The End of Anonymity

Vitaly Shmatikov

Tastes and Purchases

Social Networks

TAGGED
LinkedIn
cyworld
flickr
facebook
badoo
LiVEJOURNAL
Pinterest
BlackPlanet
twitter
deviantART
my HERITAGE
MULTIPLY

Health Care and Genetics

23andMe

wellcome trust sanger institute

patientslikeme

CERNER All Together

deCODE genetics

Navigenics

Web Tracking

exelate

ZEDO
Ad Solutions for The New Internet

atlas.

bluekai

[x+1]

PubMatic
Make every impression count

LOTAME™

rightmedia
from YAHOO!

quantcast
It's your audience. We just find it.™

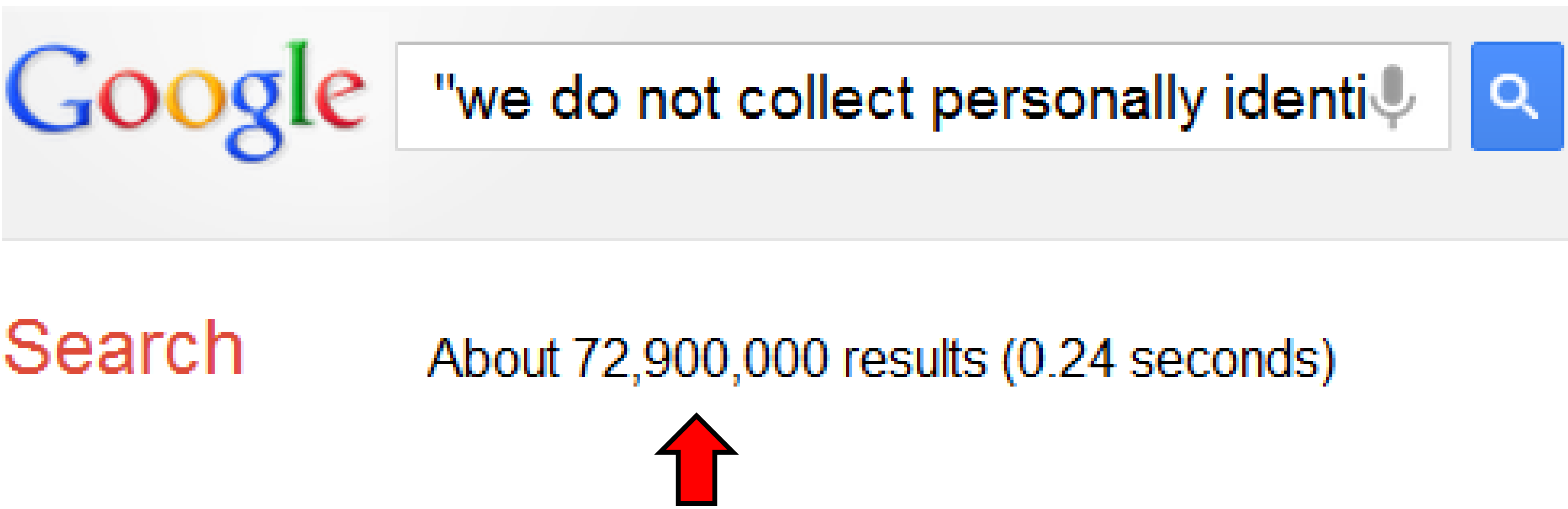
ValueClick media

Solution: Anonymity!

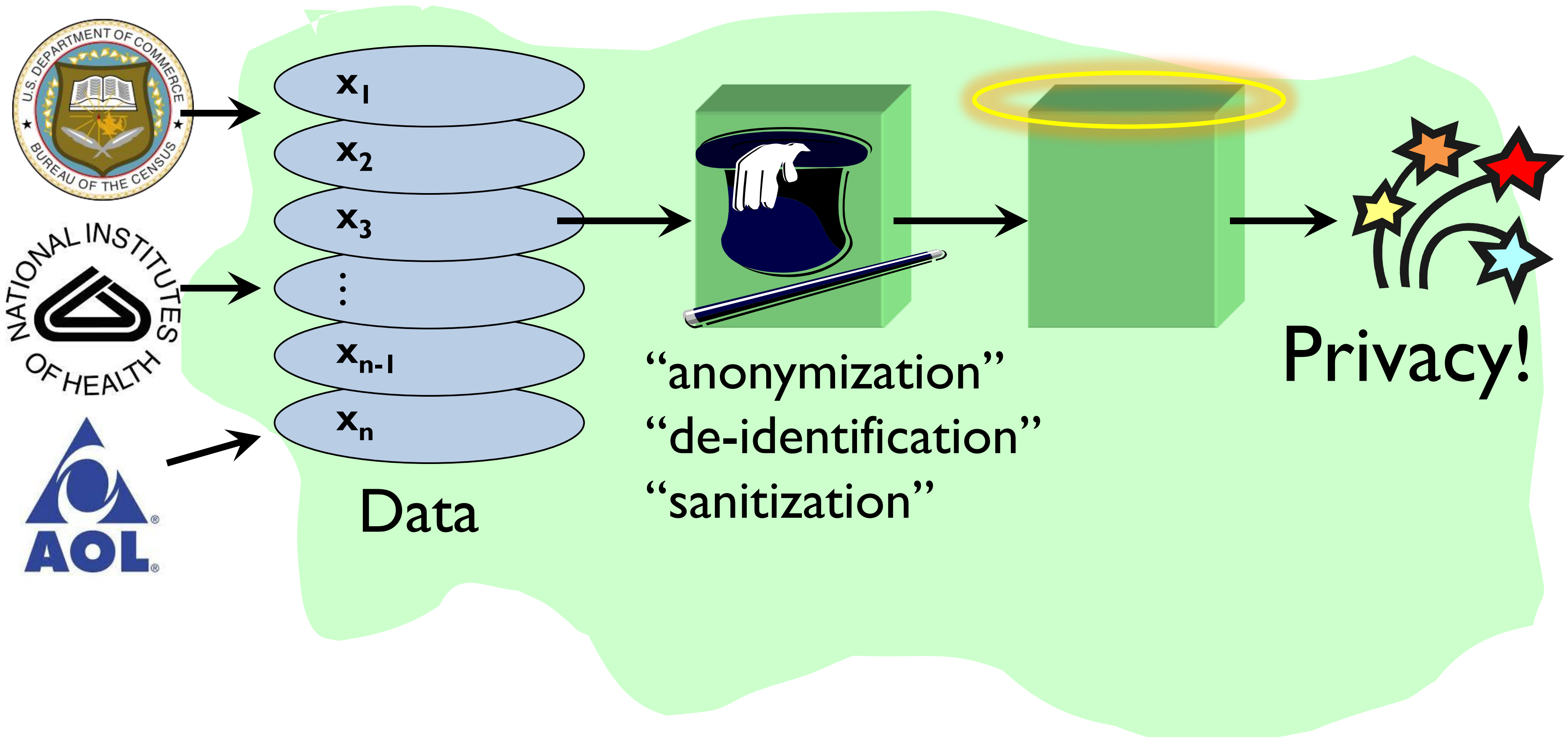
"… breakthrough technology that uses social graph data to dramatically improve online marketing … "Social Engagement Data" consists of anonymous information regarding the relationships between people"

"The critical distinction … between the use of personal information for advertisements in personally-identifiable form, and the use, dissemination, or sharing of information with advertisers in non-personally-identifiable form."

Phew…


Google
"we do not collect personally identi
Search
About 72,900,000 results (0.24 seconds)

“Privacy-Preserving” Data Release



Some Privacy Disasters

Forbes

3/12/2010 @ 12:35PM | 1,098 views

Netflix Settles Privacy Lawsuit, Cancels Prize Sequel

Taylor Buley, Forbes Staff

NEWS

Comment 3

AOL Proudly Releases Massive Amounts of Private Data

The New York Times

WORLD | U.S. | N.Y. / REGIO | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS

otect Medical Data

What went wrong?

Genomics Law Report

Back to the Future: NIH to Revisit Genomic Data-Sharing Policy

THE CHRONICLE of Higher Education

Subscri

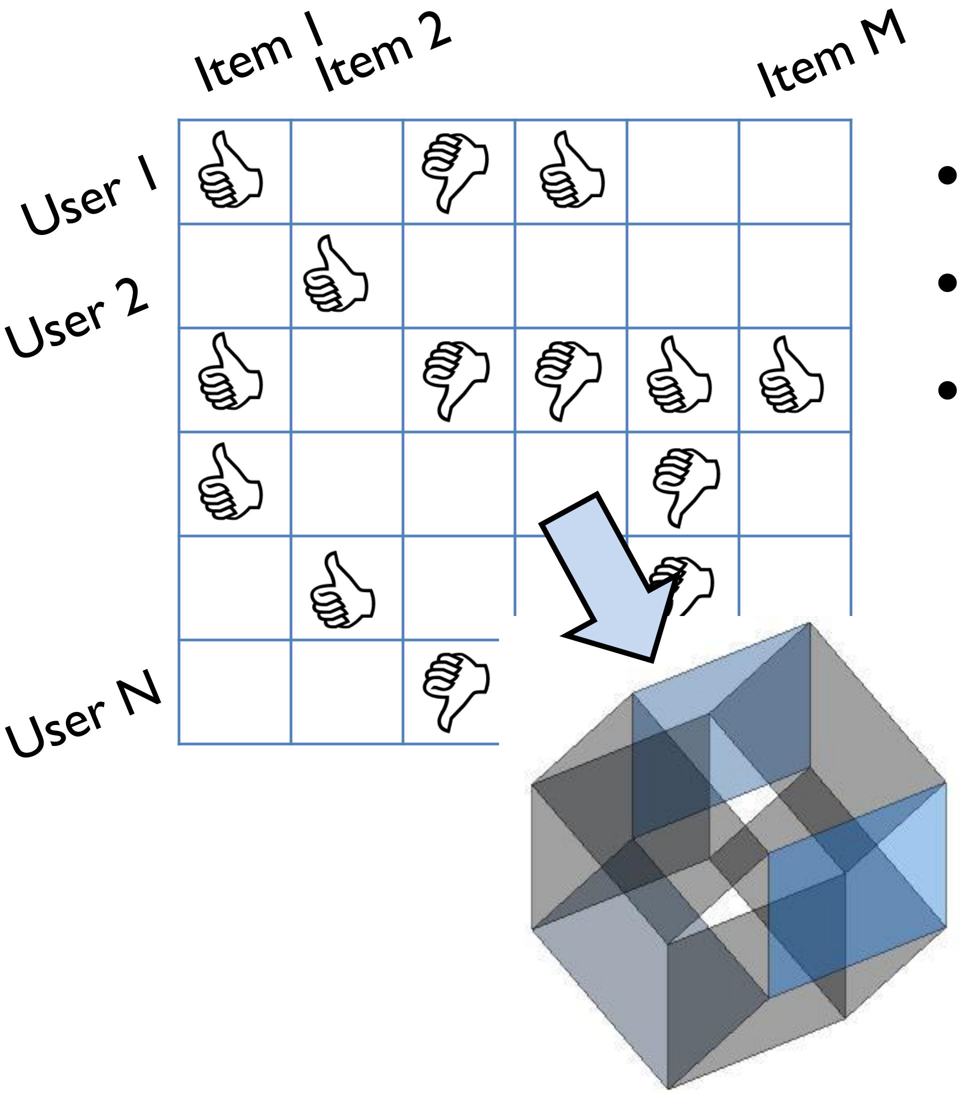
Harvard's Privacy Meltdown, Revisited: Controversial Facebook Data Yield New Paper

TARGET

The Myth of the PII

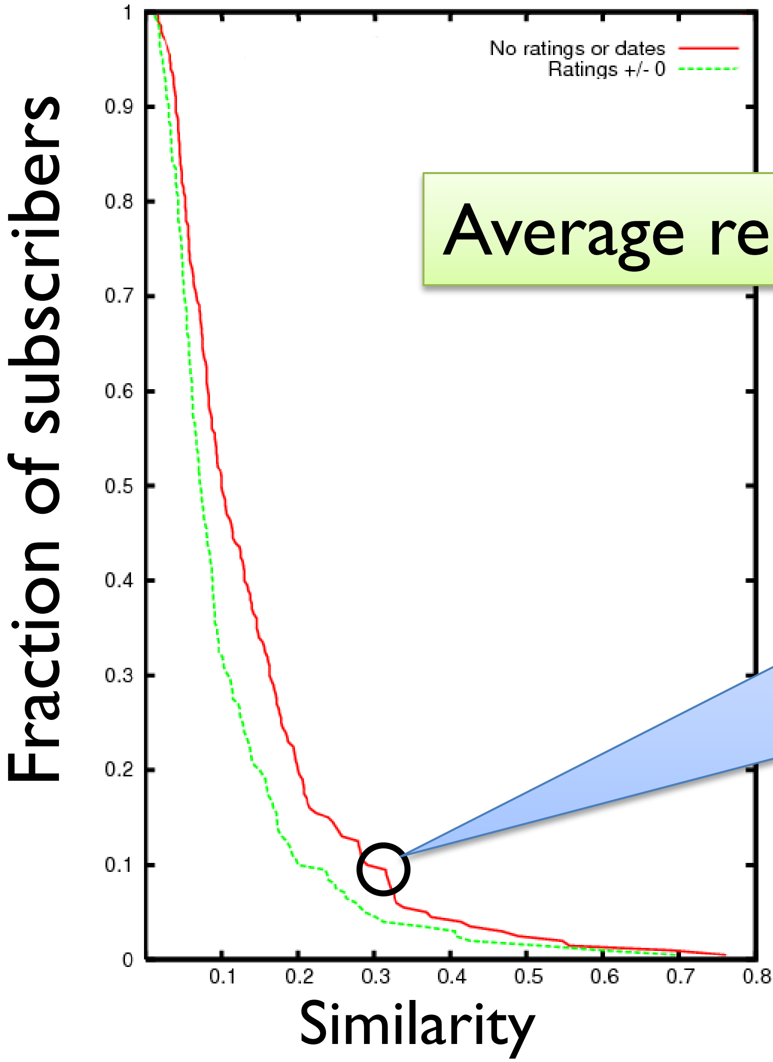
- Data are "anonymized" by removing personally identifying information (PII)
  - Name, Social Security number, phone number, email, address… what else?
- Problem: PII has no technical meaning
  - Defined in disclosure notification laws (if certain information is lost, consumer must be notified)
  - In privacy breaches, any information can be personally identifying

The Curse of Dimensionality



- Row = user record
- Column = dimension
- Thousands or millions of dimensions
  - Netflix movie ratings: 35,000
  - Amazon purchases: 10^7

Sparsity and “Long Tail”

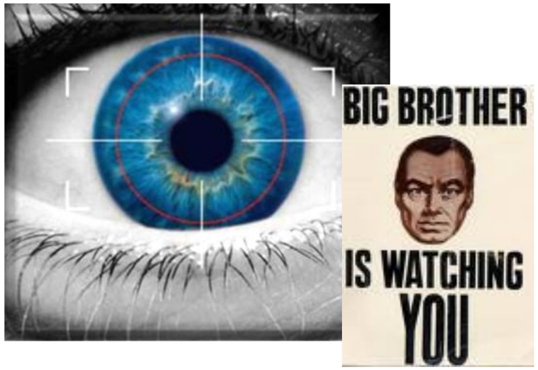


Average record has no “similar” records

Netflix Prize dataset:

Considering just movie names, for 90% of records there isn’t a <u>single</u> other record which is more than 30% similar

Privacy Threats



Global surveillance

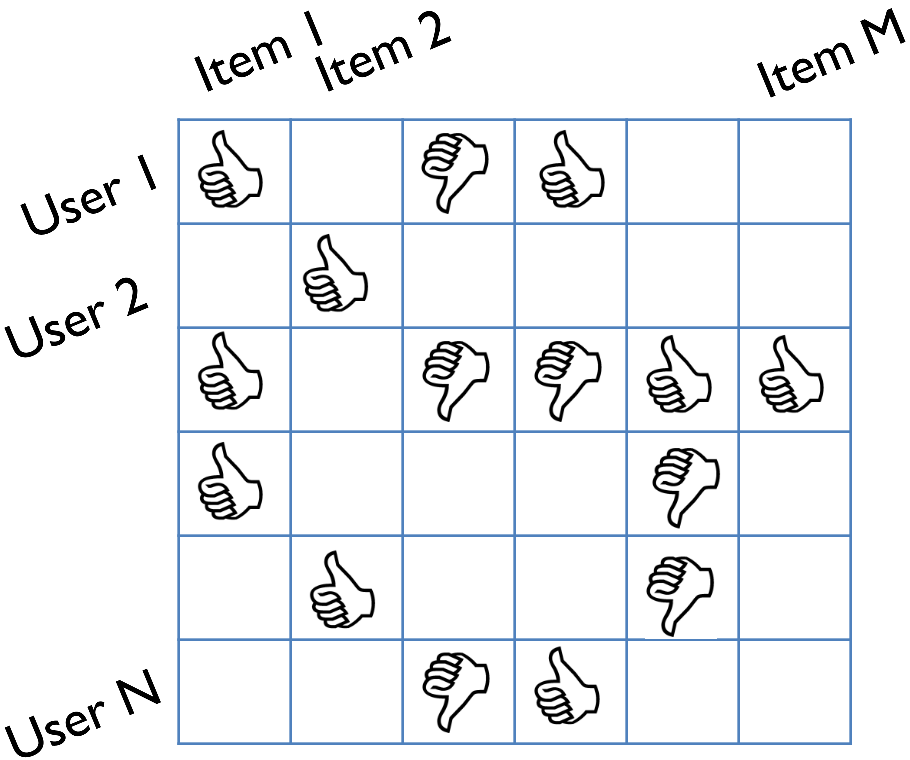
Spammers

Abusive advertisers and marketers

Phishing

Employers, insurers, stalkers, nosy friends

It's All About the Aux

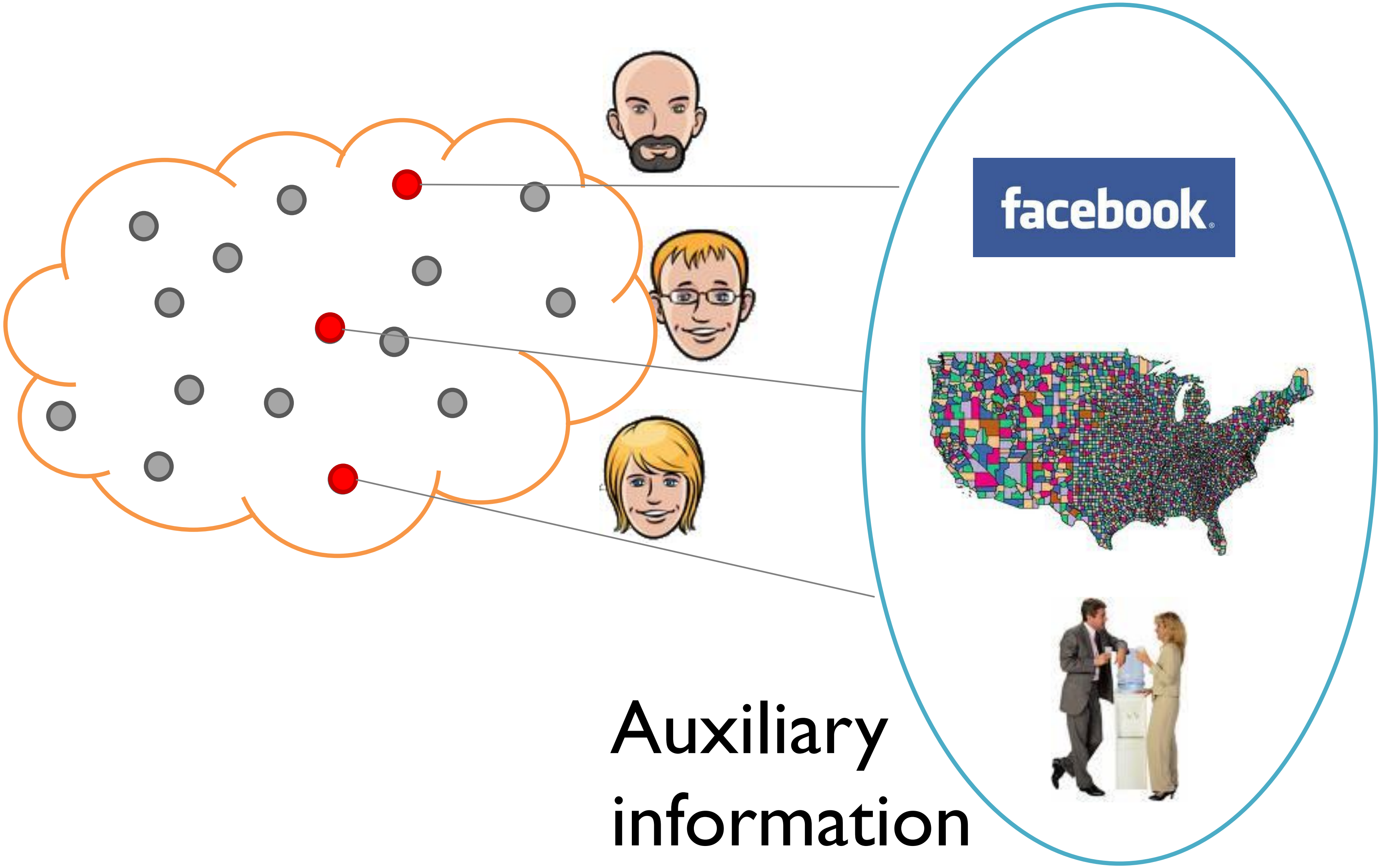


No explicit identifiers

What can the adversary learn by combining this with auxiliary information?

Information available to adversary outside of normal data release process

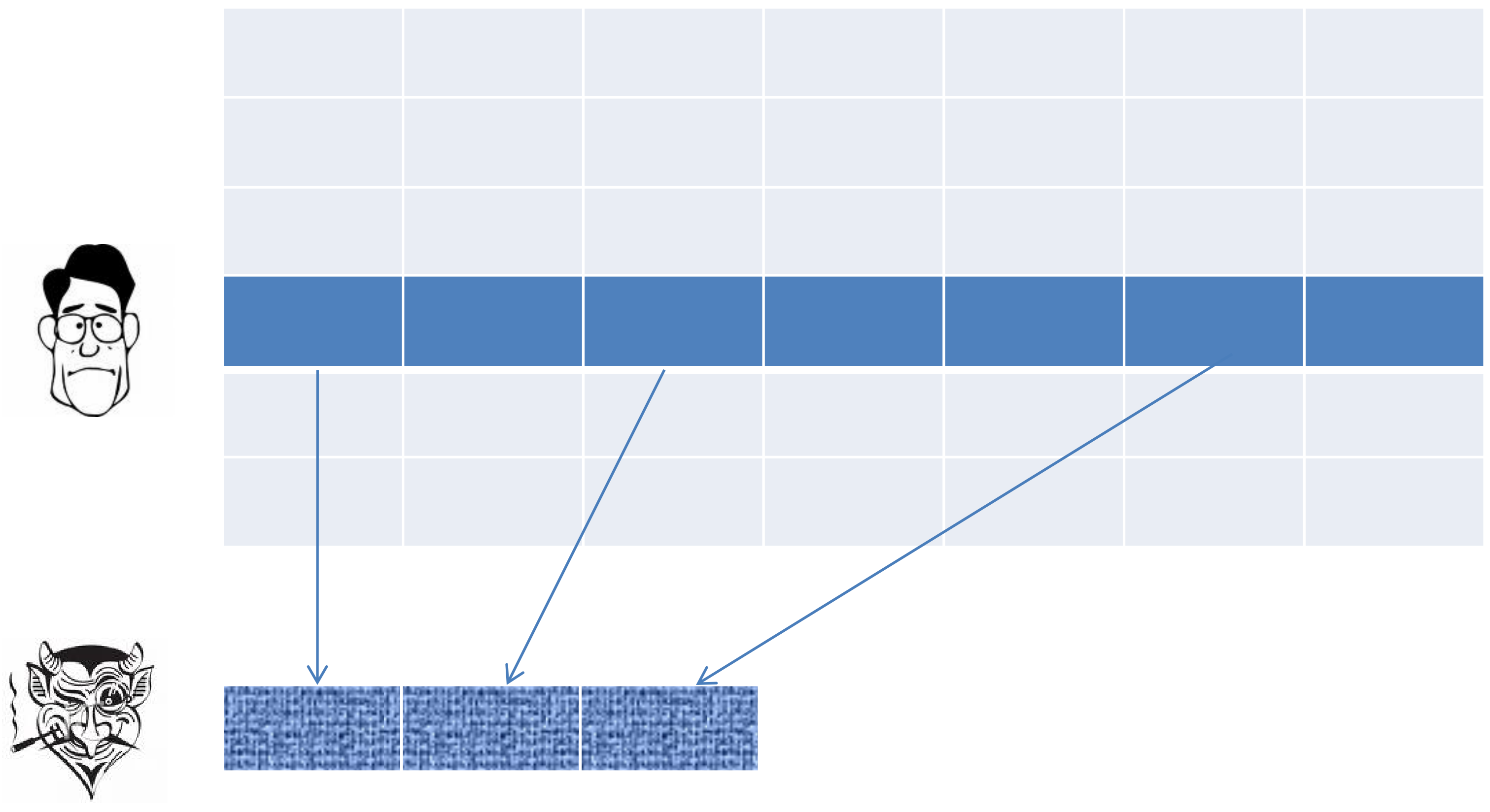
De-anonymizing Sparse Datasets



De-anonymization Objectives

- Fix some target record r in the original dataset
- Goal: learn as much about r as possible
- Subtler than "identify r in the released dataset"
  - Don't fall for the k-anonymity fallacy!
    - Silly example: released dataset contains k copies of each original record – this is k-anonymous!
  - Can't identify the "right" record, yet the released dataset completely leaks everything about r

Aux as Noisy Projection



How Much Aux Is Needed?

- How much does the adversary need to know about a record to find a very similar record in the released dataset?
  - Under very mild sparsity assumption, O(log N), where N is the number of records
- What if not enough Aux is available?
  - Identifying a small number of candidate records similar to the target still reveals a lot of information

De-Anonymization in Practice

- Sweeney (1998):
  Massachusetts hospital discharge dataset + voter database
- Narayanan and Shmatikov (2006):
  Netflix Prize dataset + IMDb
- Narayanan and Shmatikov (2009):
  social networks

Netflix Prize: Home - Microsoft Internet Explorer provided by Comcast High-Speed Internet

http://www.netflixprize.com/

NETFLIX

Netflix Prize

Home | Rules | Leaderboard | Register | Update | Submit | Download

Welcome!

The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.

Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.

You should also read the frequently-asked questions about the Prize. And check out how various teams are doing on the Leaderboard.

Good luck and thanks for helping!

FAQ | Forum | Netflix Home

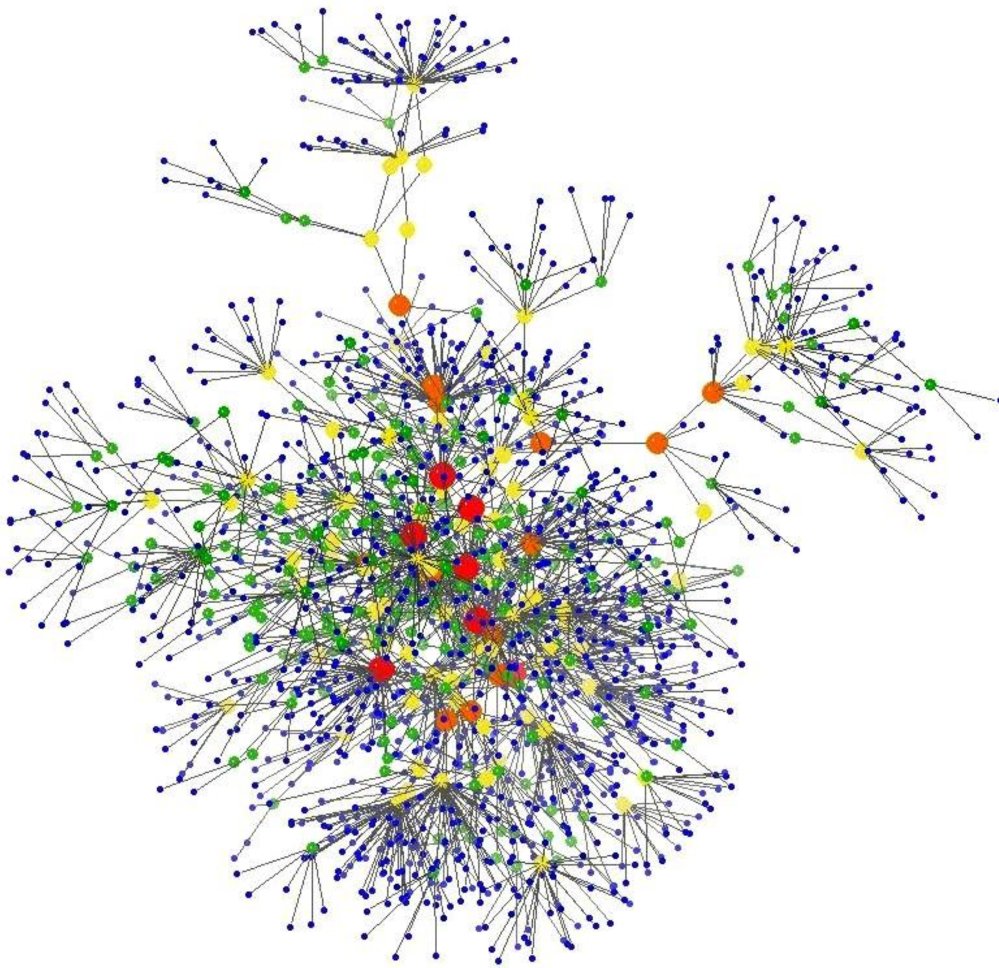
start Netflix Prize: Home - ... 9:53 AM

De-anonymizing the Netflix Dataset

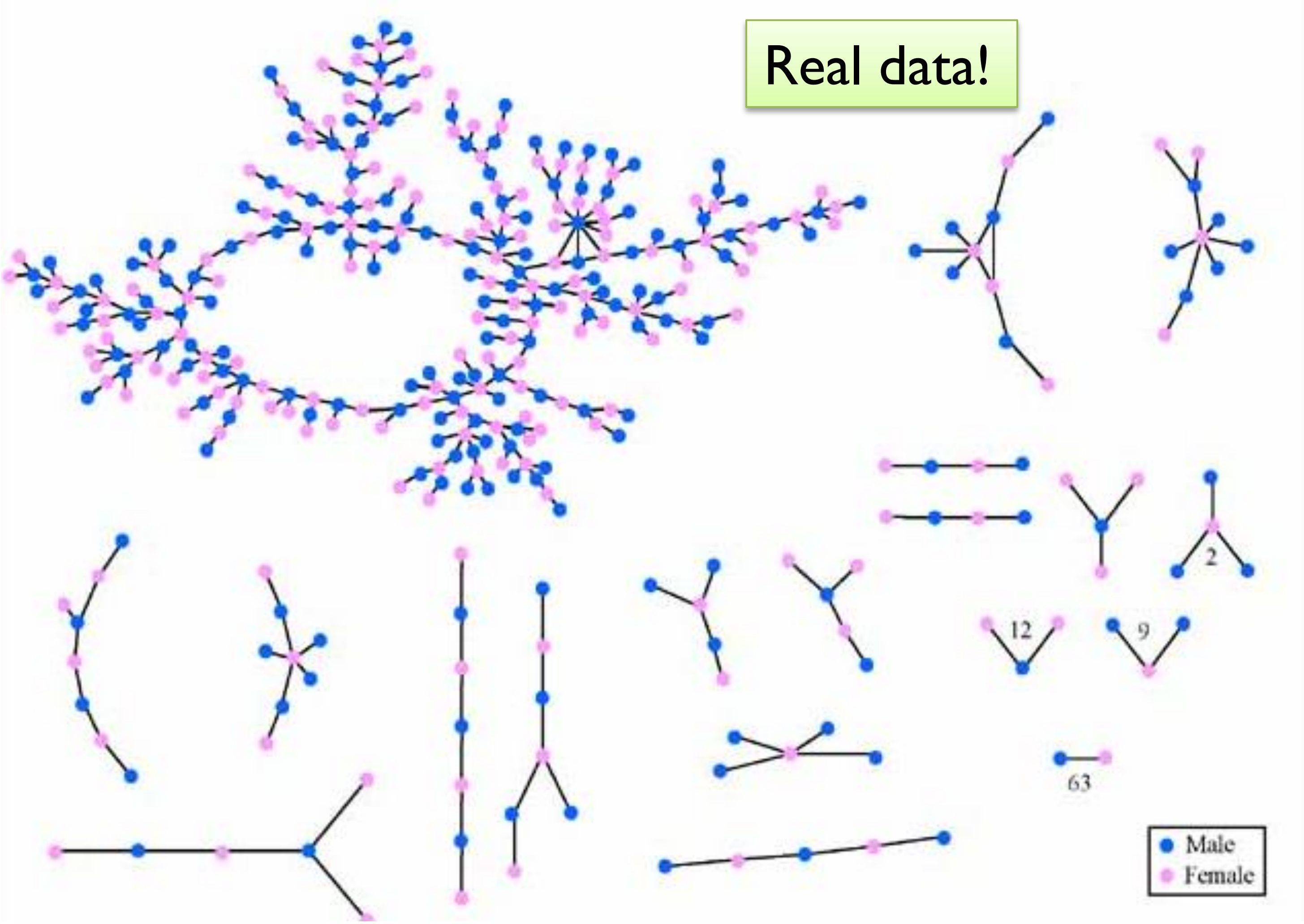
- 500K users, 18,000 movies
- 213 dated ratings per user, on average
- Two is enough to reduce to 8 candidate records
- Four is enough to identify uniquely (on average)
- Works even better with relatively rare ratings
    - "The Astro-Zombies" rather than "Star Wars"

Long Tail effect:
most people watch obscure crap

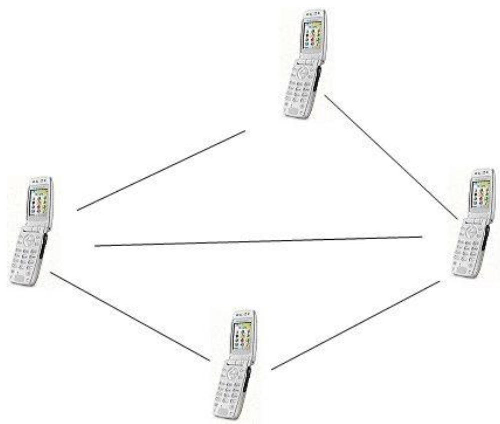
Exploiting Data Structure



“Jefferson High”: Romantic and Sexual Network



Phone Call Graphs



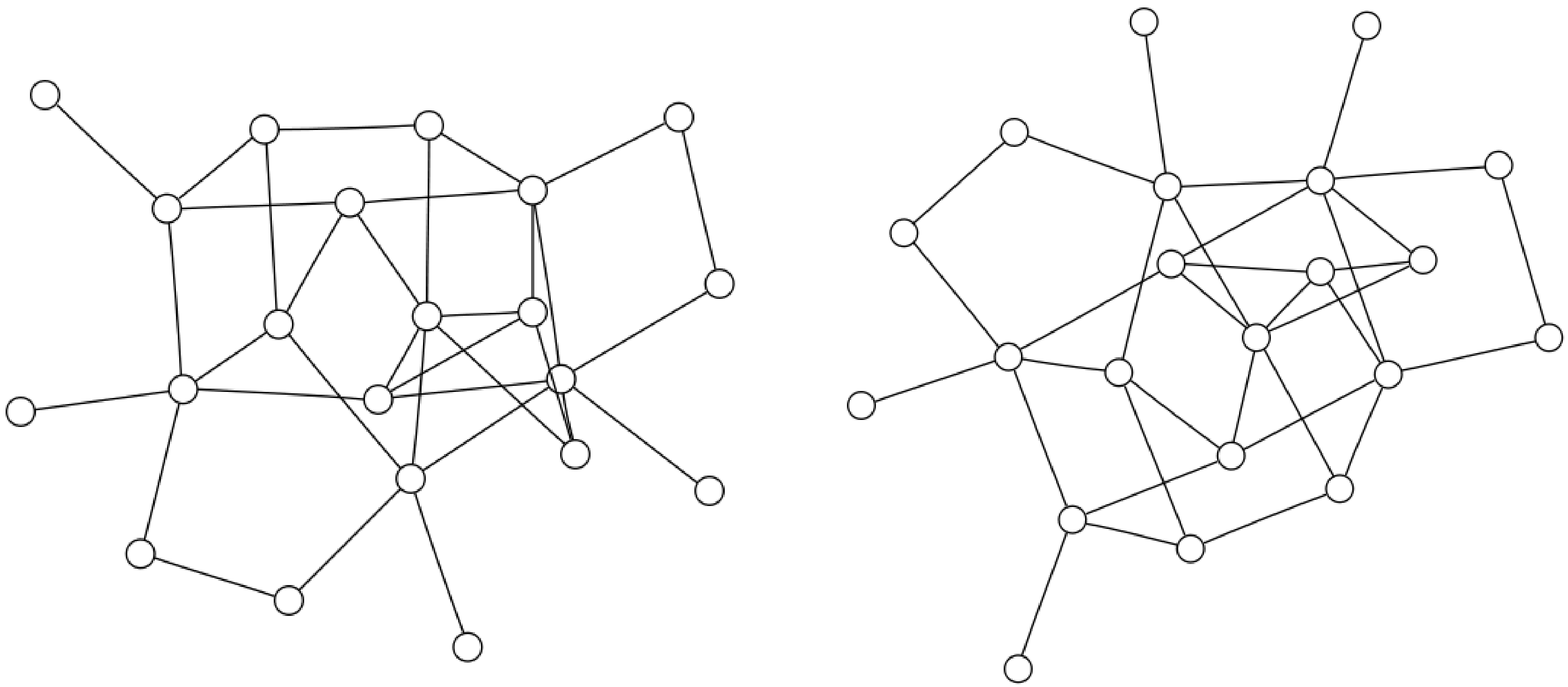
2 **trillion** edges

at&t

Examples of outsourced call graphs	
Hungary	2.5M nodes
France	7M nodes
India	3M nodes

3,000 companies providing wireless services in the U.S

Structural De-anonymization



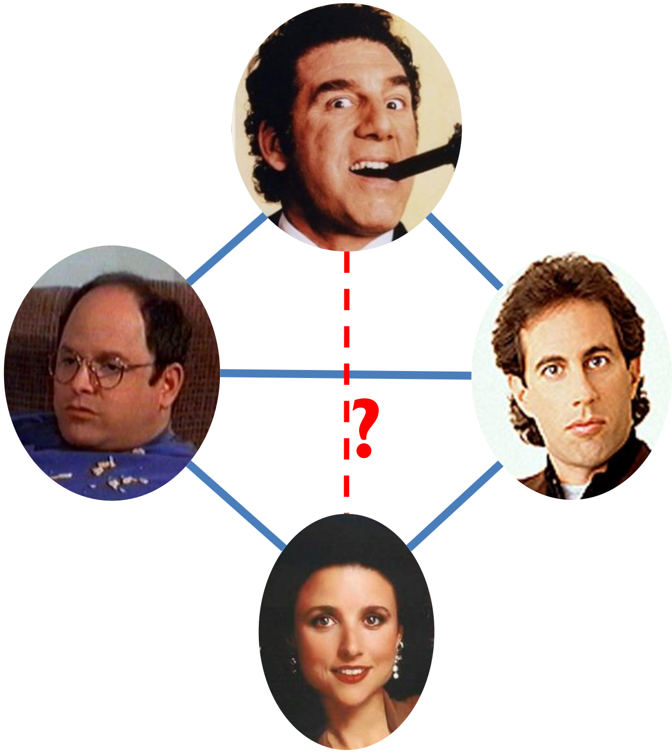
Goal: structural mapping between two graphs

For example, Facebook vs. anonymized phone call graph

Winning the IJCNN/Kaggle Social Network Challenge

[Narayanan, Shi, Rubinstein]

- "Anonymized" graph of Flickr used as challenge for a link prediction contest
- De-anonymization = "oracle" for true answers
  - 57% coverage
  - 98% accuracy



More De-Anonymization

- Social networks – again and again
- Stylometry (writing style)
- Location data
  - De Montjoye et al. (2013): mobility traces from a cell phone carrier - 4 points is enough
- Credit card transaction meta-data
  - De Montjoye et al. (2015) – 4 purchases is enough

Lesson #1: De-anonymization Is Robust

- 33 bits of entropy
  - 6-8 movies, 4-7 friends, etc.
- Perturbing data to foil de-anonymization often destroys utility
- We can estimate confidence even without ground truth
- Accretive and iterative: more de-anonymization → better de-anonymization

Lesson #2: “PII” Is Technically Meaningless

PII is info “with respect to which there is a reasonable basis to believe the information can be used to identify the individual.”

Any piece of data can be used for re-identification!

Narayanan, Shmatikov
CACM column, 2010

“blurring of the distinction between personally identifiable information and supposedly anonymous or de-identified information”