Helping Users Understand Their Web Footprints

Lisa Singh[*], Hui Yang[†], Micah Sherr[‡], Yifang Wei, Andrew Hian-Cheong,
Kevin Tian, Janet Zhu, Sicong Zhang, Tavish Vaidya, Elchin Asgarli
Department of Computer Science, Georgetown University, Washington, DC 20057

ABSTRACT

To help users better understand the potential risks associated with publishing data publicly, and the types of data that can be inferred by combining data from multiple online sources, we introduce a novel information exposure detection framework that generates and analyzes the *web footprints* users leave across the social web. We propose to use probabilistic operators, free text attribute extraction, and a population-based inference engine to generate the web footprints. Evaluation over public profiles from multiple sites shows that our framework successfully detects and quantifies information exposure using a small amount of non-sensitive initial knowledge.

1. INTRODUCTION

This poster examines this problem of quantifiably measuring online privacy risks by proposing a framework that constructs *web footprints* (as would an adversary stalking a user) and reports to the user her particular level of vulnerability based on her publicly shared information. Our framework (see Figure 1) first creates the user's web footprint by combining publicly accessible information from social media, micro-blogs, data aggregation sites, etc. Since much web data is unstructured, we also introduce a *pattern-based attribute extractor* that bootstraps patterns from text and then extracts structured attribute values based on them, thereby increasing the amount of usable information for web footprint construction. In addition, probabilistic inference logic is applied to supplement web footprints with probable attribute value pairs learned via algebraic dependencies between attribute values in profiles on different sites. Finally, we infer the user's attribute values by site-level population.

2. OUR APPROACH

We assume a person P, about whom an adversary is attempting to learn information, has publicly revealed certain 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). We frame the public information exposure (PIE) 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 has some background knowledge about P (e.g., P's name) and that he uses only publicly available information about P to form *beliefs* about P that are not originally known to the adversary. To learn information about P, we allow the adversary to query a set of sites $S = \{s_1, s_2, \dots, s_q\}$, e.g. online social networks, search engines, and data aggregation sites.

Algorithm Overview. There has been an emerging interest in linking individuals across online social networks (OSNs), including [1–5]. Our approach augments traditional structured attribute inference with three complementary methods: pattern-based inference (to extract attributes from text), distributed probabilistic-join inference (to map profiles across sites), and population-based inference (to incorporate information about norms in the population). Our experiments show that augmenting standard record linkage with these inference techniques increases the number of discovered beliefs and more accurately models a real adversary. To the best of our knowledge, we are the first to propose this holistic methodology for problems in this area.

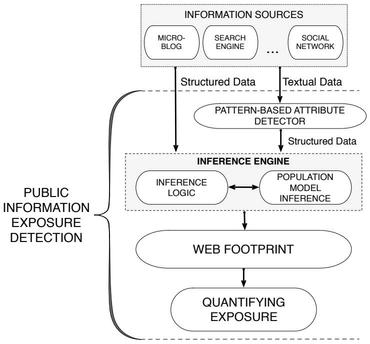

Figure 1: The PIE framework.

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 for probabilistic joins ($\theta_{\text{cross-site}}$) and population inferences (θ_{site}), and the set of public websites to search, S. The algorithm outputs person P's web footprint $\mathcal{W}$.

The algorithm begins by assigning the initial set of beliefs based on $\mathcal{B}_{\text{core}}$ to the web footprint $\mathcal{W}$. Each site is queried to find profiles that contain the attribute values in $\mathcal{B}_{\text{core}}$, adding the resulting profiles to set p. Next, it iterates through all the unstructured (text) attributes in any of the returned profiles and uses a pattern-based attribute detection algorithm to identify and extract missing struc-

[*] singh@cs.georgetown.edu
[†] huiyang@cs.georgetown.edu
[‡] msherr@cs.georgetown.edu

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742763.

Algorithm 1 Information Exposure Detection Algorithm

```
Input: $\mathcal{B}_{\text{core}}$, $\theta_{\text{cross-site}}$, $\theta_{\text{site}}$, $S$
Output: $\mathcal{W}$

$\mathcal{W} \leftarrow \mathcal{B}_{\text{core}}$
$\mathcal{B}_{\text{cand}} \leftarrow \varnothing$
$p \leftarrow \varnothing$                                          ▷ set of profiles to consider
for all $s_i$ in $\mathcal{S}$ do                    ▷ find profiles on site $s_i$ that match $\mathcal{B}_{\text{core}}$
    $p \leftarrow p \cup$ GATHER_PROFILES($\mathcal{B}_{\text{core}}$, $s_i$)
for all $p_i$ in $p$ do                              ▷ infer values from unstructured text
    for all $\langle A_j, \alpha_j \rangle$ in $p_i$ s.t. $A_j$ is an unstructured attribute do
        EXTRACT_STRUCTURED_VALUES($\alpha_j, p_i$)
repeat
    for all $\alpha_j^i$ in $p$ do                       ▷ iterate over values in all profiles
        $\mathcal{B}_{\text{cand}} \leftarrow$ DETERMINE_DEPENDENCIES($\alpha_j^i, p$)
        $\mathcal{W} \leftarrow$ UPDATE_WEBFOOTPRINT($\mathcal{B}_{\text{cand}}$, $\theta_{\text{cross-site}}$)
        $\mathcal{B}_{\text{cand}} \leftarrow \mathcal{B}_{\text{cand}} - \mathcal{W}$          ▷ remove beliefs where conf ≥ $\theta_{\text{cross-site}}$
    for all $b_j$ in $\mathcal{B}_{\text{cand}}$ do                  ▷ iterate over low confidence beliefs
        $\mathcal{B}_{\text{cand}} \leftarrow$ COMPUTE_POPULATION_INFERENCE($b_j$)
        $\mathcal{W} \leftarrow$ UPDATE_WEBFOOTPRINT($\mathcal{W}$, $\mathcal{B}_{\text{cand}}$, $\theta_{\text{site}}$)
until $\mathcal{W}$ does not change
return $\mathcal{W}$
```

Table 1: 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

tured attribute values (the detection algorithm uses bootstrapped patterns found in a large corpus to find user data that match the patterns and extracts the attributes within the patterns). Learned structured attributes are "inserted" into the corresponding profiles.

The algorithm then applies site-level and cross-site inference techniques to infer additional attribute values and assign confidences to those values. The resulting set of beliefs are added to the web footprint iff the belief's confidence exceeds a minimum threshold. For the set of beliefs that 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 based on norms found in population data. The above process repeats until no new information can be added to the web footprint.

3. EXPERIMENTS

We evaluated our approach to PIE detection using public profile data from Google+, LinkedIn, Twitter, and FourSquare. We generated a ground truth data set using the about.me API that maps actual accounts on different sites for specific individuals. Table 1 summarizes the number of profiles collected for each site and the number of ground truth individuals for each site. Our population inference engine is based on 100,000 public profiles from Google+ and 49,823 public profiles from LinkedIn.

Public Information Exposure and Accessibility. We test information exposure breaches by considering different initial $\mathcal{B}_{\text{core}}$ sets, beginning with just first name and last name, and then consider attribute cores that include one or more additional attributes (location, education, city, relationship status, birthday, college, gender). We compute three PIE scores for each attribute core averaged over all of the ground truth users that are on all four sites: the number of true beliefs, information accessibility (the weighted sum of the learned beliefs and the confidence values), and information exposure (the fraction of beliefs in $\mathcal{W}$ that are accurate, weighted by attribute importance). Due to space limitations, we cannot present all results for all combinations tested. When adding more attributes the number of true beliefs increases from 6 (when using only name as the initial core beliefs) to between 7 and 27 (when using name and relationship status). We also find that information accessibility is 16 when using only name as the initial core beliefs). It sometimes decreases to as low as 11 when additional attributes are added, but usually increases (to as high as 38). Finally, we find 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 *there is enough variation in common attributes to uniquely identify people with high accuracy if the adversary knows a small number of these attributes*.

Table 2: 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

We also compare our approach to a gold standard for accuracy that uses exact-match record linkage (string matching) across the profiles from different sites to find new beliefs. The gold standard adds an attribute, attribute value pair 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. Our interest is in understanding the impact of using this strict approach on the number of true beliefs discovered. Table 2 shows this comparison. We see that 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.

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%; many of these beliefs would not be discovered without the combined framework.

4. CONCLUSION

There has been little work that examines how much information can be derived from the data that we publish *openly* and *publicly* online. This poster proposes an approach to determine a user's *web footprint*– beliefs inferred by an adversary. An empirical analysis across multiple social networking sites highlights how easy it is to re-identify people using our approach. We hope that our framework will make the risks of data leakage more transparent to web users.

Acknowledgments

This work was supported by NSF CNS-1223825 and CNS-1149832.

References

[1] 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.

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

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

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

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