Assessing Online Information Exposure Using Web Footprints

- PIs: Lisa Singh, Micah Sherr, Grace Hui Yang (Georgetown University)

http://webfootprinting.cs.georgetown.edu/

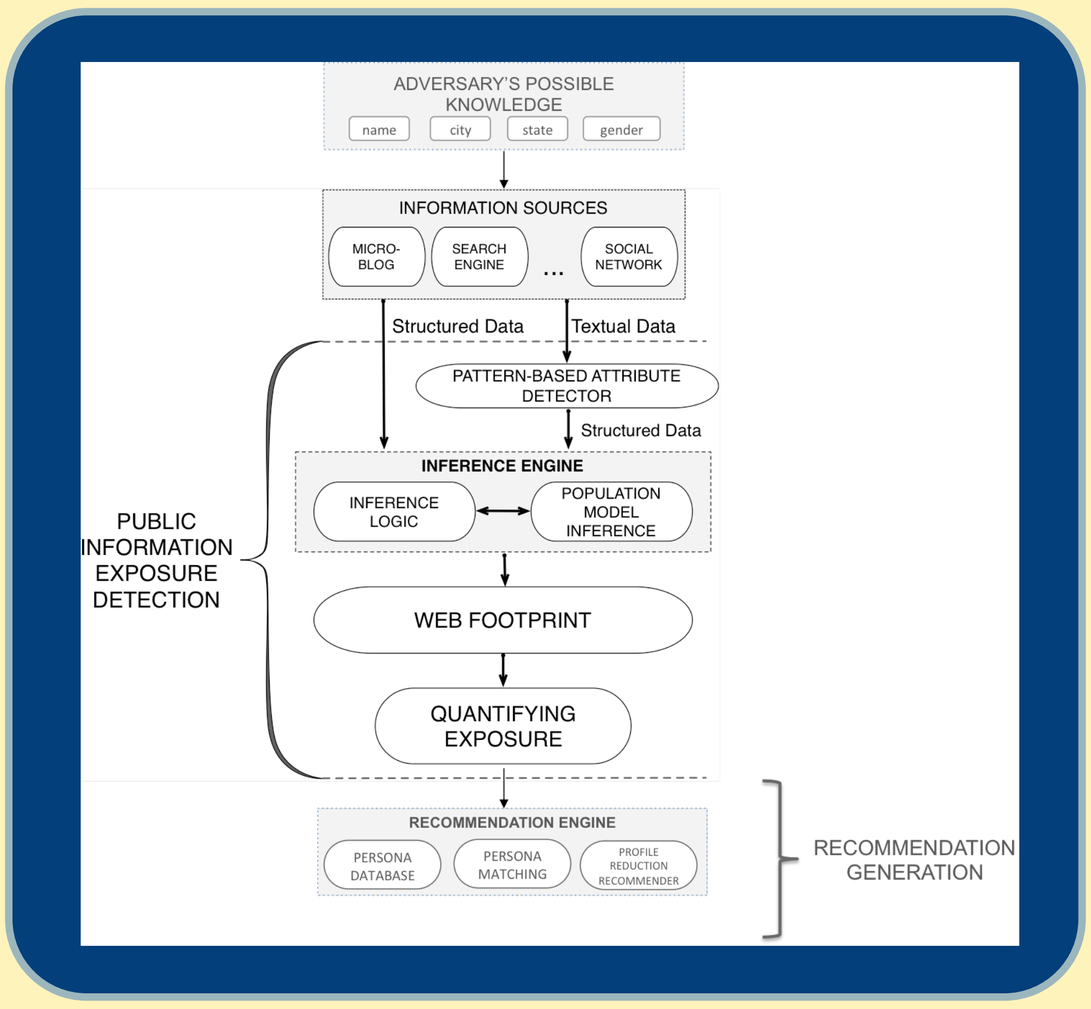
Challenge:

- While people share large amounts of information publicly, they may not understand the potential risks of doing so (stalking, identity theft, job loss, etc.).

Goal:

- Make the risks of data leakage more transparent to web users. Then they can make more informed decisions about what types of information they want to share.

We introduce a novel information exposure detection framework and application that generates and analyzes users' *web footprints*. Then given a user's level of information exposure, we make recommendations about which attributes to remove from her public profiles to reduce the overall inference potential.



Approaches

Public Information Exposure Detection

Our work has introduced probabilistic operators, free text attribute extraction methods, and a population-based inference engine that uses site level statistics to improve inference to generate the web footprints.

Recommendation Generation

We have also introduced persona-based recommendations that reduce the identifiability of the individual, while maintaining utility.

Data Set

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

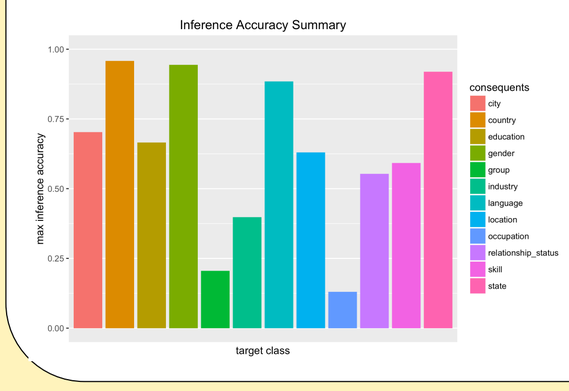
We evaluated our approach to PIE detection using public profile data from Google+, LinkedIn, Twitter, and FourSquare. We generated a ground truth data set using the about.me API that maps actual accounts on different sites for specific individuals.

PIE Scores

Initial Beliefs ($\mathcal{B}_{core}$)	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

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 the web footprint that are accurate, weighted by attribute importance).

Population Inference Engine



As part of our overall framework, the population inference engine allows for different data preparation methods, machine learning algorithms, and ensembles to be exploited, to fully leverage the inference potential of public social media data. The engine begins by learning population norms from public social media data to develop a set of background knowledge. It then applies this background knowledge to infer hidden attributes about a target using the targets public attributes. We see that some attributes are more readily predicted using population level data than others.

Persona-based Recommendations

All Modifications	
Location	1039
Occupation	602
Education	584
Gender	359
Company	344
College	120
Industry	115
City	99
High School	88
Graduation Year	85
Language	67
Relationship Status	62
State	54
Group	51
Country	31
Total	3700

Removals	
Location	740
Occupation	553
Education	469
Company	317
College	110
Industry	103
City	94
Graduation Year	85
High School	83
Language	60
Relationship Status	53
Group	51
State	46
Gender	29
Country	21
Total	2814

Changes	
Location	162
Education	115
Occupation	49
Company	27
Industry	12
Language	6
College	4
Country	1
State	1
High School	1
Total	378

Additions	
Gender	330
Location	137
Relationship	9
Country	9
State	7
College	6
City	5
High School	4
Language	1
Total	378

Personas are frequently occurring sets of attribute value pairs. In this experiment, we make modifications to match profiles to pre-computed personas.

- Personas are generated using 30,000 profiles. Each persona contains at least 30 individuals.
- Experiment set contains 1600 individuals that have attributes on 3 social media sites.
- While different attributes are involved in modification, location is most common.

Interested in meeting the PIs? Attach post-it note below!

The 3rd NSF Secure and Trustworthy Cyberspace Principal Investigator Meeting
January 9-11, 2017
Arlington, Virginia

National Science Foundation
WHERE DISCOVERIES BEGIN

GEORGETOWN UNIVERSITY