Information Sharing with Privacy by Design

Case Study: Voter Registration Modernization

National Academy of Science
State and Local Governments Use of Alternative and Multiple Data Sources

Jeff Jonas, IBM Fellow
Chief Scientist, Context Computing
http://www.twitter.com/jeffjonas
www.jeffjonas.typepad.com

Jeff Jonas

IBM Fellow
Chief Scientist, Context Computing

- Founded SRD in early 1980's
- Architected, designed, developed 100+ systems over three decades
- Funded by In-Q-Tel in 2001 & 2003
- Acquired by IBM in 2005
- Selected affiliations:
  - Board Member of EPIC and USGIF
  - Advisory Board Member of EFF and Privacy International
  - Senior Associate at CSIS, Transnational Threats Group
  - Adjunct at Singapore Management University, School of Information Systems
- Current focus: Sensemaking Systems w/ Privacy by Design

INTRODUCING ENTITY RESOLUTION

Who is Fang Wong?

@FangWong
2.5M Followers

Fang Wong
Top 100 Customer

FangWong@Email.com
Newsletter Subscriber

F A Wong
Seattle, DOB: 6/12/82
Former Customer

Fang Wong
FangWong@Email.com
Marketing Department's
Prospect List

Resolving the Fang Wong

@FangWong
2.5M Followers

Fang Wong
Top 100 Customer

FangWong@Email.com
Newsletter Subscriber

F A Wong
Seattle, DOB: 6/12/82
Former Customer

Fang Wong
FangWong@Email.com
Marketing Department's
Prospect List

Resolving the Fang Wong

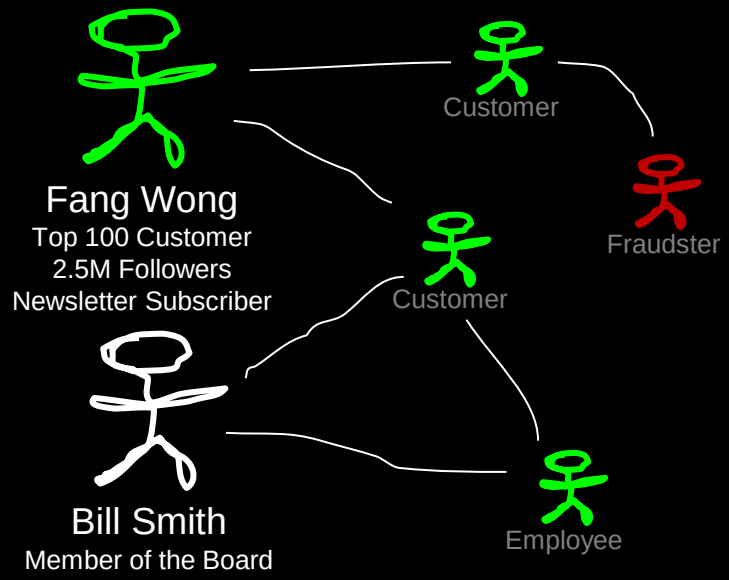
Fang Wong

Top 100 Customer

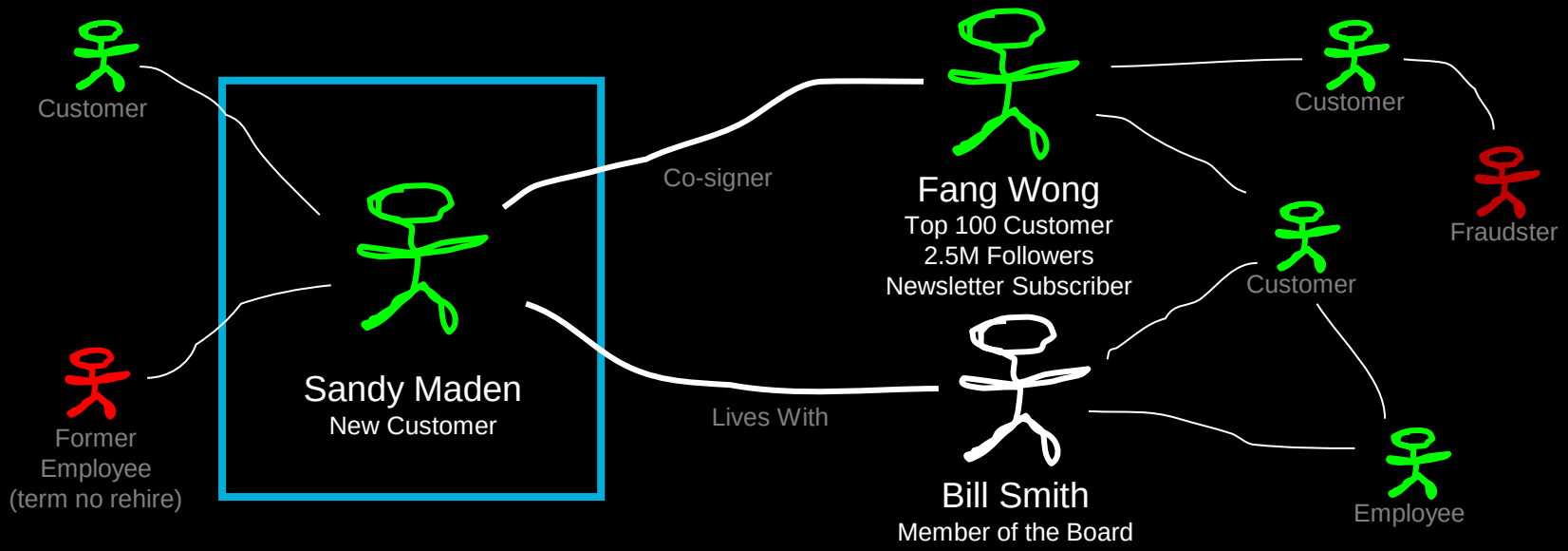
2.5M Followers

Newsletter Subscriber

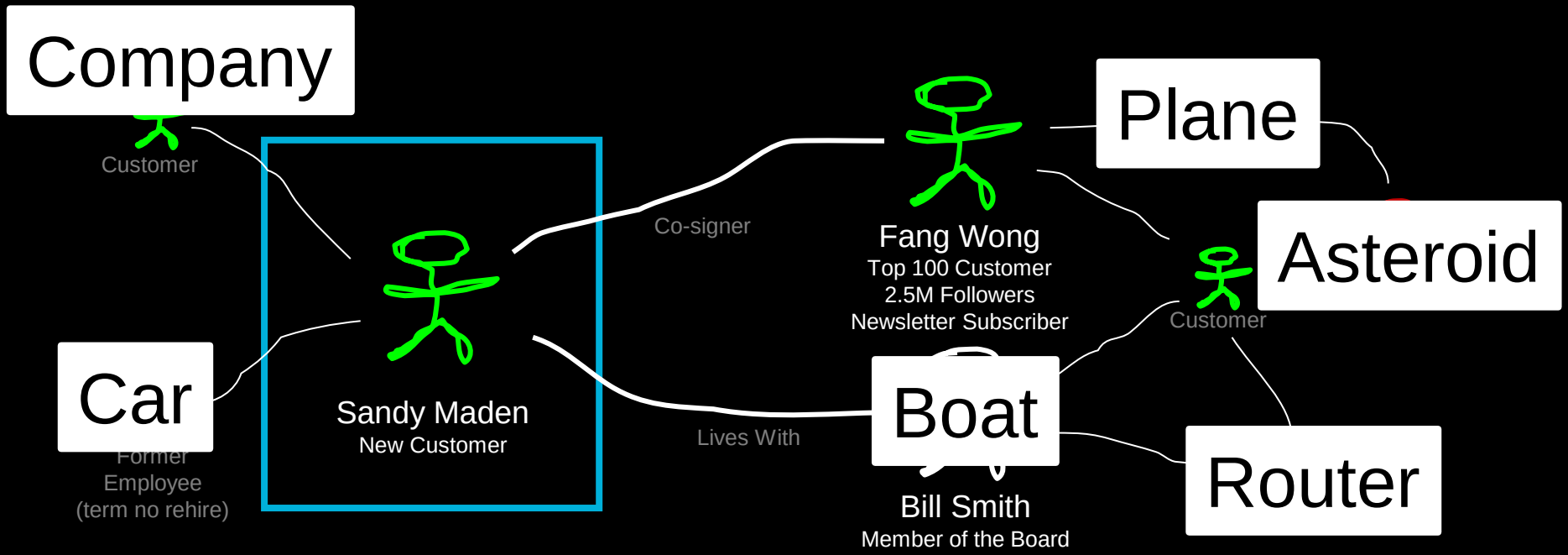
Graphing the (resolved) Fang Wong



Sandy Maden in Context



“Entities”


Company
Customer
Plane
Co-signer
Fang Wong
Top 100 Customer
2.5M Followers
Newsletter Subscriber
Asteroid
Customer
Car
Former
Employee
(term no rehire)
Sandy Maden
New Customer
Lives With
Boat
Bill Smith
Member of the Board
Router

Use Cases

- Insider Threat Detection for Las Vegas
- Fraud Detection at MoneyGram
- Post Katrina re-unifications for Louisiana Governor’s Office
- Maritime Domain Awareness for Singapore Ministry of Defense
- Foreign Corrupt Practices Act (FCPA) Risk Assessment in IBM

My Basic Privacy by Design (PbD) Principles

1. Full Attribution (mandatory)
2. Data Tethering
3. **Selective Anonymization**
4. Tamper Resistant Audit Log
5. False Negative Favoring (mandatory)
6. Self-Correcting False Positives (mandatory)
7. Information Transfer Accounting

VOTER REGISTRATION MODERNIZATION USING “SELECTIVE ANONYMIZATION”

Is this Voter Deceased?

Because many people share a name and year of birth, this is a “maybe.”

With thousands of “maybes” there are not enough humans to review them all.

VOTER	Voted 2012
Bill F Balston Born: 1951 D/L: 4801 13070 Karen St Apt #7	

DECEASED	Died 2015
William Balston Born: 1951 SSN: 5598	

Tertiary Data Assists

VOTER	Voted 2012
Bill F Balston Born: 1951 D/L: 4801 13070 Karen St Apt #7	

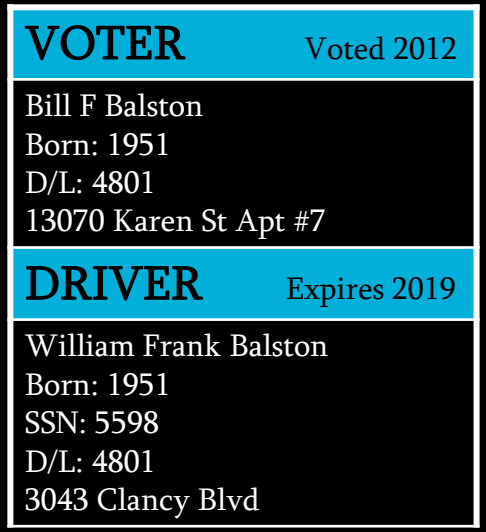
DECEASED	Died 2015
William Balston Born: 1951 SSN: 5598	

Fortunately, this DMV record comes along and resolves to the Voter.

DRIVER	Expires 2019
William Frank Balston Born: 1951 SSN: 5598 D/L: 4801 3043 Clancy Blvd	

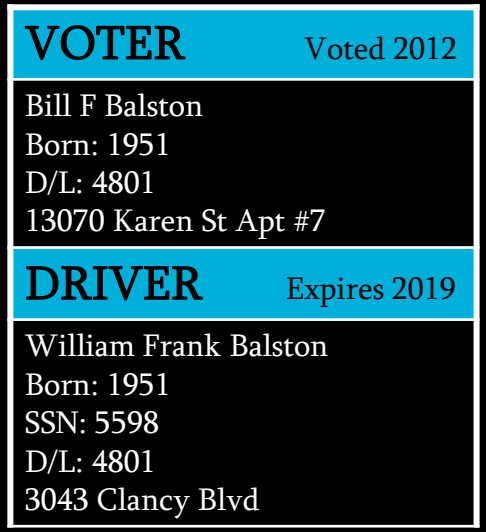
Tertiary Data Assists

Fortunately, this DMV record comes along and resolves to the Voter.



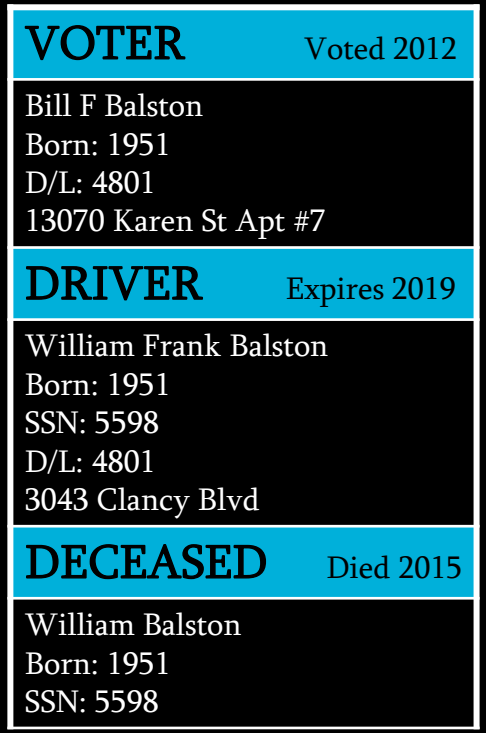
Tertiary Data Assists

The combined record has ‘learned’ an SSN, making it possible to assert this is the deceased person.



Tertiary Data Assists

The combined record has ‘learned’ an SSN, making it possible to assert this is the deceased person.



Insight … Revealed

VOTER	Voted 2012
Bill F Balston Born: 1951 D/L: 4801 13070 Karen St Apt #7	
DRIVER	Expires 2019
William Frank Balston Born: 1951 SSN: 5598 D/L: 4801 3043 Clancy Blvd	
DECEASED	Died 2015
William Balston Born: 1951 SSN: 5598	

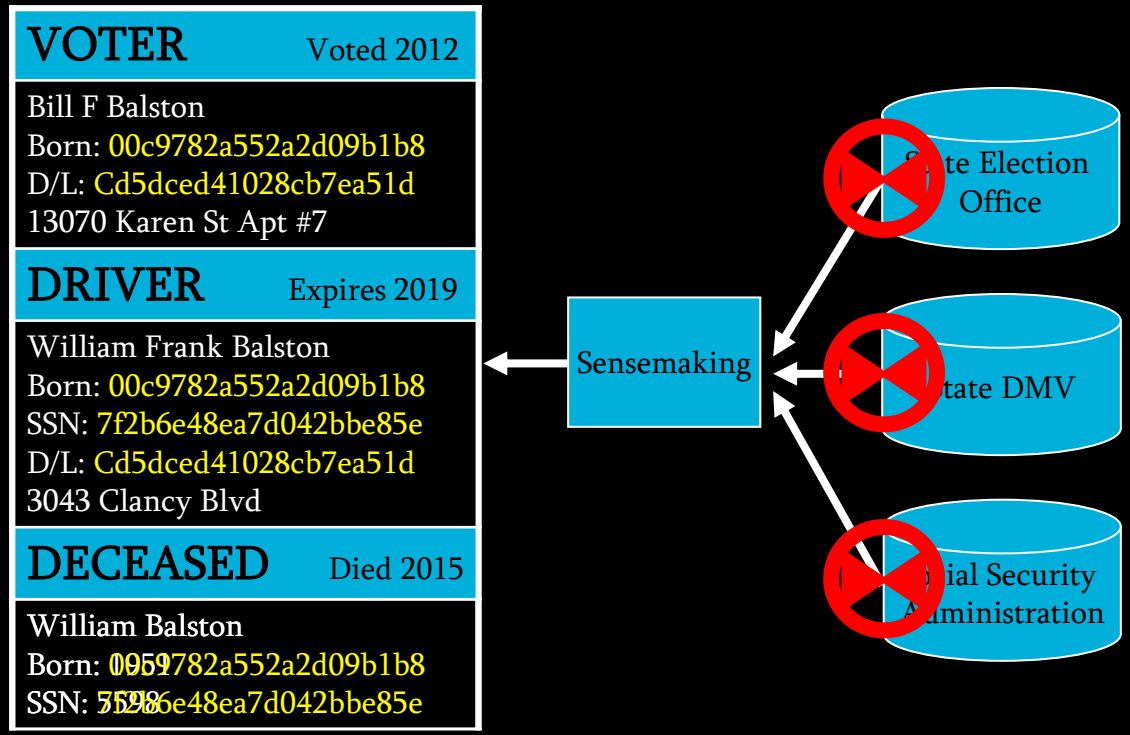
This voter is not expected to vote in 2016!

Information Sharing with Privacy by Design (PbD)

Using “**Selective Anonymization**” to reduce the risk of unintended disclosure.

This is not de-identification.

Every entity is re-identifiable *i.e.*, auditable.



Benefits of Anonymous Resolution

- Unlike other anonymization techniques, identities can be resolved after anonymization (versus k-anonymity techniques)
- Guaranteed source attribution – resulting in a fully auditable and reconcilable system
- Domain-specific implementations virtually eliminates the opportunities for re-purposing the anonymized database
- Where there is sharing or there are pressures to share sensitive identity data, presents an important alternative to “wide open” sharing
- Enhanced protections against unintended disclosure (*e.g.,* insider threat resistant)

Vulnerabilities: Various Crypto Attacks

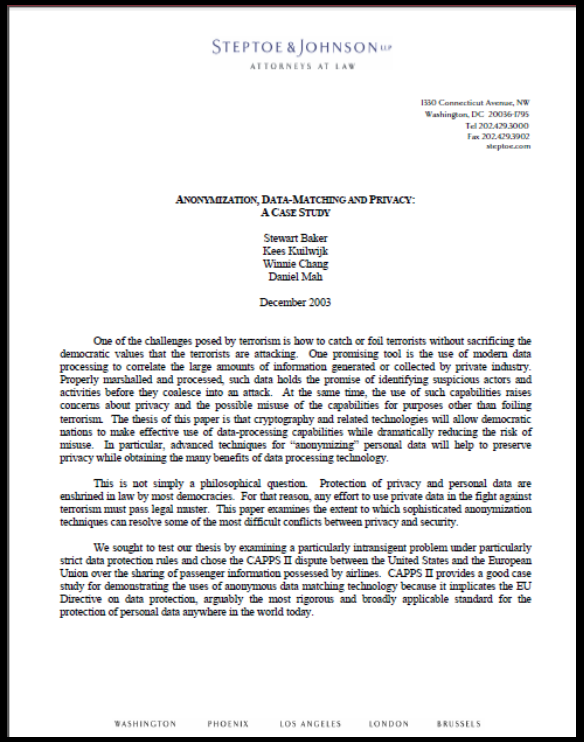
- Dictionary attacks against the whole anonymized database
- Chosen text attacks carried out by users with Anonymizers
- Statistical or traffic analysis attacks
- Others

Different Missions Necessitate Different Measures

- Information sharing with oneself
- Information sharing with similar organizations (*e.g.,* private-private or public-public)
- Information sharing across organization types (*e.g.,* private-public)
- Information sharing across friendly governments
- Information sharing across other entities with high levels of bilateral distrust

Legal Analysis and Opinion Papers

As related to the EU Data Protection Act



STEPTOE & JOHNSON LLP
ATTORNEYS AT LAW

1330 Connecticut Avenue, NW
Washington, DC 20036-1795
Tel 202.429.3000
Fax 202.429.3902
steptoe.com

ANONYMIZATION, DATA-MATCHING AND PRIVACY:
A CASE STUDY

Stewart Baker
Kees Kuilwijk
Winnie Chang
Daniel Mah

December 2003

One of the challenges posed by terrorism is how to catch or foil terrorists without sacrificing the democratic values that the terrorists are attacking. One promising tool is the use of modern data processing to correlate the large amounts of information generated or collected by private industry. Properly marshalled and processed, such data holds the promise of identifying suspicious actors and activities before they coalesce into an attack. At the same time, the use of such capabilities raises concerns about privacy and the possible misuse of the capabilities for purposes other than foiling terrorism. The thesis of this paper is that cryptography and related technologies will allow democratic nations to make effective use of data-processing capabilities while dramatically reducing the risk of misuse. In particular, advanced techniques for "anonymizing" personal data will help to preserve privacy while obtaining the many benefits of data processing technology.

This is not simply a philosophical question. Protection of privacy and personal data are enshrined in law by most democracies. For that reason, any effort to use private data in the fight against terrorism must pass legal muster. This paper examines the extent to which sophisticated anonymization techniques can resolve some of the most difficult conflicts between privacy and security.

We sought to test our thesis by examining a particularly intransigent problem under particularly strict data protection rules and chose the CAPPS II dispute between the United States and the European Union over the sharing of passenger information possessed by airlines. CAPPS II provides a good case study for demonstrating the uses of anonymous data matching technology because it implicates the EU Directive on data protection, arguably the most rigorous and broadly applicable standard for the protection of personal data anywhere in the world today.

WASHINGTON PHOENIX LOS ANGELES LONDON BRUSSELS

http://www.steptoe.com/publications/279d.pdf

As related to US HIPAA law

IBM

Research Report:
Application of IBM Anonymous Resolution to the Health Care Sector

February, 2006

Peter P. Swire
C. William O'Neill Professor of Law
The Ohio State University

ON DEMAND BUSINESS

Other Reference Material

Blog

- To Anonymize or Not Anonymize, That is the Question

Video

- Modernizing Voter Registration in America

Papers

- Privacy by Design in the Era of Big Data
- Anonymous linking project - final report

jeffjonas@us.ibm.com

www.jeffjonas.typepad.com

http://www.twitter.com/jeffjonas

Information Sharing with Privacy by Design

Case Study: Voter Registration Modernization

National Academy of Science
State and Local Governments Use of Alternative and Multiple Data Sources

Jeff Jonas, IBM Fellow
Chief Scientist, Context Computing
http://www.twitter.com/jeffjonas
www.jeffjonas.typepad.com