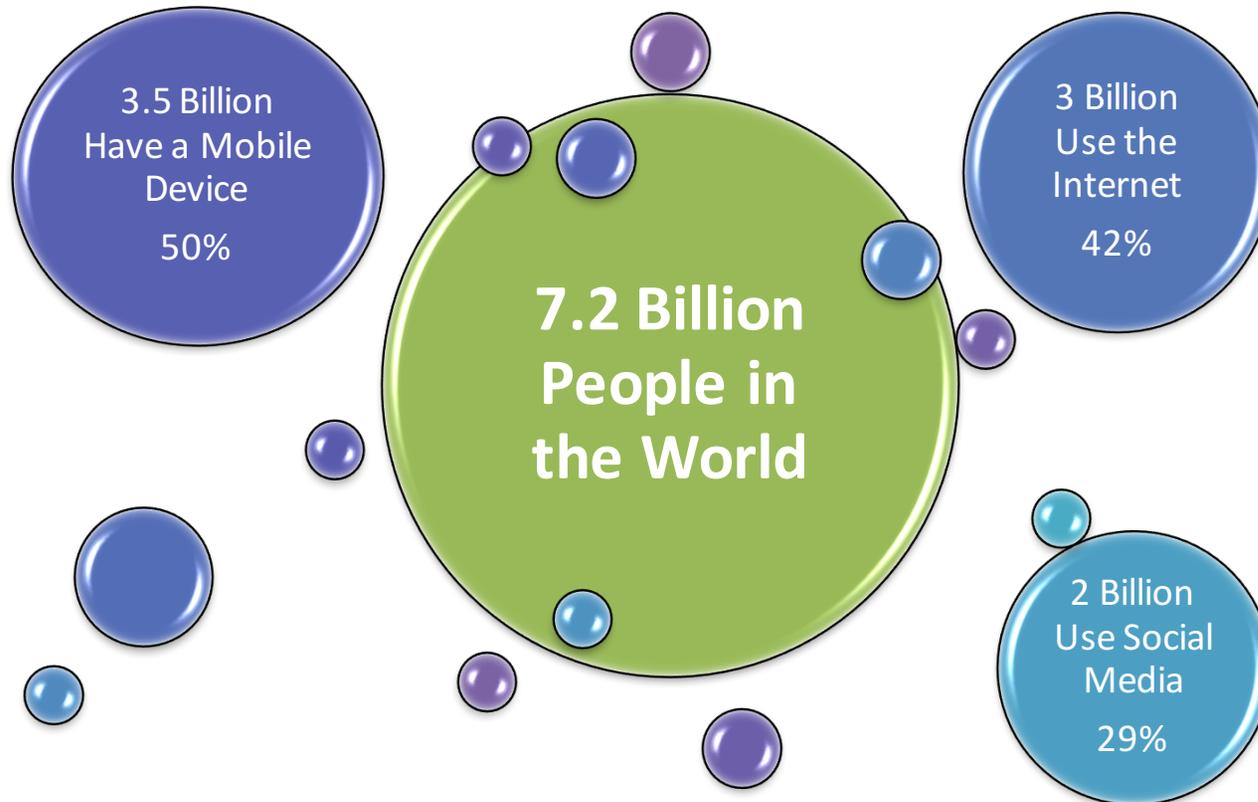# Privacy & Social Media

Lisa Singh, PhD

Department of Computer Science

Georgetown University

GEORGETOWN
UNIVERSITY

Outline

- Our world on the Internet
- Data privacy in a public profile world
- Methods for determining our web footprints
- Taking control of our web identities

Our presence on the Internet and social media

3.5 Billion
Have a Mobile
Device

50%

7.2 Billion
People in
the World

3 Billion
Use the
Internet

42%

2 Billion
Use Social
Media

29%

Data, so much data…

Users share 70 billion pieces of content each month on Facebook

190 million tweets are sent per day

65 hours of video are uploaded to YouTube every minute

Image from http://www.playbuzz.com/jaylam10/which-social-media-fits-your-personality

Privacy settings and social media

- 25% of Facebook users do not bother with any privacy settings (velocitydigital.co.uk, 2013)

- 37% of Facebook users have used the site's privacy tools to customize how much information apps are allowed to see (Consumer reports, 2012)

- 40% of teen Facebook users DO NOT set their Facebook profiles to private (friends only) (Pew Study 2013)
  - 71% post their school name
  - 71% post the city or town where they live
  - 53% post their email address
  - 20% post their cell phone number

Consequences of Over-sharing

- Identity theft
- Online and physical stalking
- Blackmailing
- Negative employment consequences
- Enabling of snoopers

GEORGETOWN UNIVERSITY

Data Privacy Expectations

- We should expect data privacy

- We should expect freedom from unauthorized use of our data

- We should expect freedom from data intrusion.

© MARK ANDERSON   WWW.ANDERTOONS.COM

"Before I write my name on the board, I'll need to know how you're planning to use that data."

GEORGETOWN UNIVERSITY

How informative, linkable, or sensitive is your public profile – your web footprint?

Divorced

Gay

Department of Defense

Washington, DC

Spanish-speaking

Georgetown University

John Smith

Catholic

Software Developer

Republican

John Smith

Your name

Linking data

Facebook

First Name: Sally
Last Name: Smith
Gender: Female
Location: Georgetown
Hometown: Pittsburgh
Favorite Sports Team: Seahawks
Religion: Atheist

Google+

First Name: Sally
Last Name: Smith
Gender: Female
Location: Georgetown
Occupation: Dentist
Relationship Status: Married
Zip code: 22033

Linking data

Facebook Adversary's Beliefs Google+

First Name: Sally
Last Name: Smith
Gender: Female
Location: Georgetown
Hometown: Pittsburgh
Favorite Sports Team: Seahawks
Religion: Atheist

First Name: Sally
Last Name: Smith
Gender: Female
Location: Georgetown
Hometown: Pittsburgh
Occupation: Dentist
Favorite Sports Team:
Religion: Atheist
Relationship Status: Married
Zip Code: 22033

First Name: Sally
Last Name: Smith
Gender: Female
Location: Georgetown
Occupation: Dentist
Relationship Status: Married
Zip Code: 22033

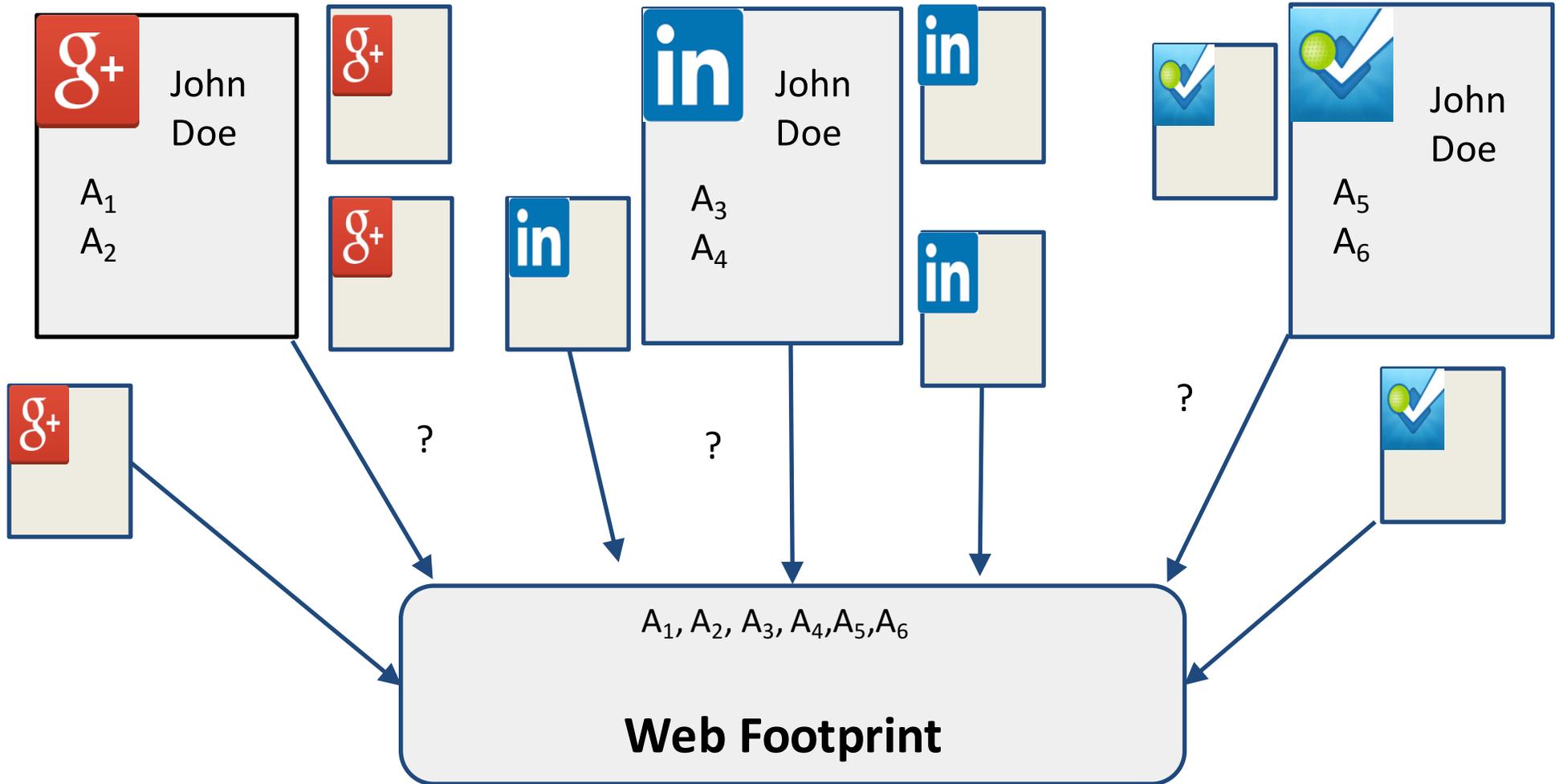# What about friends?

site 1

Starting user

List of names of friends

List of names of friends for given user

site 2
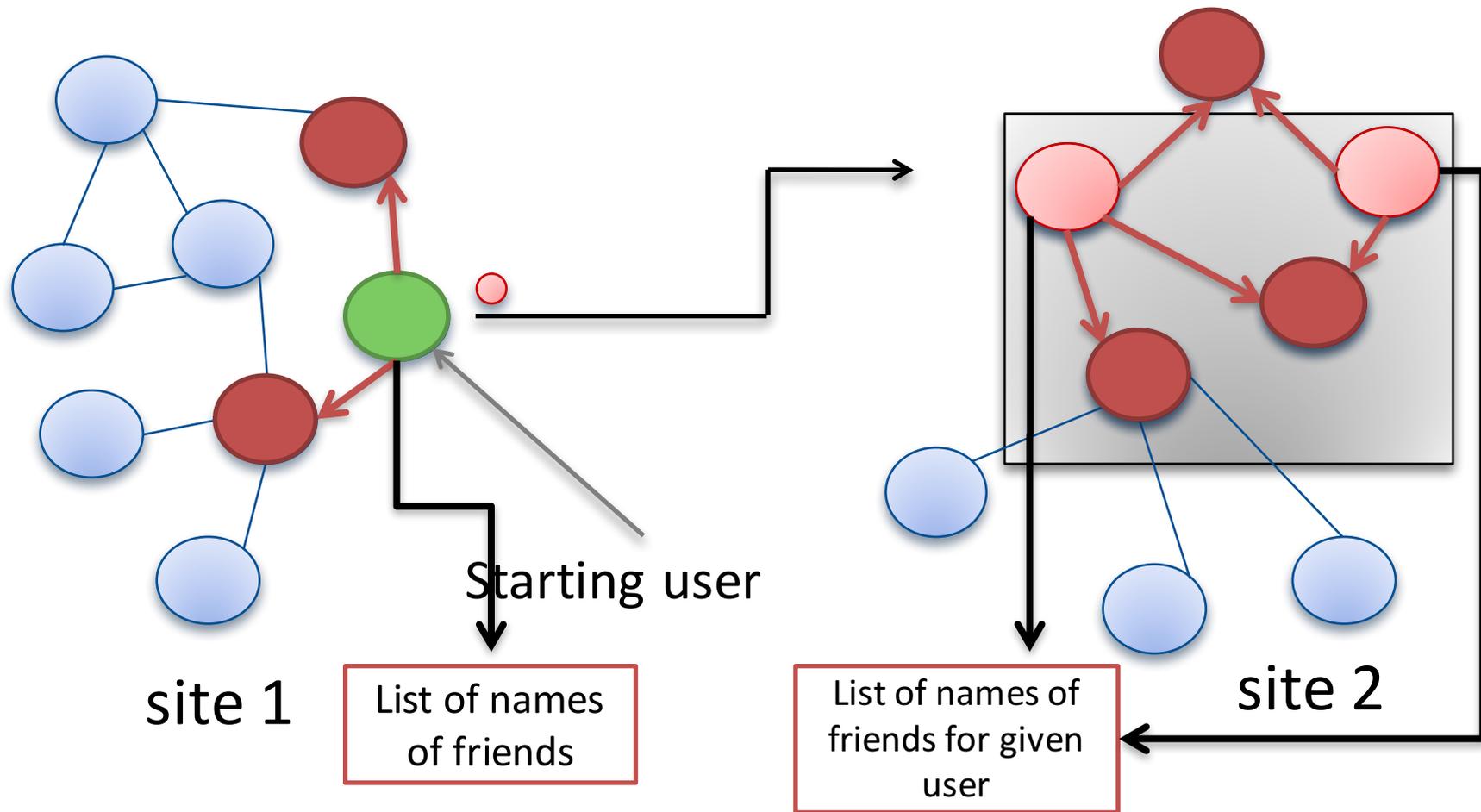
match = number overlapping friends between users

[Ramachandran et al., 2012]

GEORGETOWN UNIVERSITY

Really linking data

Shared Public Attributes

Google+	LinkedIn	FourSquare
• Company	• Company	• Facebook id
• Occupation	• Location	• Twitter handle
• Education	• Education	• Email
• Location	• Email	• Gender
• Birthdate	• Occupation	• Location
• Relationship status	• Skills	• Phone number
• Gender	• Industry	• Relationship status
• Graduation Year	• Website	
	• Languages	

What do group memberships tell us?

What about tweets?

- A special wish for a special girl #HappyBirthday
- I love #Starbuck #MangoTeaLemonade
- Go #Bears!!!!

[Singh et al., 2015]

GEORGETOWN
UNIVERSITY
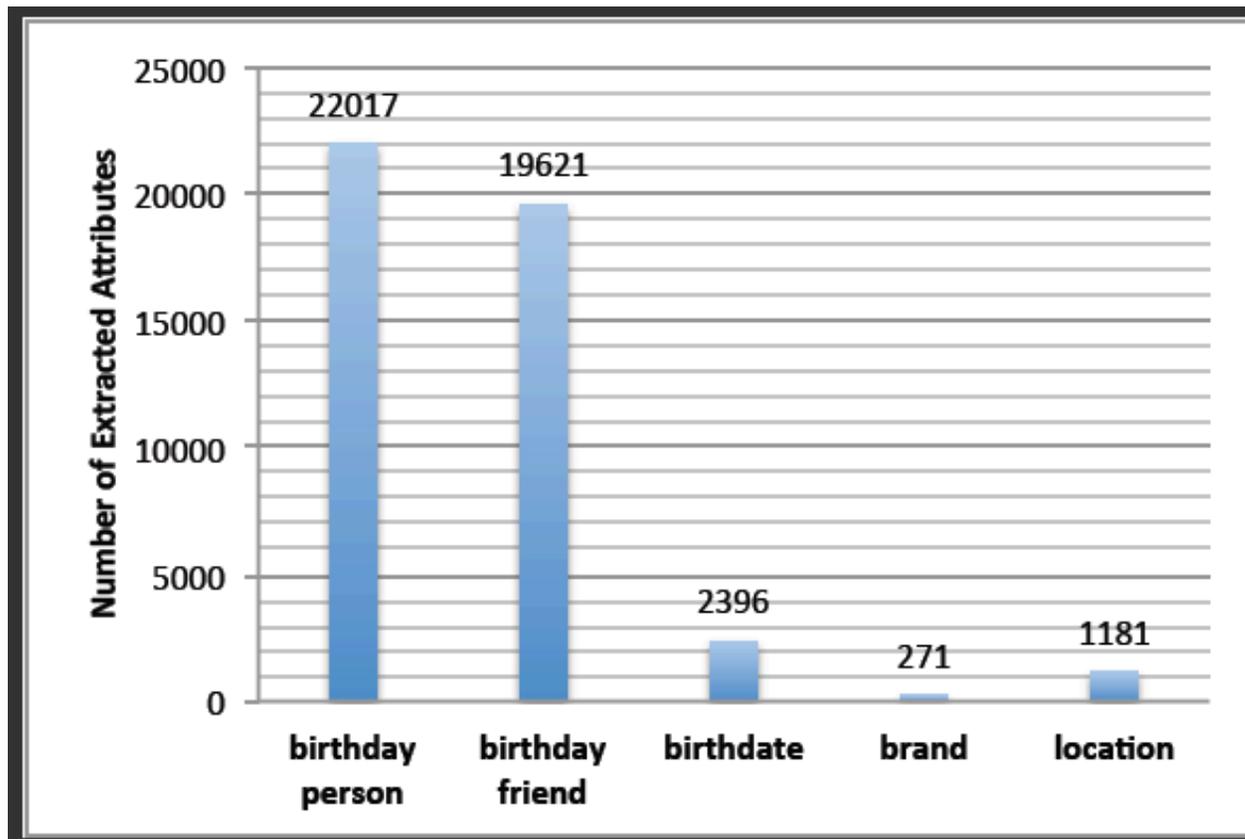
What about the population?

Facebook
- Birthday
- Gender
- Address

LinkedIn
- Skills
- Title
- Industry

Twitter
- Thoughts
- Ideas
- Interests

To what degree can site level data be leveraged to determine the undisclosed attributes of a user?

Methodology

Public Profiles

Inference Engine

Inference Model

$$a,b \rightarrow c$$
$$a,c,d \rightarrow e$$
$$a,d \rightarrow b$$
$$b,c,d,f \rightarrow a$$

User Profile

Inference Model

Hidden Attribute-Values

Step 1:
Subpopulation
Sampling

Step 2:
Inference Engine
Construction

Step 3:
Determination of Hidden
Attribute-Values

- Use user profiles to construct an inference engine containing a set of inference rules.
- Make inferences using the inference engine.

LinkedIn dataset:
  91,150 public profiles
  12 attributes per profile

Inference gain

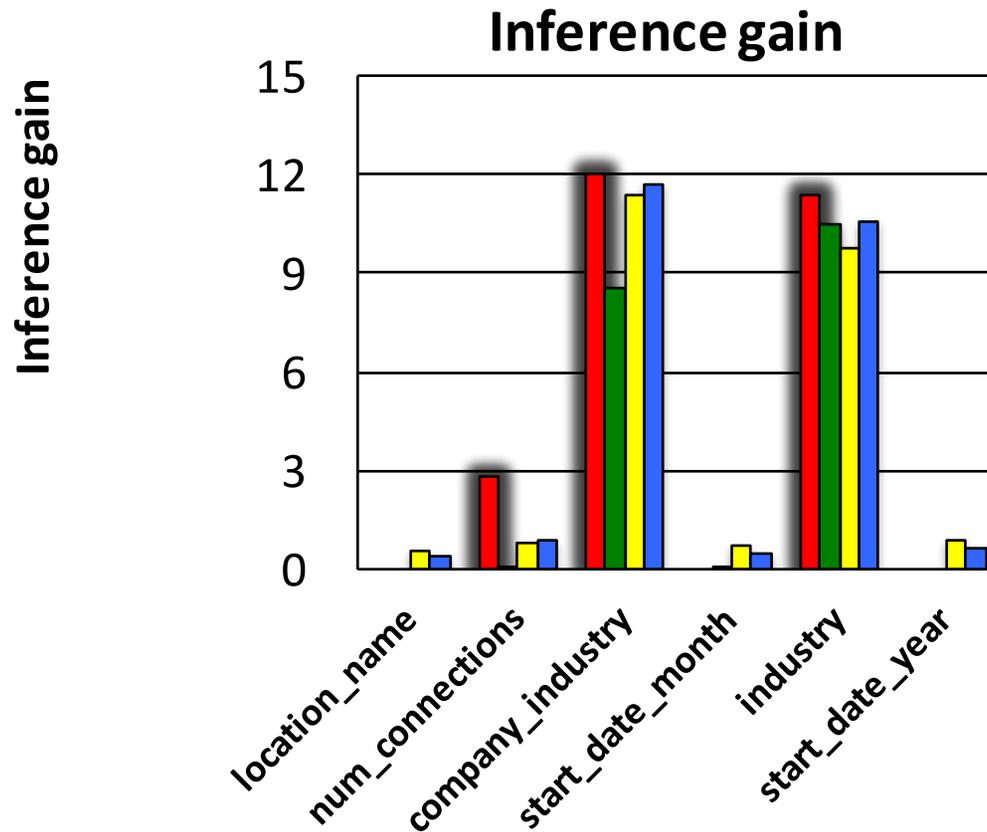**Legend:** Association Rule, Naïve Bayes, LDA, Ensemble

[Moore et al., 2013]

What can be inferred from the population?

GEORGETOWN
UNIVERSITY
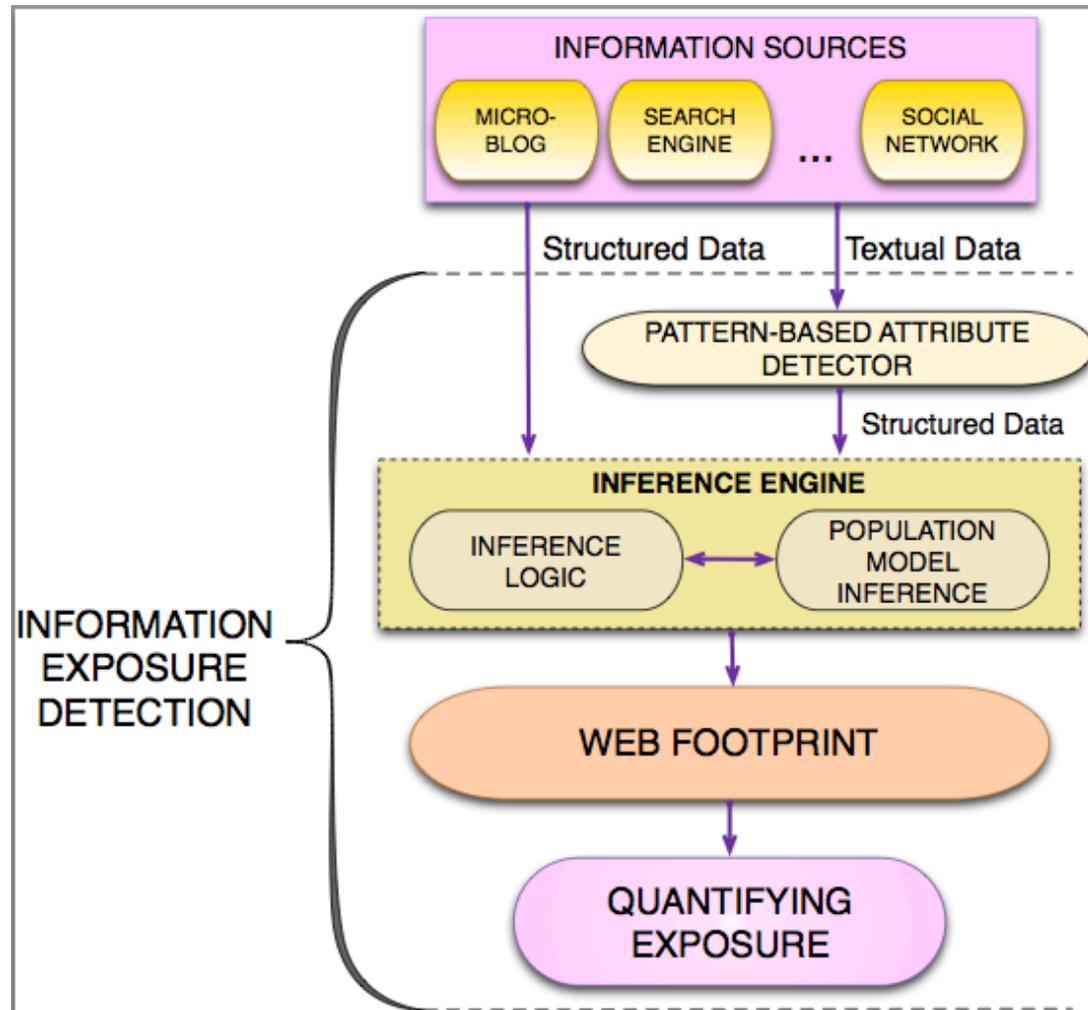
Web Footprinting

Experiments for Understanding Public Profiles

- About.me - personal website hosting site
  - Each user can make a custom webpage about themselves
  - Can list links to their social media profiles on multiple websites

- Using their API, we collected 124,497 people's information -> Ground Truth
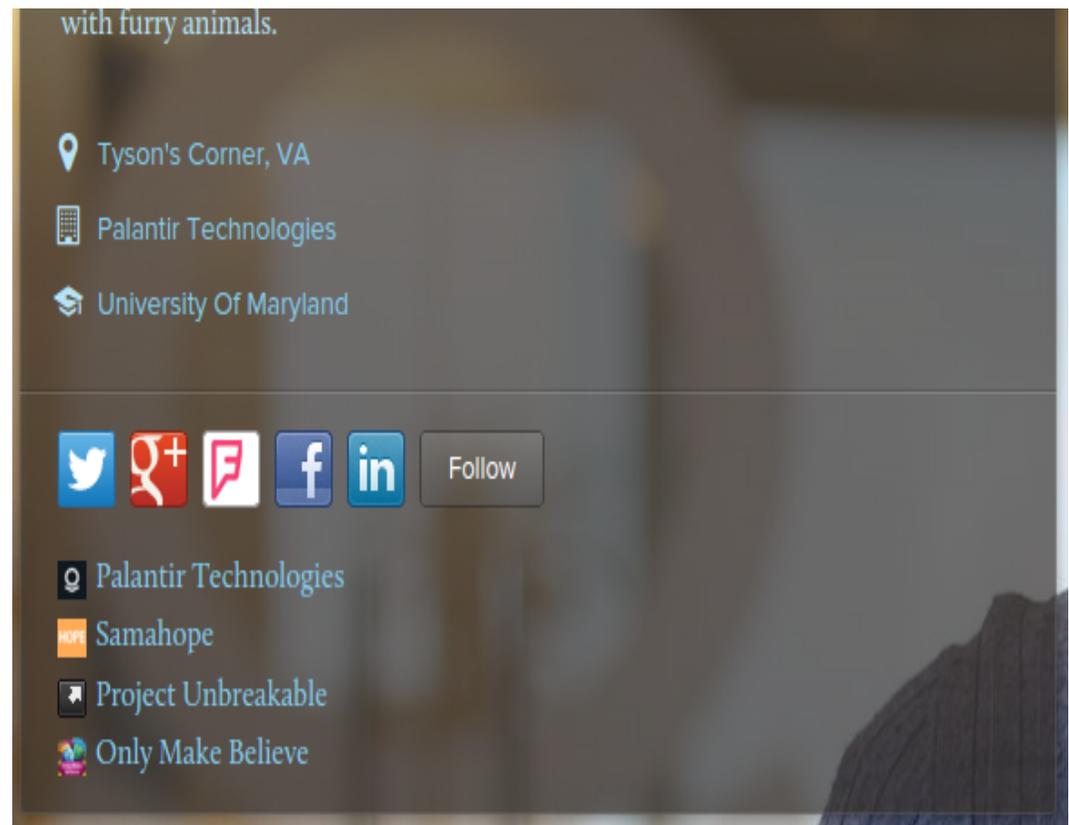
Creating Web Footprints Using Google+, Foursquare, LinkedIn Profiles

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

[Singh et al., 2015]

GEORGETOWN UNIVERSITY

Synonyms can be found

Dbpedia

Meronym

Ontology Property Name	Property Value
owl:sameAs	http://fr.dbpedia.org/resource/UniversitÃÍ_de_Georgetown
dbpedia-owl:wikiPageRedirects	dbpedia:Georgetown_University,_Washington,_D.C.
dbpprop:parent	dbpedia:Georgetown_University_Law_Center
foaf:name	Collegium Georgiopolitanum Ad Ripas Potomaci in Marylandia
dbpprop:nickname	dbpedia:Georgetown_Hoyas

Synonyms

Using an Ontology
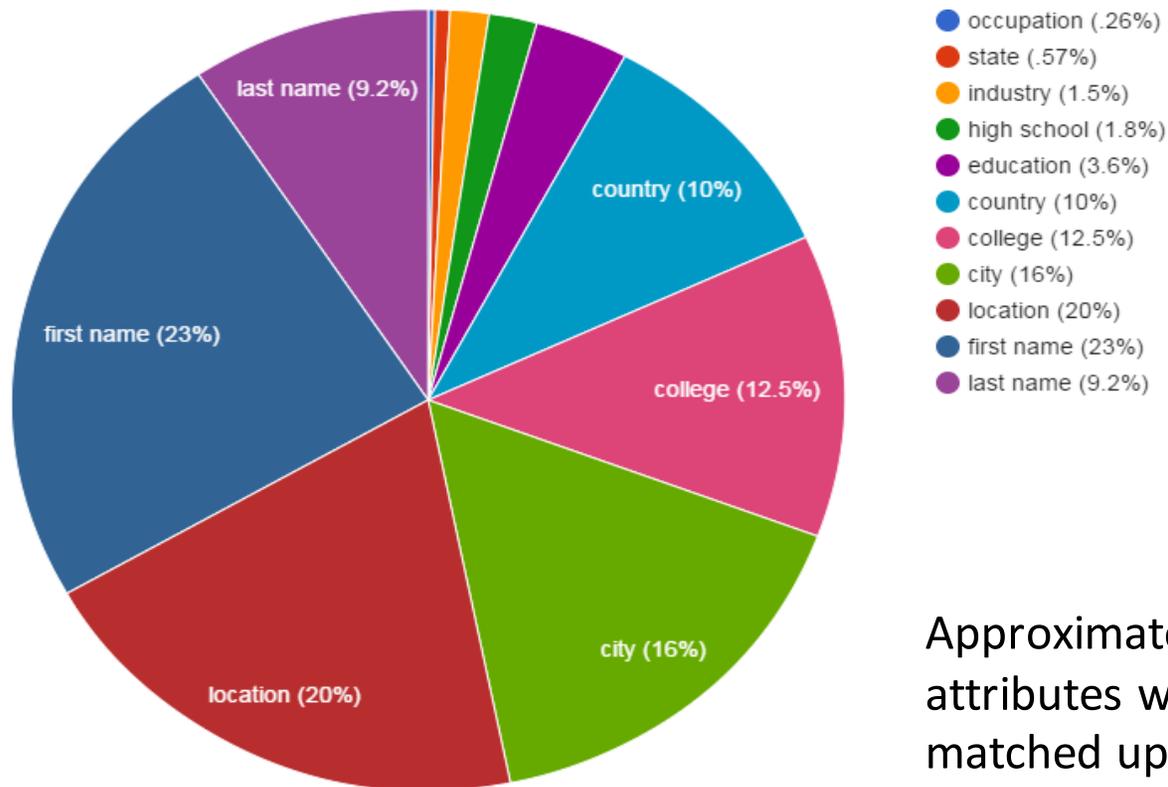
Approximately 8000
attributes were
matched up from
the ontology

Taking Control of Our Web Identity and Data

1. Keep your public profile professional.
2. Change all your social media account settings that have personal information on them from public to private.
3. Choose your friends wisely – add them selectively.
4. Join groups related to your professional interests.
5. Make it difficult for automated tools to link your accounts, e.g. use different account user names, share different information, etc.
6. Install ad blockers to reduce data about your click through habits.
7. Set your browser to not accept cookies from sites that you have not visited before.

The world around us

DATAFICATION

Data Ethics

- ## Regulation
  - We need to hold companies to higher standards.

- ## Data ethics standards
  - We need discussion, debate, and possibly a new discipline.

- ## Catalog of personal data
  - Individuals should be able to see, correct and/or remove data companies have about them.

[Singh, 2016]

GEORGETOWN
UNIVERSITY

Final Thoughts

- There is a cultural acceptance of sharing private data publicly.
- This is a problem - I have shown you different techniques for generating web footprints. It is too easy!!
- We need new ways to help users understand what data can be determined about them and help them take control of their information.
- We need to pause and debate online privacy and ethical uses of large-scale human behavioral data.
- We need to develop guidelines and regulations that protect users.

We need to take back control of our data.

References

J. Zhu, S. Zhang, L. Singh, H. Yang, and M. Sherr. "Generating Risk Reduction Recommendations to Decrease Identifiability of Public Online Profiles." under submission.

A. Hian-Cheong, L. Singh. M. Sherr, H. Yang. "Semantics and Public Information Exposure Detection." invited.

L. Singh, H. Yang, M. Sherr, A. Hian-Cheong, K. Tian, J. Zhu, and S. Zhang. "Public Information Exposure Detection: Helping Users Understand Their Web Footprints." *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. Paris, France: EEE/ACM , 2015.

L. Singh, H. Yang, M. Sherr, Y. Wei, A. Hian-Cheong, K. Tian, J. Zhu, S. Zhang, T. Vaidya, and E. Asgarli. Helping Users Understand Their Web Footprints. *International Conference on World Wide Web - Companion Proceedings*. *World Wide Web (WWW),* Florence, Italy. Poster Paper, 2015 .

W. B. Moore, Y. Wei, A. Orshefsky, M. Sherr, L. Singh, H. Yang. "Understanding Site-Based Inference Potential for Identifying Hidden Attributes." *International Conference on Privacy, Security, Risk and Tr*ust. Alexandria, VA: IEEE Computer Society, 2013.

J. Ferro, L. Singh, M. Sherr. "Identifying individual vulnerability based on public data." *International Conference on Privacy, Security and Trust*. Tarragona, Catalonia, Spain: IEEE Computer Society, 2013.

F. Nagle, L. Singh, and A. Gkoulalas-Divanis. "EWNI: Efficient Anonymization of Vulnerable Individuals in Social Networks." *Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD*). Kuala Lumpur, Malaysia: Springer, 2012.

A. Ramachandran, L. Singh, E. Porter, and F. Nagle. "Exploring re-identification risks in public domains." *Conference on Privacy, Security and Trust (PST)*. IEEE Computer Society, 2012.

GEORGETOWN
UNIVERSITY

The Team & Support

- Faculty:
  - Lisa Singh, Micah Sherr, Grace Hui Yan

- Students & Researchers:
  - Rob Churchill, Kristen Skillman, Kevin Tian, Sicong Zhang, Yanan Zhu

- Alumni:
  - Aditi Ramachandran, Frank Nagle, John Ferro, Yifang Wei, Brad Moore, Andrew Hian-Cheong, Janet Zhu

Support: National Science Foundation

5 Reasons to Join Our Program!

1. We are research active and provide full financial RA support for PhD students for 5 year.
2. We have full and partial scholarships for Master's students.
3. Our courses span applied and theoretical areas of computer science, as well as interdisciplinary areas like data science.
4. We have exceptional job placement in top tech firms, national labs, and government agencies.
5. We have a strong community among students and faculty.

Need more information:
Website: http://cs.georgetown.edu/
Graduate Director: Lisa Singh
(singh@cs.georgetown.edu)

Application deadline April 1!

GEORGETOWN UNIVERSITY