Digital Footprints: What can be learned from the traces we leave on social networks

April Galyardt

Carson Sestili

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

[Distribution Statement A] Approved for public release and unlimited distribution.

Copyright 2018 Carnegie Mellon University. All Rights Reserved.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

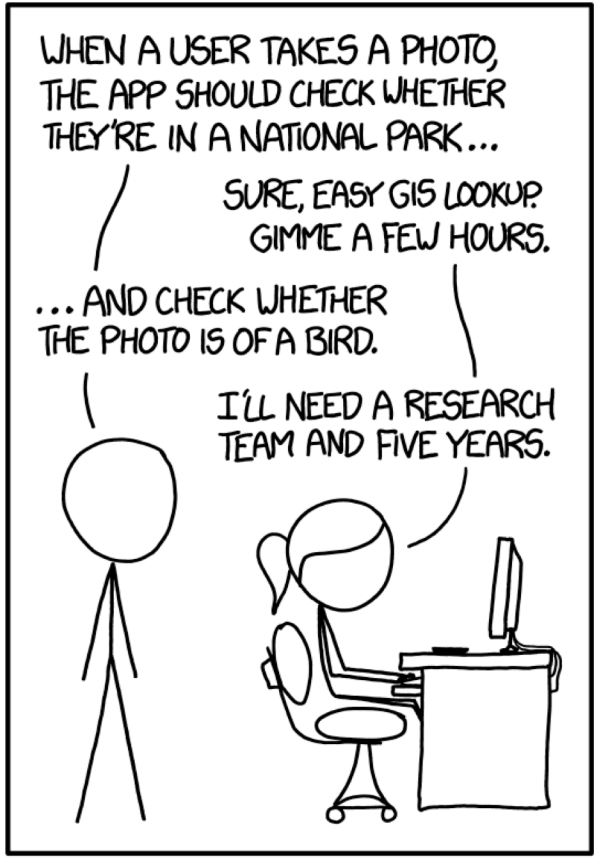
NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

Carnegie Mellon® and CERT® are registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM18-0711



IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

Today, we're gonna try.

https://xkcd.com/1425/
(CC BY-NC 2.5)

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

What is Metadata?

Metadata

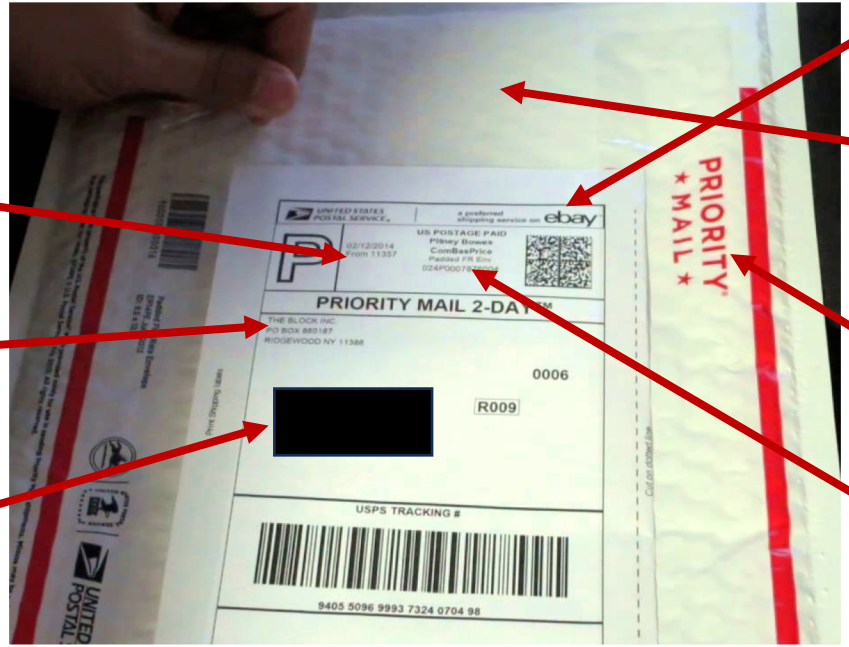
Content

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

What is Metadata?

Metadata



© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

“We don’t see any content.”

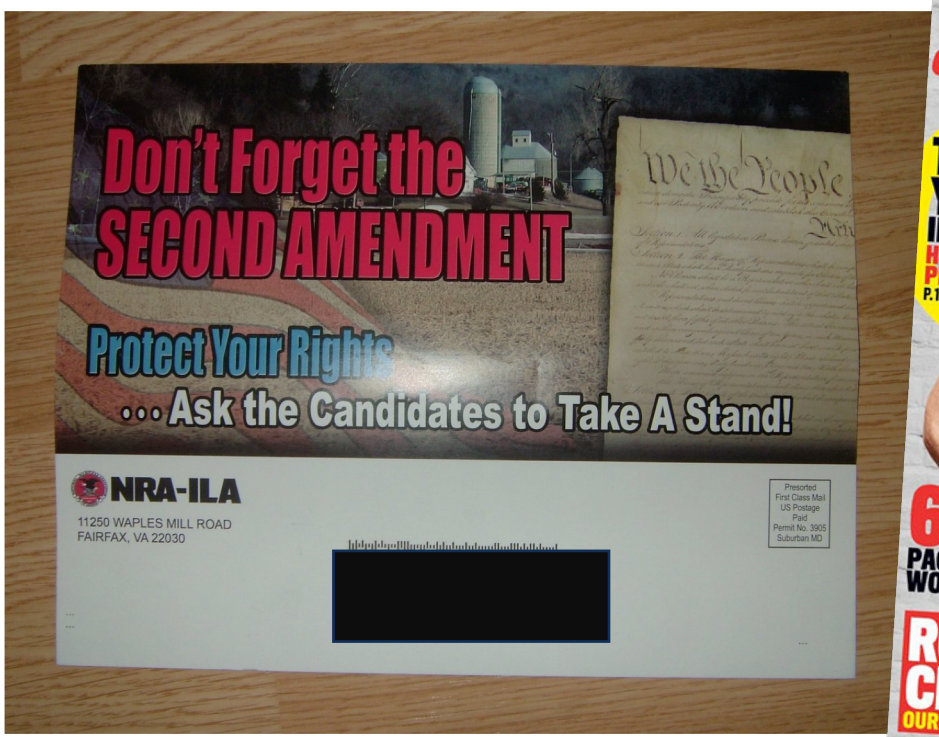
© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Well,… what might your postman infer?

© 2018 Carnegie Mellon University
[Distribution Statement A] Approved for public release and unlimited distribution.

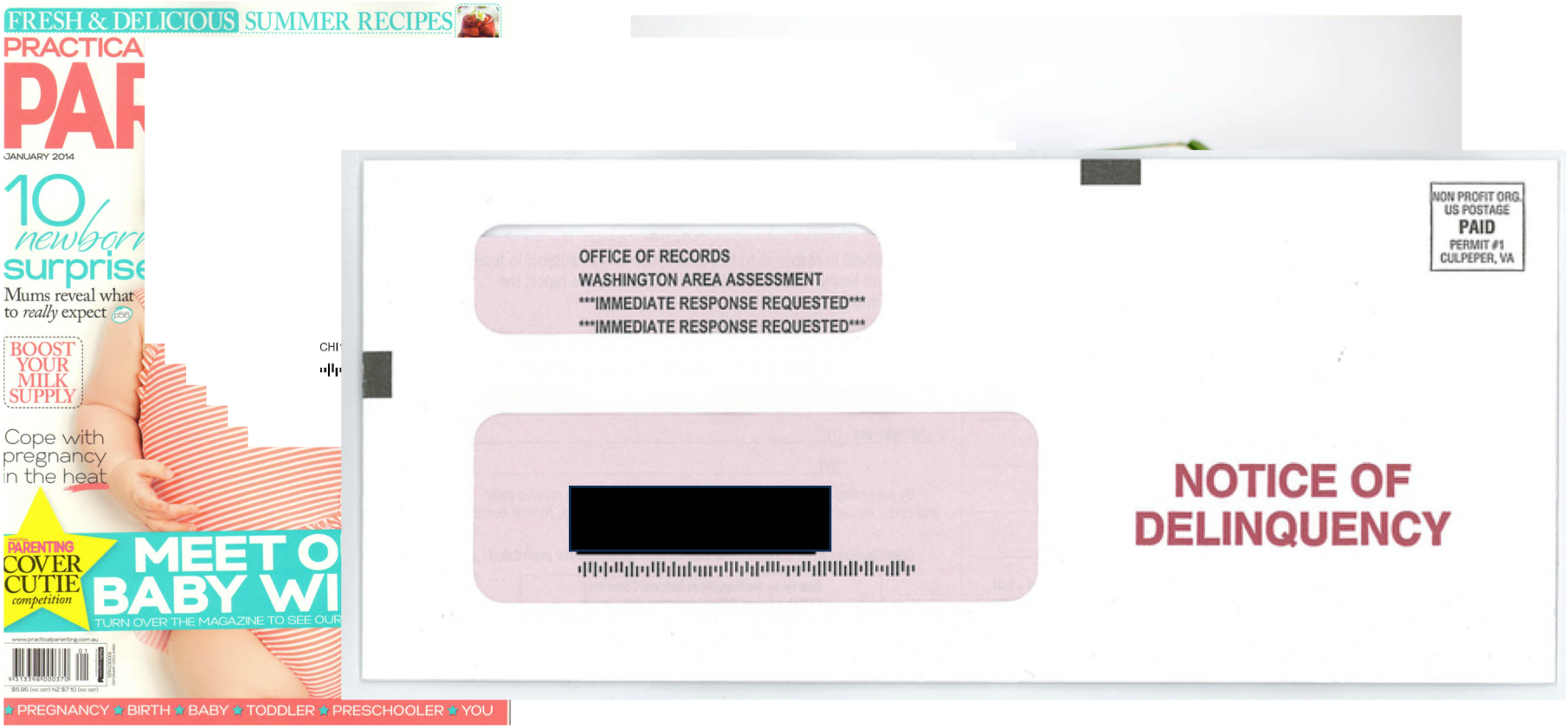
What about this person?



© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

And these residents?



© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

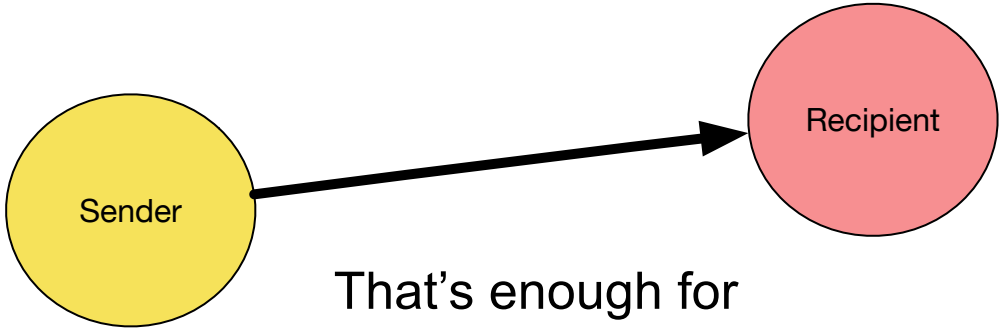
Metadata is Data

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Just 3 pieces of information

- Sender
- Recipient
- Date

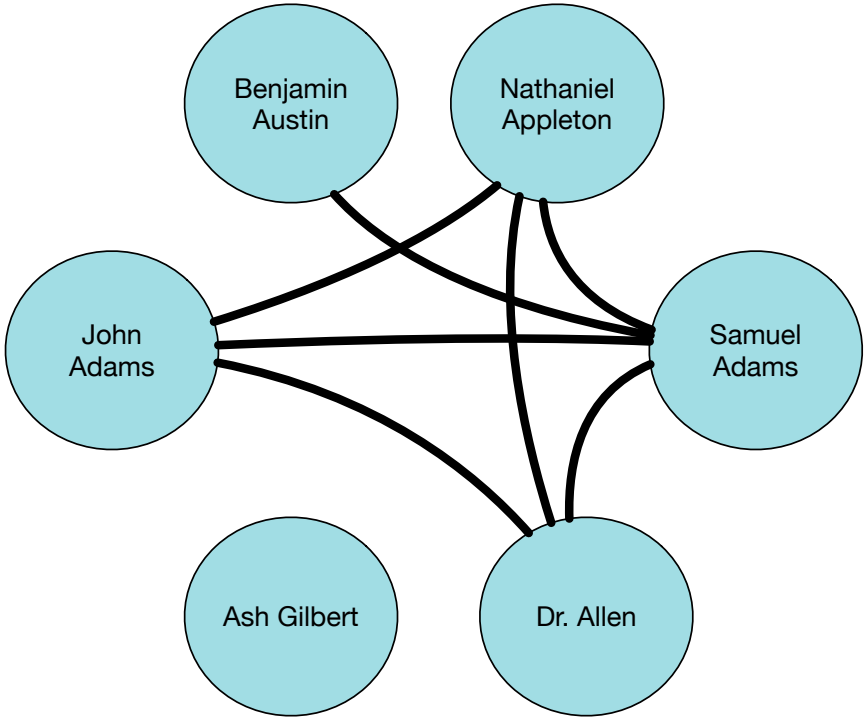


That's enough for social network analysis.

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Boston, 1772 – A piece of the picture.



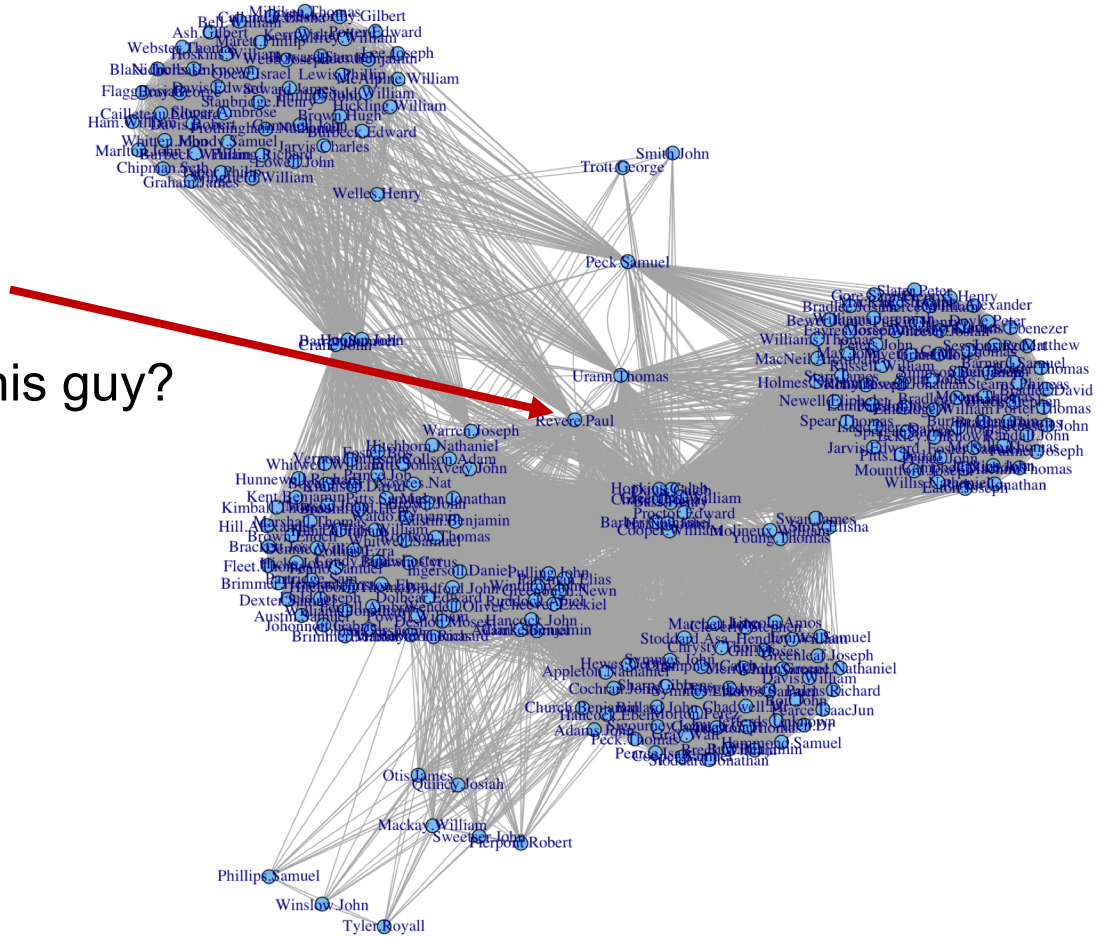
This example is due to Kieran Healy:
https://kieranhealy.org/blog/archives/2013/06/09/using-metadata-to-find-paul-revere/

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Boston, 1772

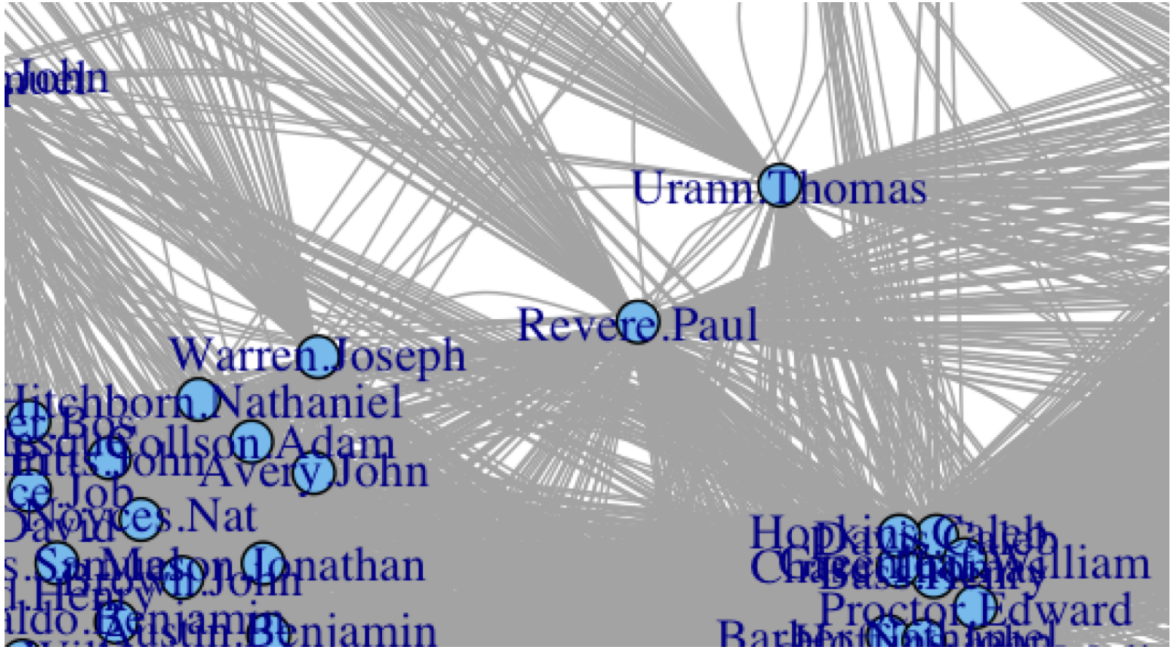
Wait, Who's this guy?



© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Boston, 1772

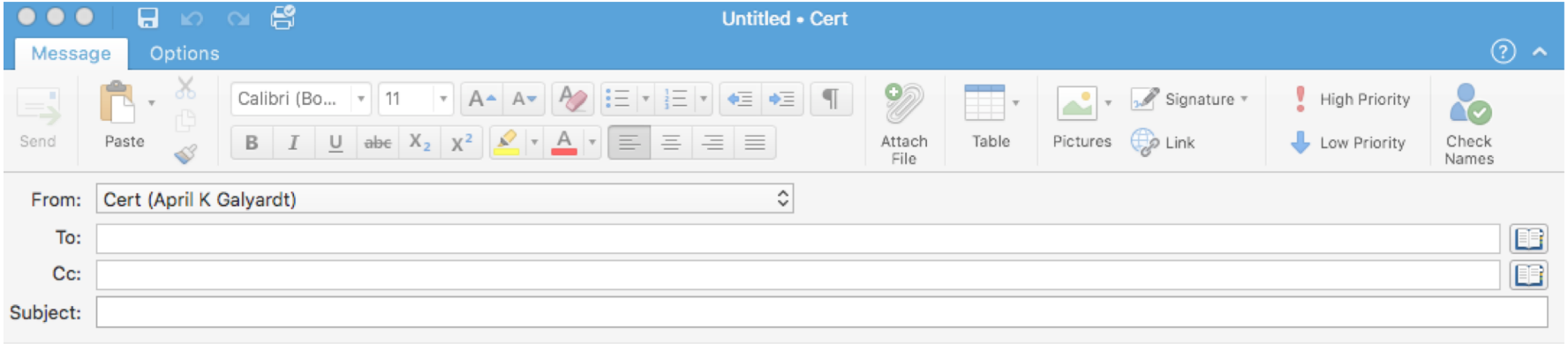


With 3 pieces of information and minimal calculations, we've identified a figure central in the "uprising" of 1776.

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Explore your own Metadata



Data Used:

From

To

CC

Timestamp

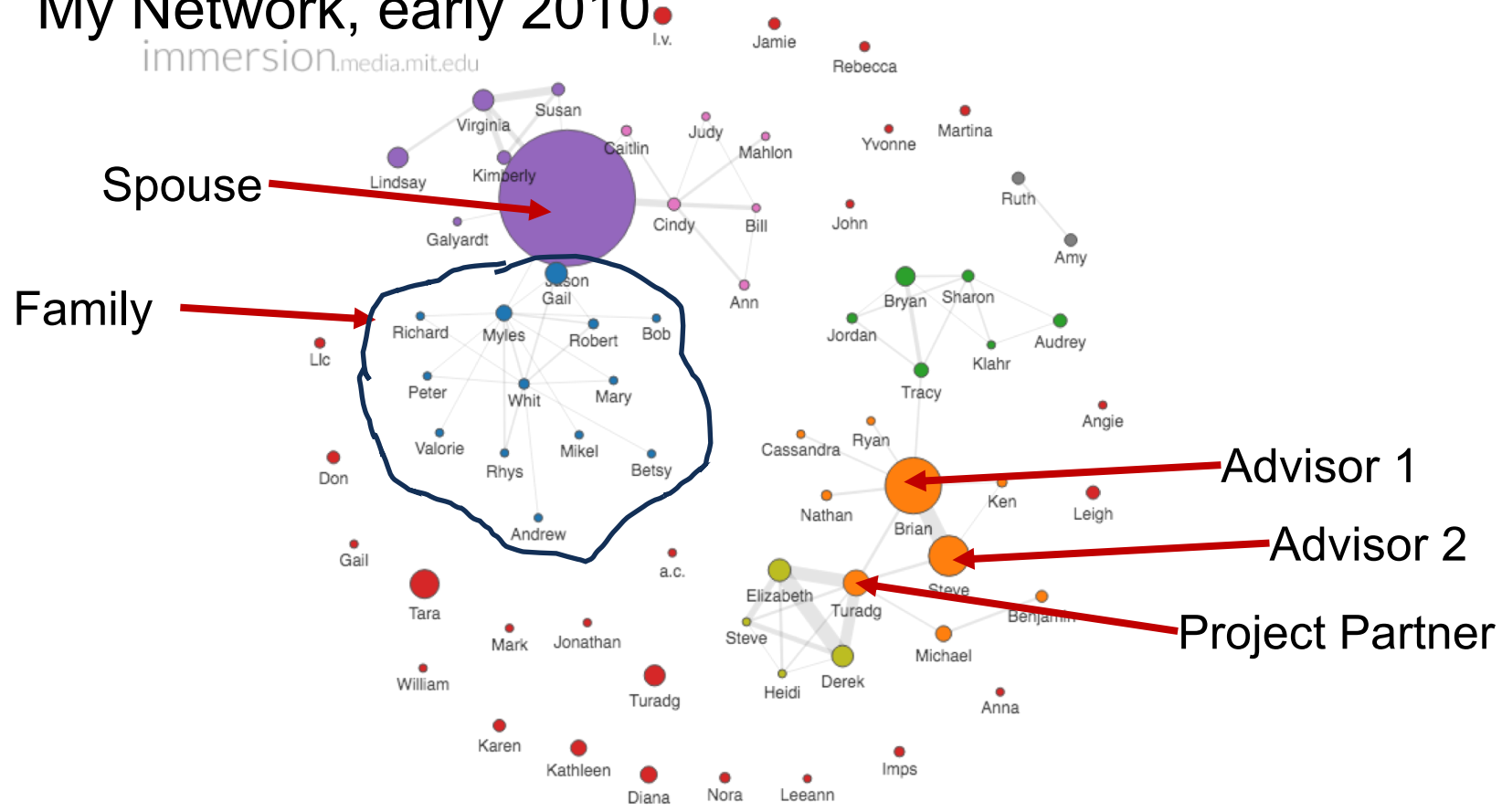
https://immersion.media.mit.edu/

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

My Network, early 2010

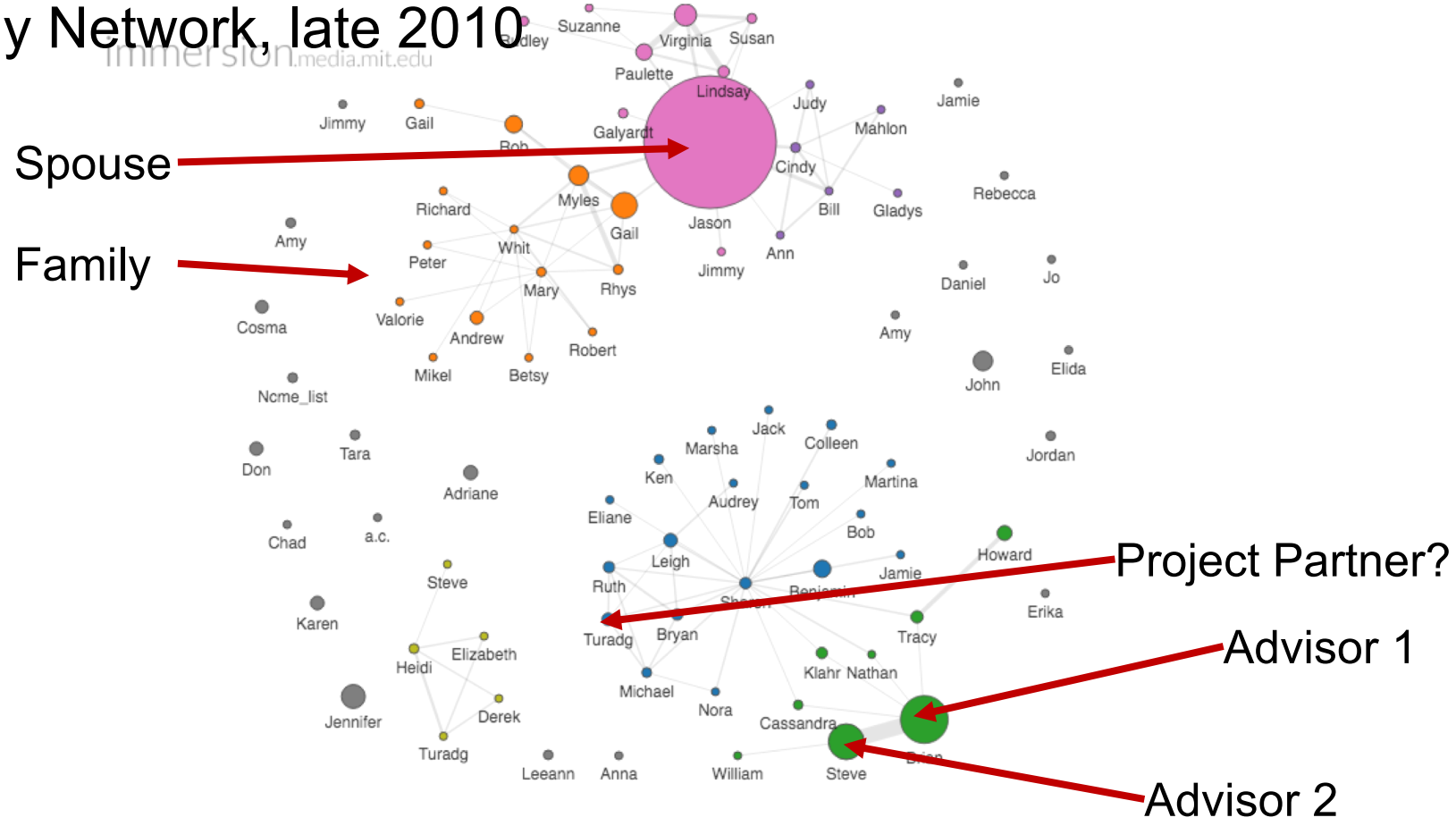
immersion.media.mit.edu



© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

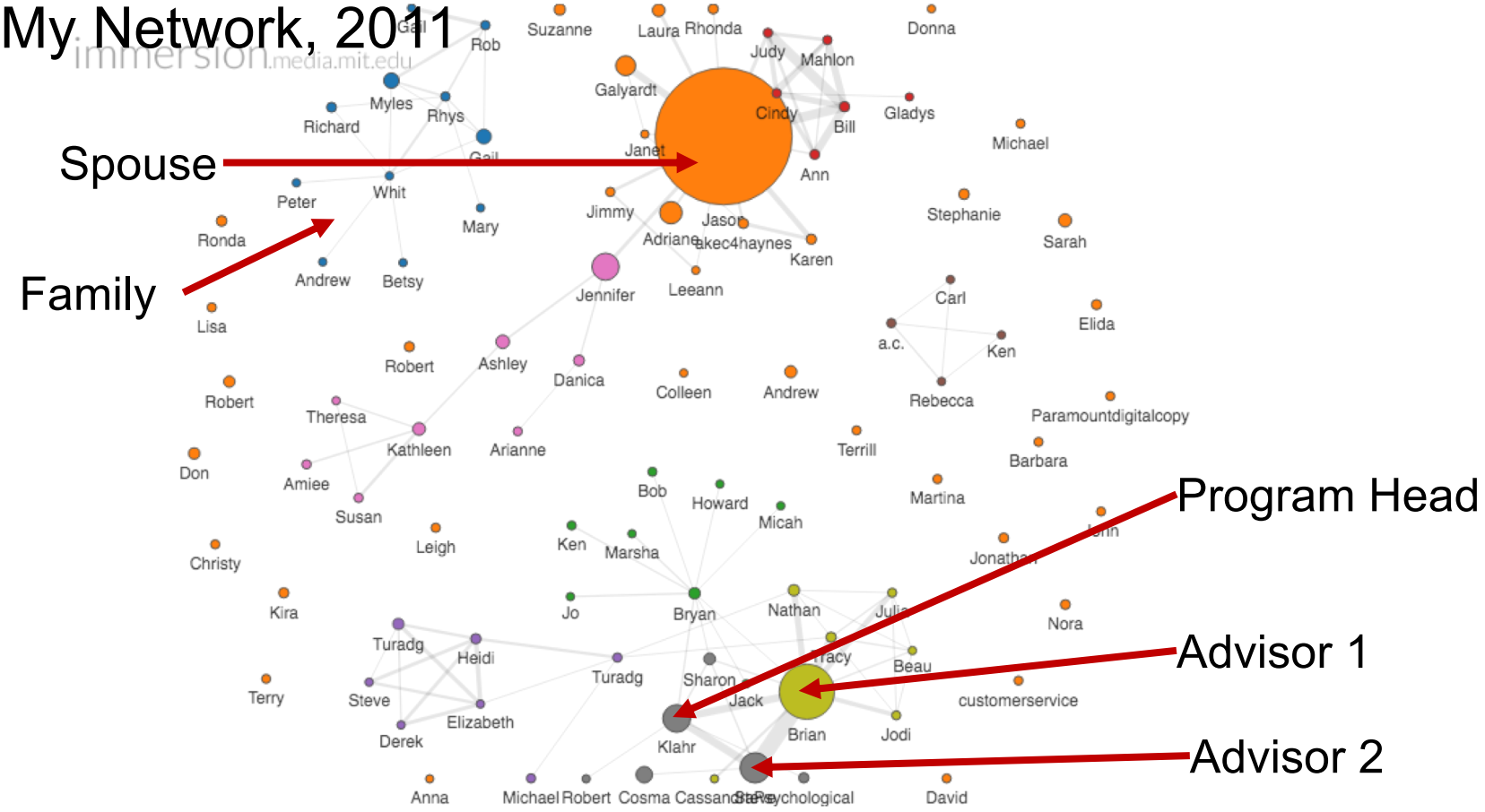
My Network, late 2010



© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

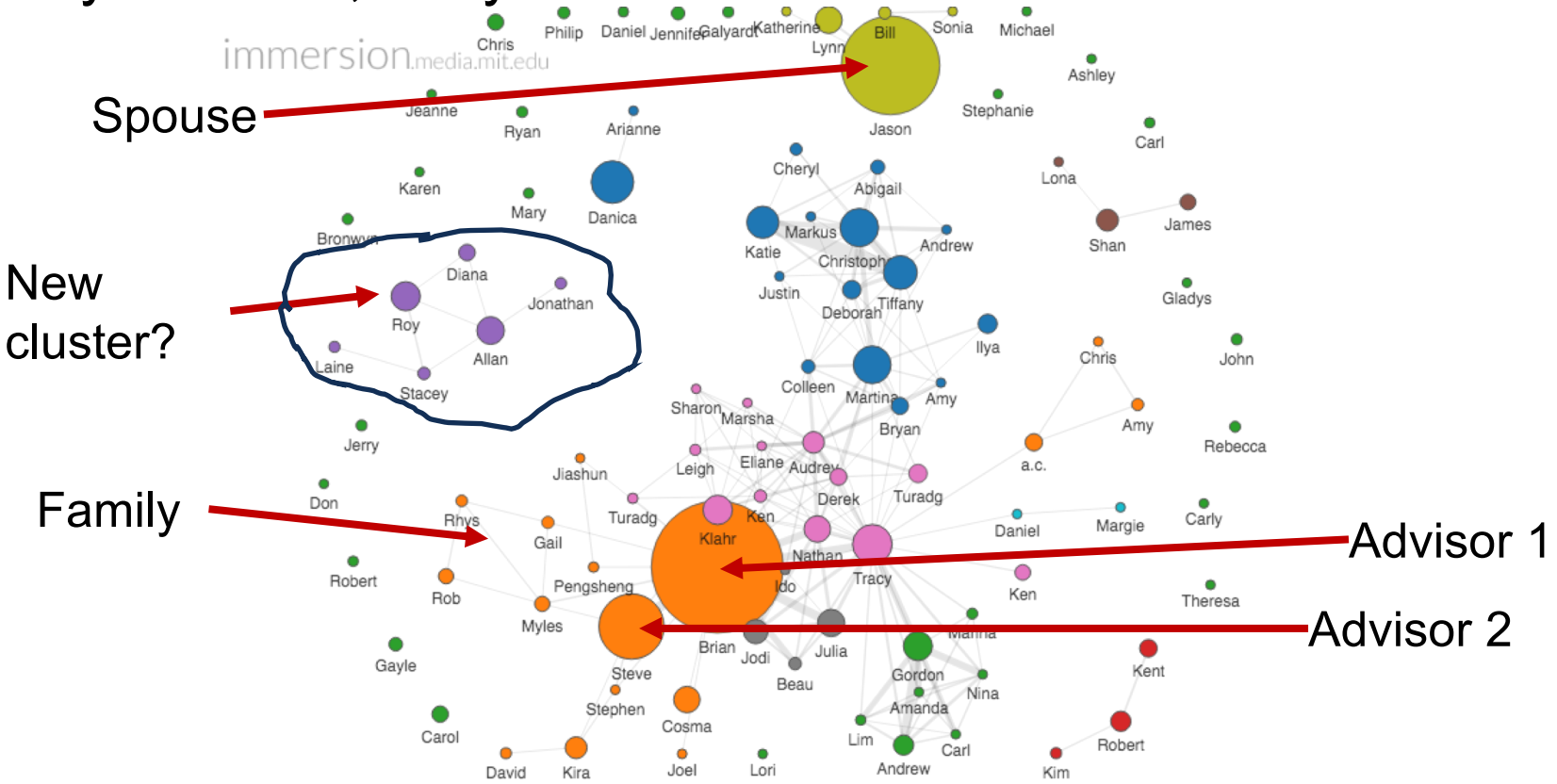
My Network, 2011



© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

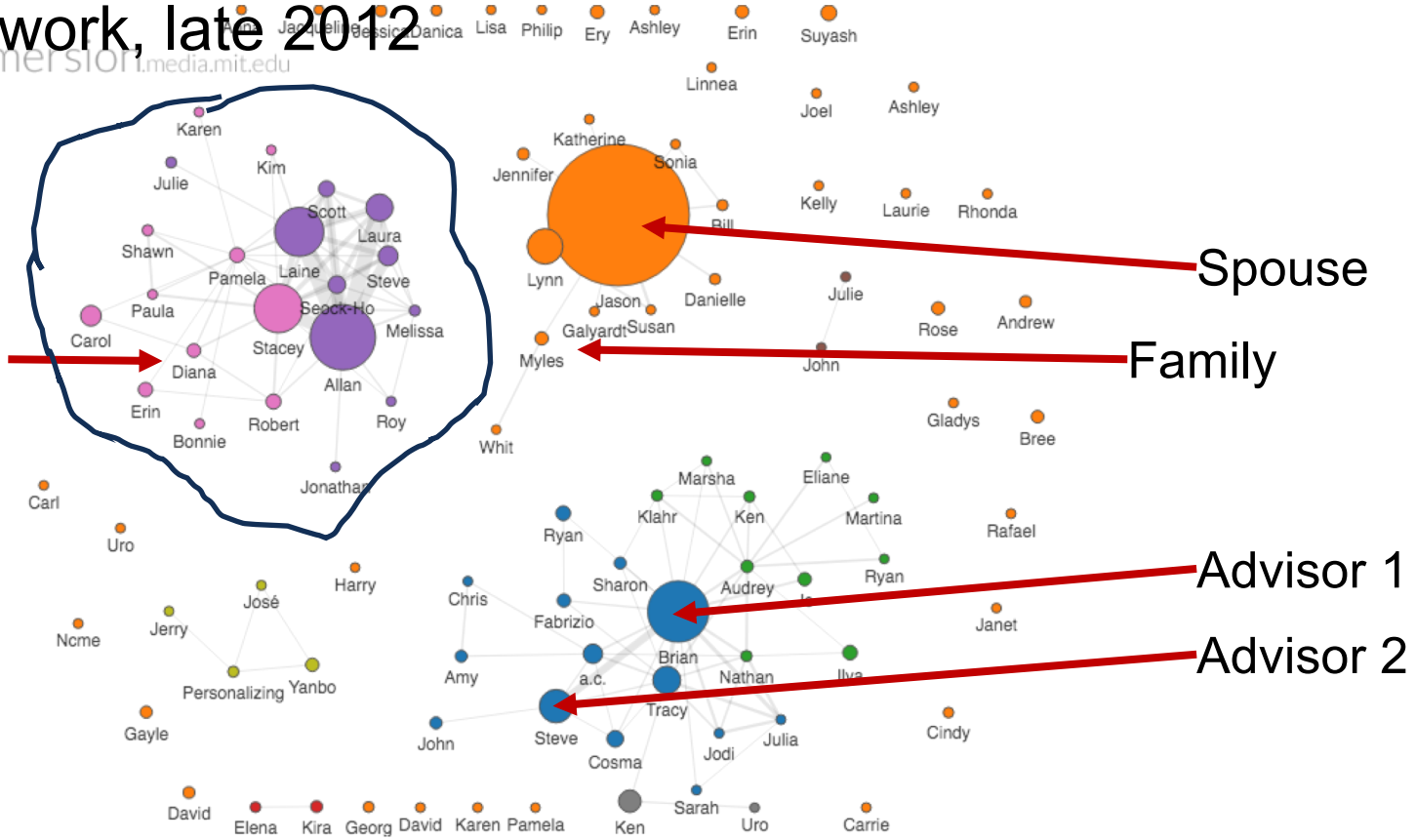
My Network, early 2012



© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

My Network, late 2012



Woah, something changed!

Spouse

Family

Advisor 1

Advisor 2

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

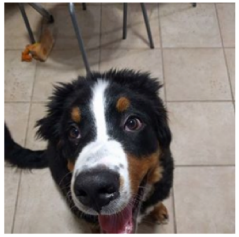
What if we only see you, and your likes?

April Galyardt reacted to [redacted]

Student Punished After Putting High School Up for Sale on Craigslist
gizmodo.com

April Galyardt likes [redacted]

April Galyardt likes [redacted]



This boy is growing like a weed! He accomplished his first tail table clear this weekend, and it was not a coffee table. No, he knocked things off the dining room table. His legs have suddenly caught up with the rest of his body.

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

October 2012

PNAS PNAS

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski[a,1], David Stillwell[a], and Thore Graepel[b]

[a]Free School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and [b]Microsoft Research, Cambridge CB1 2FB, United Kingdom

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The analysis presented is based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests. The proposed model uses dimensionality reduction for preprocessing the Likes data, which are then entered into logistic/linear regression to predict individual psychodemographic profiles from Likes. The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrat and

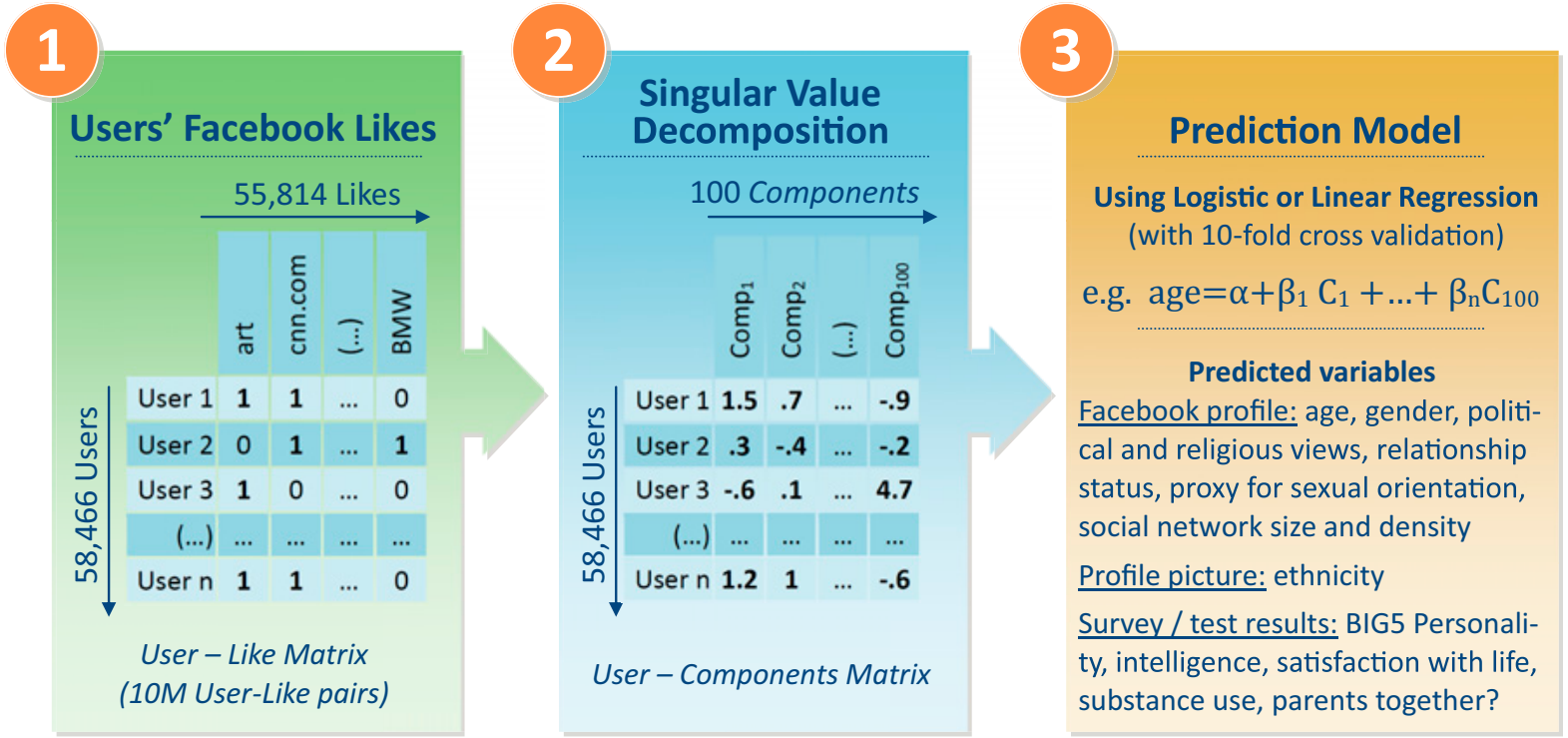
browsing logs (11–15). Similarly, it has been shown that personality can be predicted based on the contents of personal Web sites (16), music collections (17), properties of Facebook or Twitter profiles such as the number of friends or the density of friendship networks (18–21), or language used by their users (22). Furthermore, location within a friendship network at Facebook was shown to be predictive of sexual orientation (23).

This study demonstrates the degree to which relatively basic digital records of human behavior can be used to automatically and accurately estimate a wide range of personal attributes that people would typically assume to be private. The study is based on Facebook Likes, a mechanism used by Facebook users to express their positive association with (or "Like") online content, such as photos, friends' status updates, Facebook pages of prod-

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Kosinski, Stillwell, & Graepel (2012, PNAS)



© 2018 Carnegie Mellon University
[Distribution Statement A] Approved for public release and unlimited distribution.

Kosinski, Stillwell, & Graepel (2012, PNAS)

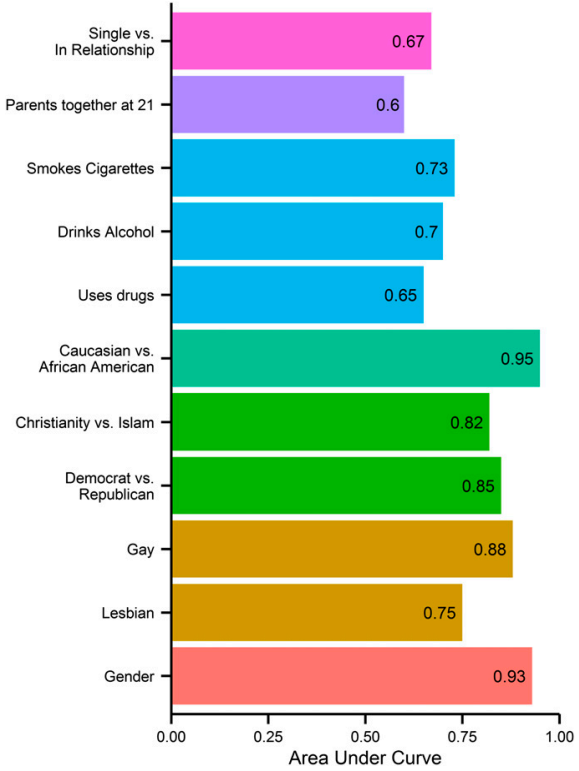


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Kosinski, Stillwell, & Graepel (2012, PNAS)

Strong Predictors of male homosexuality included

- “No H8 Campaign,”
- “Mac Cosmetics,” and
- “Wicked The Musical,”

Strong Predictors of male heterosexuality included

- “Wu-Tang Clan,”
- “Shaq,” and
- “Being Confused After Waking Up From Naps.”

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Kosinski, Stillwell, & Graepel (2012, PNAS)

How hard do you think it would be to predict "parent"?

"gun owner?"

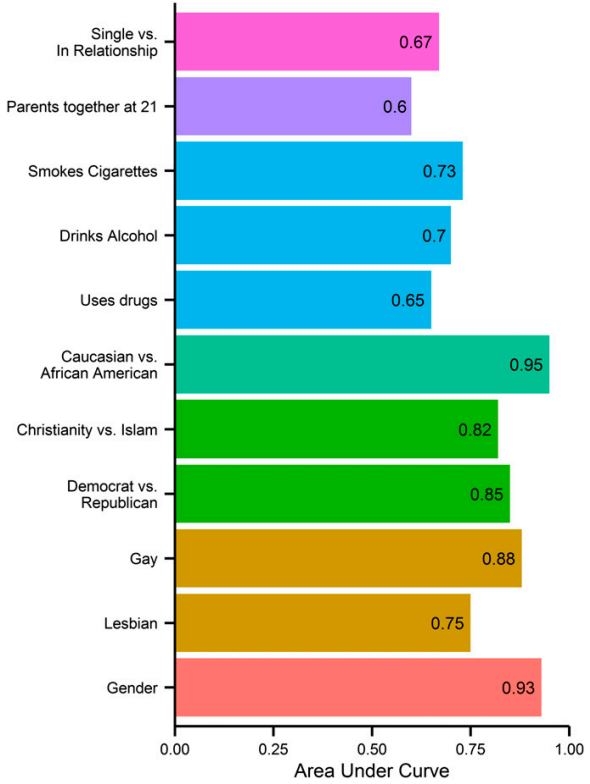


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

© 2018 Carnegie Mellon University

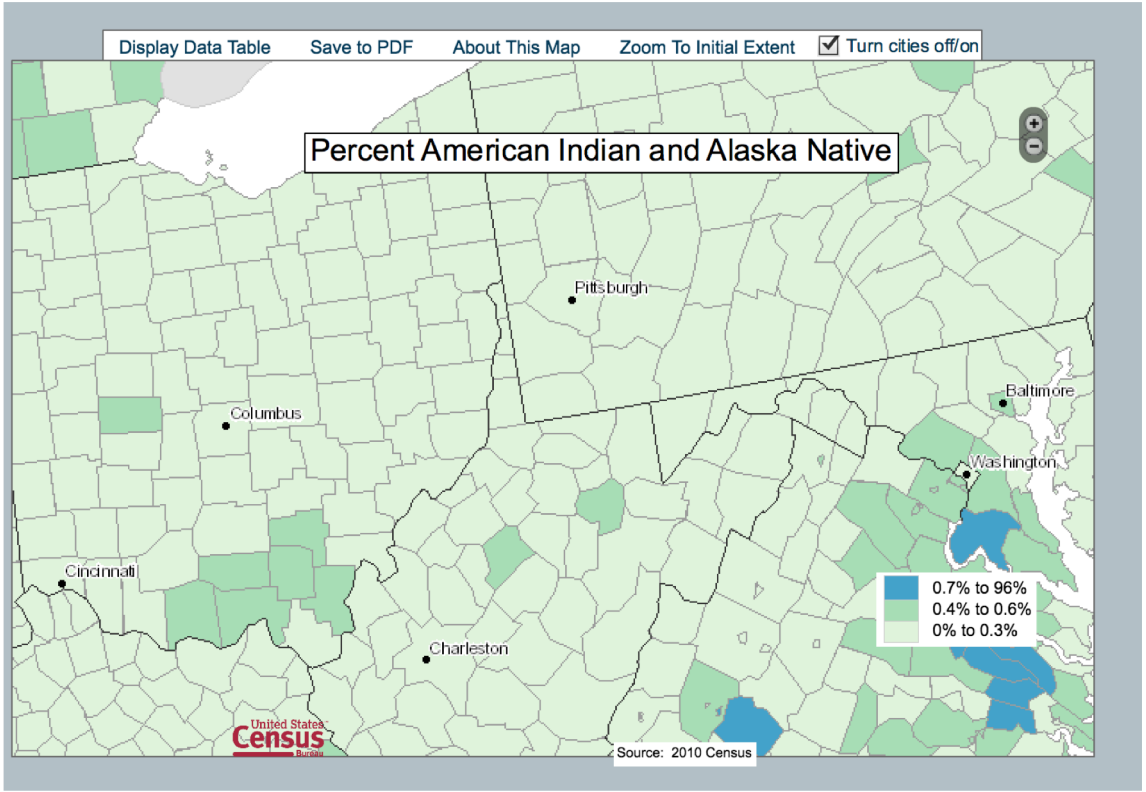
[Distribution Statement A] Approved for public release and unlimited distribution.

Anonymization is Easy to Break

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Anonymization is easily broken



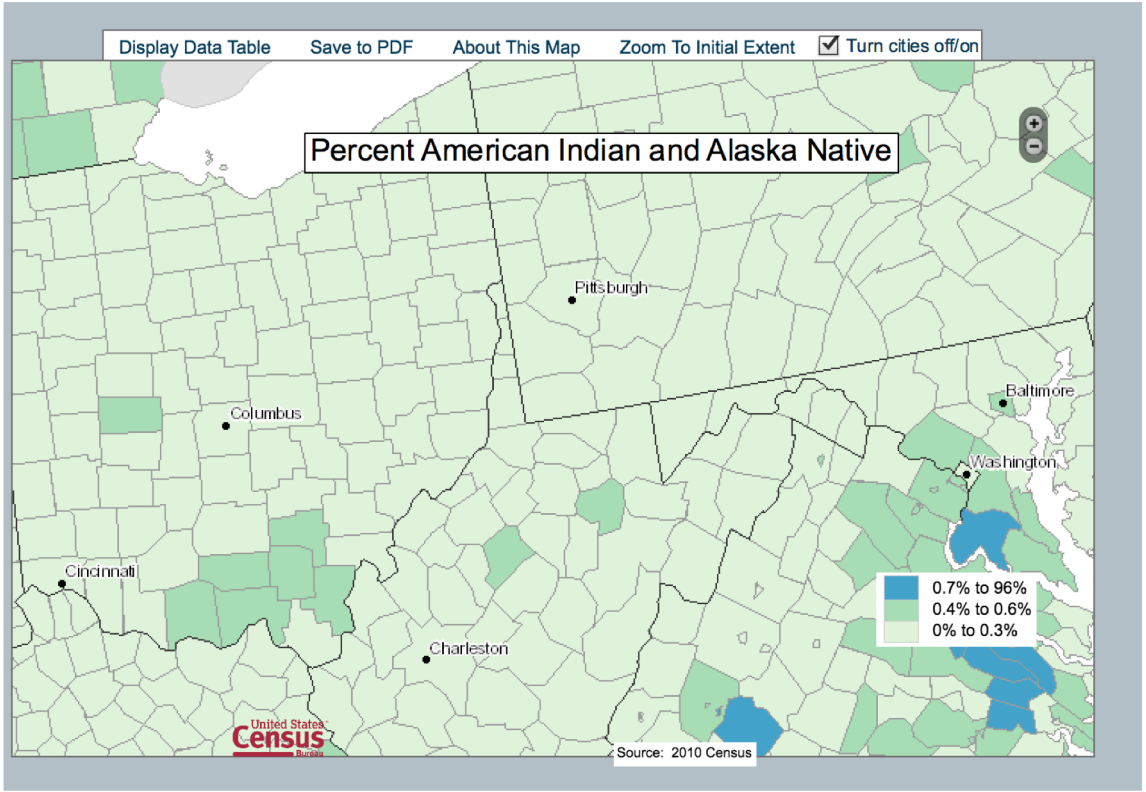
Earlier versions of the Census Data Mapper allowed you to map demographic information on a neighborhood level.

Now, only county-level maps are available.

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

The US Census



53% of the U.S. population can be uniquely identified if you have

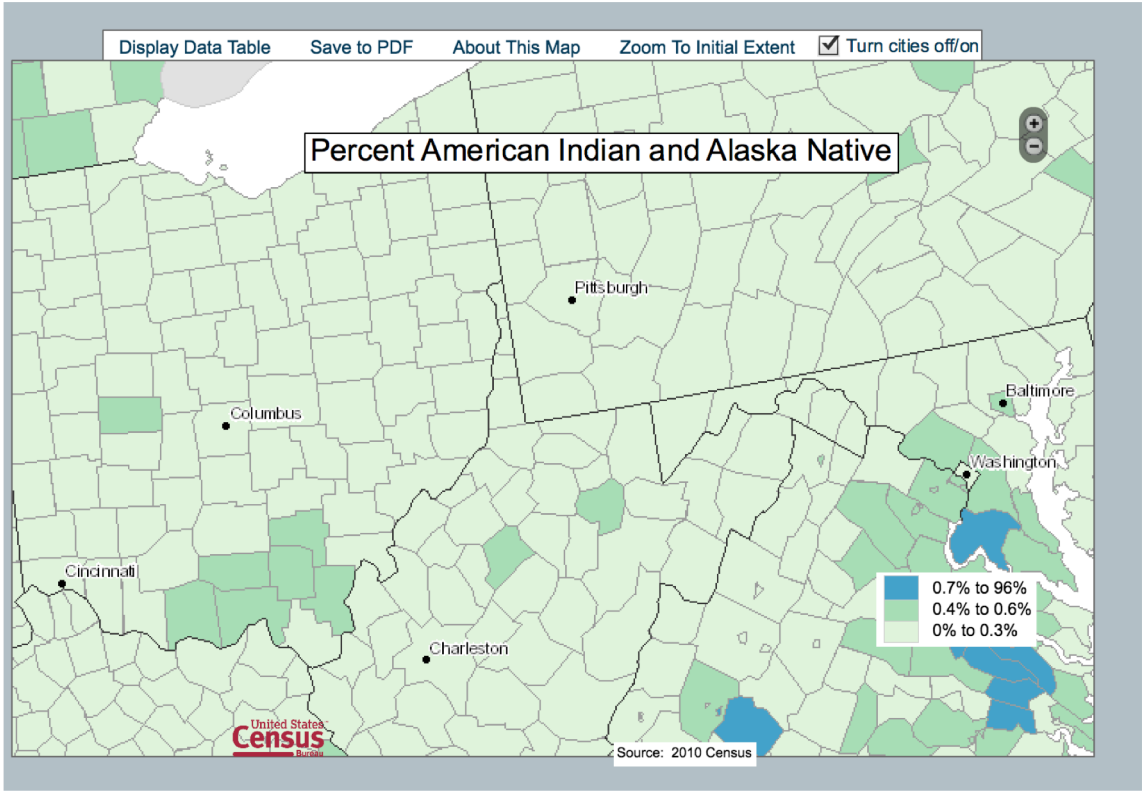
- place,
- gender,
- date of birth

Sweeny (2000)

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

The US Census



87% of the U.S. population can be uniquely identified if you have

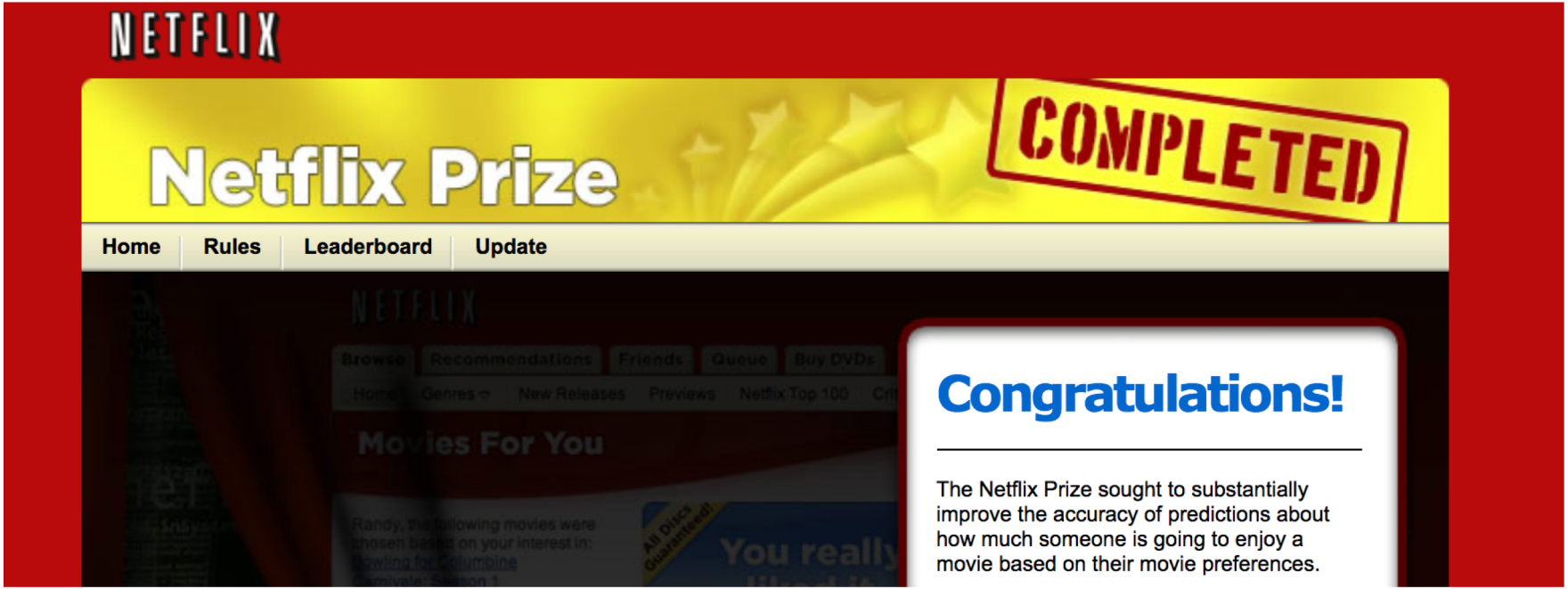
- place => Zip Code,
- gender,
- date of birth

Sweeny (2000)

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

The Netflix Prize



© 2018 Carnegie Mellon University
[Distribution Statement A] Approved for public release and unlimited distribution.

The Netflix Prize

arXiv.org > cs > arXiv:cs/0610105

Computer Science > Cryptography and Security

How To Break Anonymity of the Netflix Prize Dataset

Arvind Narayanan, Vitaly Shmatikov

(Submitted on 18 Oct 2006 (v1), last revised 22 Nov 2007 (this version, v2))

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistak We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ra world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowle records of known users, uncovering their apparent political preferences and other potentially sensitive informa

Subjects: **Cryptography and Security (cs.CR)**; Databases (cs.DB)
Cite as: **arXiv:cs/0610105 [cs.CR]**
(or **arXiv:cs/0610105v2 [cs.CR]** for this version)

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

The record of movies you watch is a fingerprint

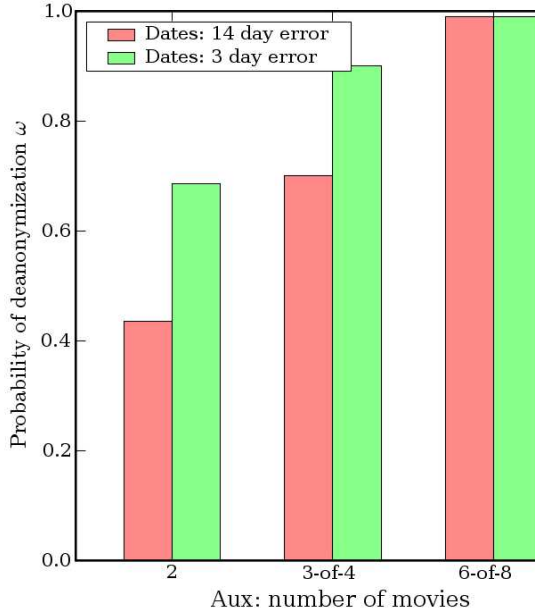


Figure 1: De-anonymization: adversary knows exact ratings and approximate dates.

68% of records can be re-identified with 2 movie ratings and dates accurate to within 3 days

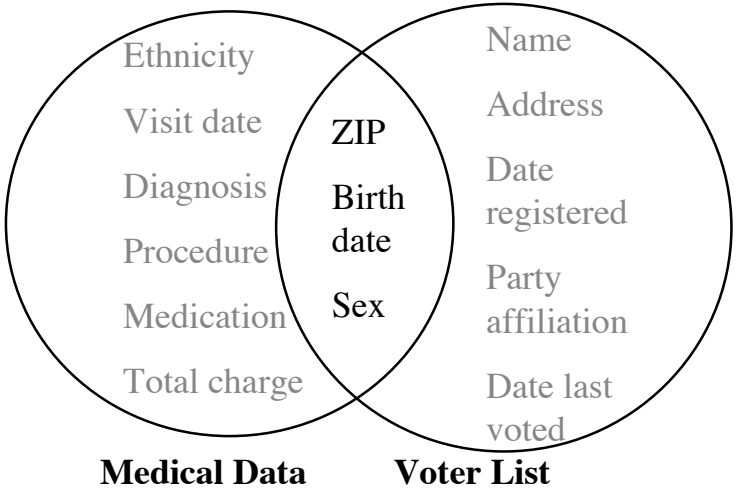
99% of records can be re-identified with 8 movie ratings (2 of which might be wrong) and dates accurate to within 2 weeks.

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Health Records

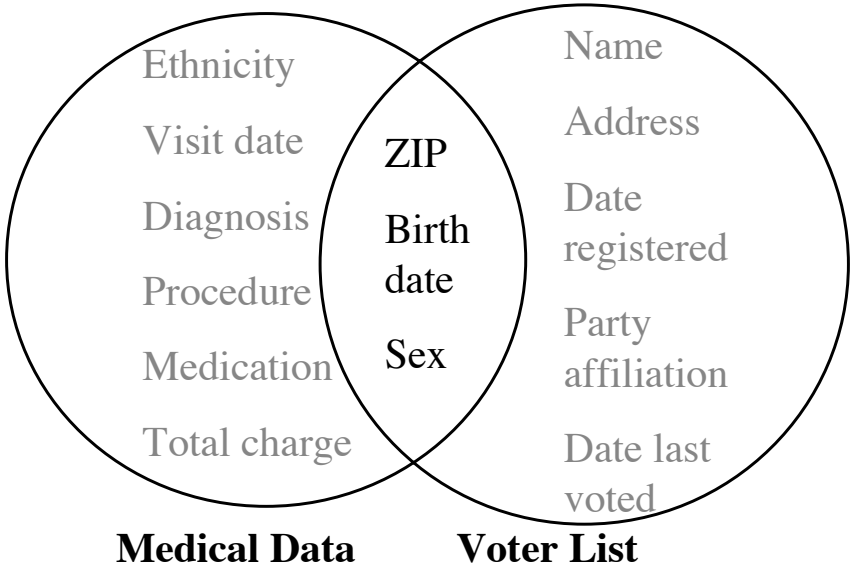
1997, Group insurance commission released *anonymized* health records from Massachusetts.



© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Health Records



Sweeny (2002), identified Governor Weld's medical records:

- 6 people shared his birth date
- Only 3 were men
- He was the only one in his zip code.

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

There are Real Consequences

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

There are real consequences

Marketing for the wrong products

- Annoyance for most of us
- Reveals a medical condition at work

Someone gets categorized as “Diabetes Interest”

- Ads for sugar- free products?
- High risk insurance category?

© 2017 Carnegie Mellon University
[DISTRIBUTION STATEMENT Please copy and paste the appropriate distribution statement into this space.]

There are real consequences

Facilitates

- harassment
- stalking,

Exposes

- domestic violence victims,
- law enforcement officers,
- prosecutors,
- public officials
- …

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Real Consequences

Real personal information, including

- Names
- Addresses
- Social security numbers,
- Bank account numbers

Attached to fake debts.

Packaged and sold to debt collectors.

FTC Privacy & Data Security Update: 2017

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

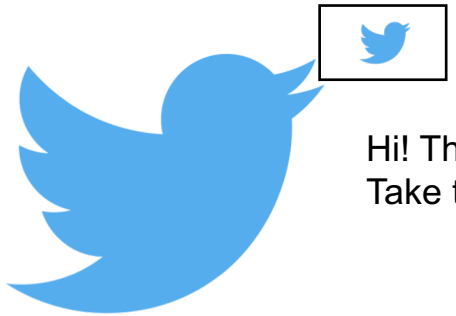
Personal Precautions

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Cookies

Small text files stored on your machine by websites you visit, to remember information specific to you



Hi! Thanks for logging in!
Take this card and show it to me next time you visit.

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

First-Party Cookies

Cookies set or requested about the domain you're visiting

Hi! Can you show me your ?
I want to see if you're logged in.

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Third-Party Cookies

Cookies set or requested about a domain other than the one you're visiting

Hi! Can you show me your ?
If you're logged into Twitter, I'll add a tweet box to this page.

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Third-Party Cookies

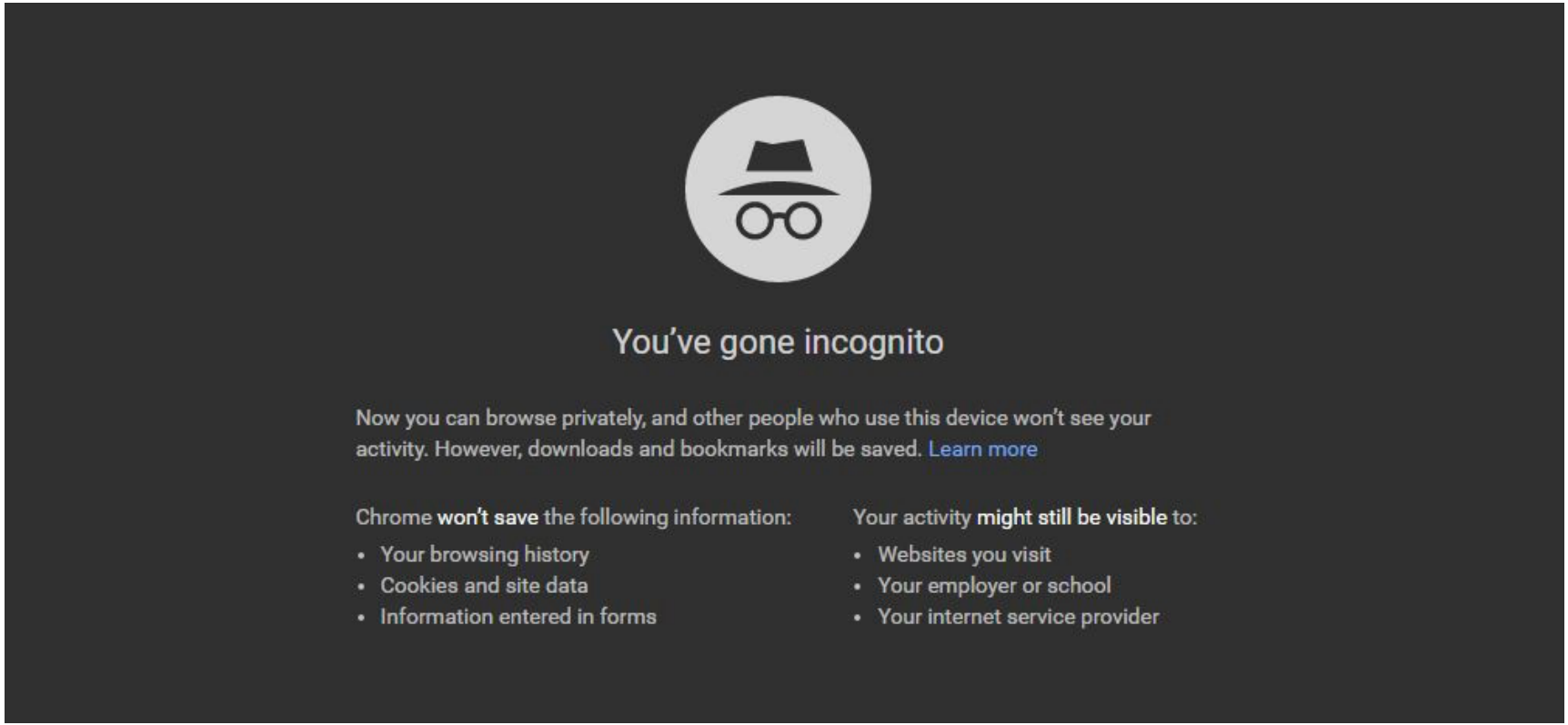
Cookies set or requested about a domain other than the one you're visiting

Hi! I see that you're on a website about Pittsburgh public transit. I'll record that on and show you relevant ads.

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Incognito/Private Browsing



© 2018 Carnegie Mellon University
[Distribution Statement A] Approved for public release and unlimited distribution.

Disabling Third-Party Cookies

First-party cookies are generally not harmful, don't need to be turned off

Third-party cookies can be disabled in each browser you use

Details vary based on which browser you use. E.g. some browsers only disable *setting* third-party cookies but don't disable *reading*

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Browser fingerprinting

Websites ask your browser for information about itself and your computer so that content can be displayed most effectively, e.g.

What browser and operating system are in use?
What plugins are installed?
What is the screen resolution and color depth?
What fonts are installed?

This information can be used to **uniquely identify you**.

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Browser fingerprinting

In one study, **79.4%** of fingerprints were unique

source: amiunique.org

© 2018 Carnegie Mellon University
[Distribution Statement A] Approved for public release and unlimited distribution.

Fingerprinting and cookies example

Wall Street Journal 2012:

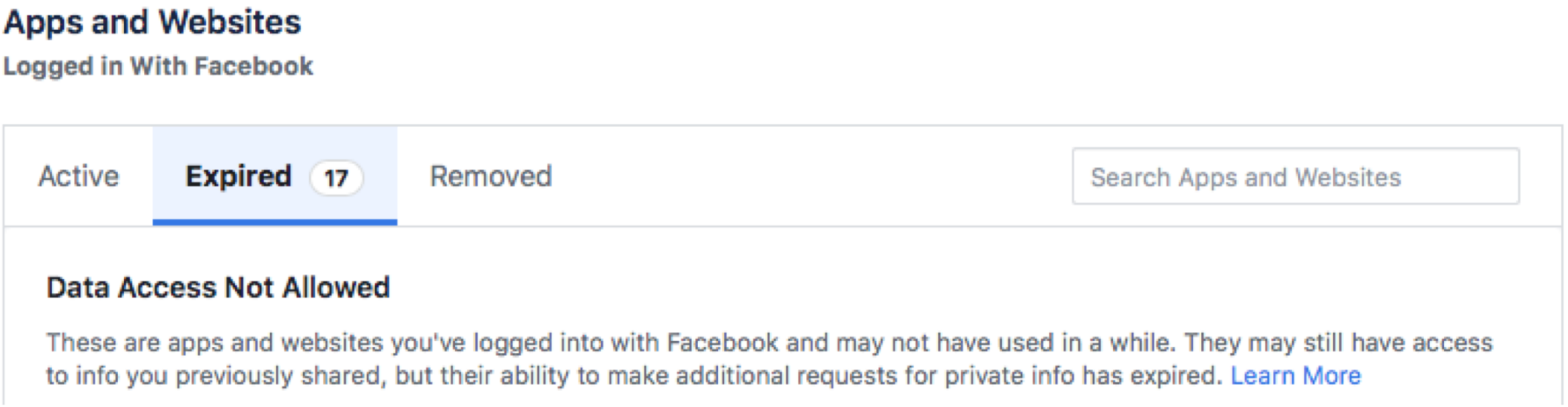
“Orbitz Worldwide Inc. has found that people who use Apple Inc.'s Mac computers spend as much as 30% more a night on hotels, so the online travel agency is starting to show them different, and sometimes costlier, travel options than Windows visitors see.”

Orbitz also used this data to influence ranking:

referring site: the site a user follows a link from to get to Orbitz
return visits: booking history and previous activity on the site
location

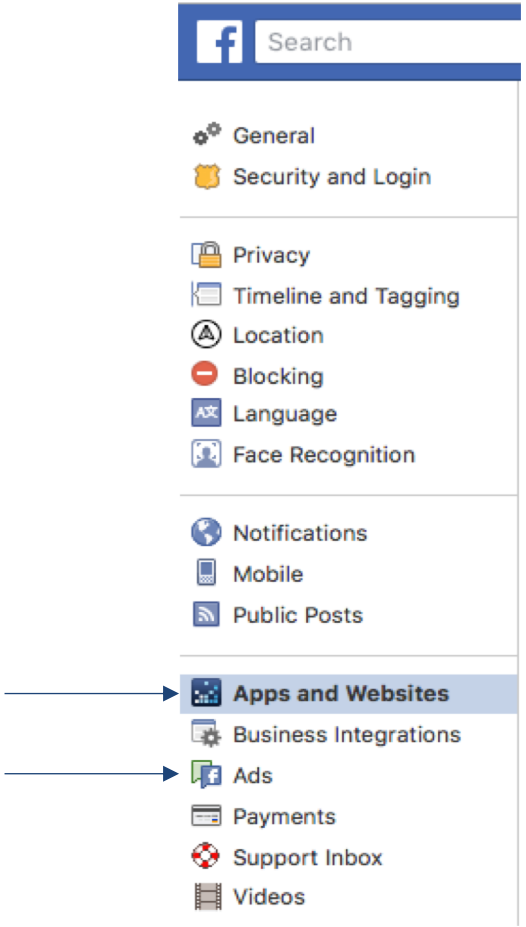
© 2018 Carnegie Mellon University
[Distribution Statement A] Approved for public release and unlimited distribution.

Revoking access



© 2018 Carnegie Mellon University
[Distribution Statement A] Approved for public release and unlimited distribution.

Seeing what's known about you, and controlling what's shared



© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Takeaways

1. Metadata is Data.
2. There’s no such thing as anonymous.
3. There are real consequences.

© 2018 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.