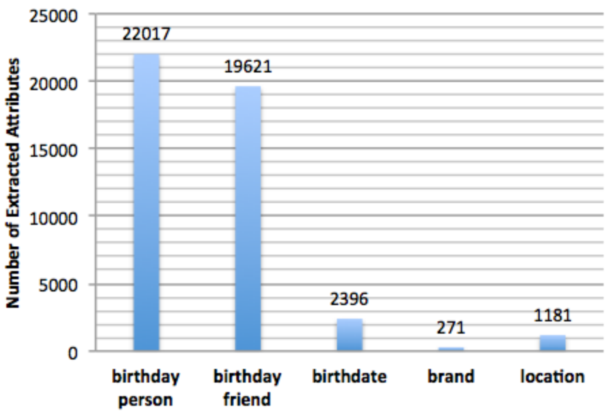

Fig. 3: Number of structured attribute values (y-axis) extracted from tweets organized by pattern type (x-axis).

someone else, leading to confusion about who is actually celebrating the birthday. For the location attribute we are interested in identifying locations visited by the tweeter. For the hand-validated sample, the tweeter had visited the location mentioned in the tweet approximately half of the time. Therefore, even though the locations extracted from the tweets are valid locations, our low precision indicates that we cannot assume the location is one visited by the tweeter. More generally, from our experiments, we see that extracting structured attributes unstructured text not only increases the number of available attributes for inference, it also increases a user's overall exposure.

VI. Conclusion

This paper introduces methods for determining the amount of information that can be ascertained using only publicly accessible data. In particular, we contribute (1) a formalism for reasoning about information exposure due to publicly available information, (2) a framework for determining a user's *web footprint*—a set of beliefs about a user's attributes that may be inferred by an adversary using only public sources of information, and (3) an extensive empirical analysis across multiple social networking sites that highlights how easy it is to reidentify people using very common, public attributes. Although we frame the PIE detection problem in an adversarial context, we emphasize that the techniques can equally be applied by individuals to assess their own exposure. This paper serves as a blueprint for making the risks of data leakage more clear and transparent to web users. Our hope is that this will encourage those with high information exposure and accessibility scores to reduce the amount of information they publicly expose on social media sites.

Acknowledgments:
This work was supported by NSF CNS-1223825. The opinions and findings described in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] J. Anderson, C. Diaz, J. Bonneau, and F. Stajano. Privacy-enabling Social Networking over Untrusted Networks. In *WOSN*, 2009.

[2] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore Art Thou r3579x?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography. In *WWW*, 2007.

[3] R. Baden, A. Bender, N. Spring, B. Bhattacharjee, and D. Starin. Persona: An Online Social Network with User-defined Privacy. In *SIGCOMM*, 2009.

[4] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting Innocuous Activity for Correlating Users Across Sites. In *WWW*, 2013.

[5] R. Gross and A. Acquisti. Information Revelation and Privacy in Online Social Networks. In *WPES*, 2005.

[6] M. A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *COLING*, 1992.

[7] M. Humbert, T. Studer, M. Grossglauser, and J.-P. Hubaux. Nowhere to Hide: Navigating around Privacy in Online Social Networks. In *ESORICS*, 2013.

[8] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff. Identifying Users Across Social Tagging Systems. In *ICWSM*, 2011.

[9] D. Irani, S. Webb, K. Li, and C. Pu. Large Online Social Footprints–An Emerging Threat. In *International Conference on Computational Science and Engineering*, 2009.

[10] P. Jain, P. Kumaraguru, and A. Joshi. @I Seek 'fb.me': Identifying Users Across Multiple Online Social Networks. In *WoLE*, 2013.

[11] A. Malhotra, L. Totti, W. Meira Jr., P. Kumaraguru, and V. Almeida. Studying User Footprints in Different Online Social Networks. In *ASONAM*, 2012.

[12] B. Moore, Y. Wei, A. Orshefsky, M. Sherr, L. Singh, and H. Yang. Understanding Site-Based Inference Potential for Identifying Hidden Attributes. In *PASSAT*, 2013.

[13] A. Narayanan and V. Shmatikov. De-anonymizing Social Networks. In *IEEE Symposium on Security and Privacy*, 2009.

[14] A. Ramachandran, L. Singh, E. Porter, and F. Nagle. Exploring Re-identification Risks in Public Domains. In *IPST*, 2012.

[15] D. Ravichandran and E. Hovy. Learning Surface Text Patterns for a Question Answering System. In *ACL*, 2002.

[16] J. Song, S. Lee, and J. Kim. I Know the Shortened URLs you Clicked on Twitter: Inference Attack using Public Click Analytics and Twitter Metadata. In *WWW*, 2013.

[17] B. H. Spitzberg and G. Hoobler. Cyberstalking and the Technologies of Interpersonal Terrorism. *New Media and Society*, 4:71–92, February 2002.

[18] L. Sweeney. k-anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10:557–570, October 2002.

[19] L. Sweeney. Protecting Job Seekers from Identity Theft. *IEEE Internet Computing*, 10(2), Mar. 2006.

[20] C. Warren. 10 People Who Lost Jobs Over Social Media Mistakes, 2011. Mashable. Available at http://mashable.com/2011/06/16/weinergate-social-media-job-loss/.

[21] H. Yang and J. Callan. A Metric-based Framework for Automatic Taxonomy Induction. In *ACL*, 2009.