The End of Anonymity

Vitaly Shmatikov


KEEP
CALM
YOU'RE
BEING
WATCHED

Tastes and Purchases

amazon.com

last.fm™
the social music revolution

hunch

hulu

NETFLIX

Social Networks

Health Care and Genetics

23andMe

wellcome trust sanger institute

patientslikeme

CERNER All Together

deCODE genetics

Navigigenics

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

Online-Offline Aggregation

datalogix™

RapLeaf

INTELIUS
Live in the know.™

ChoicePoint

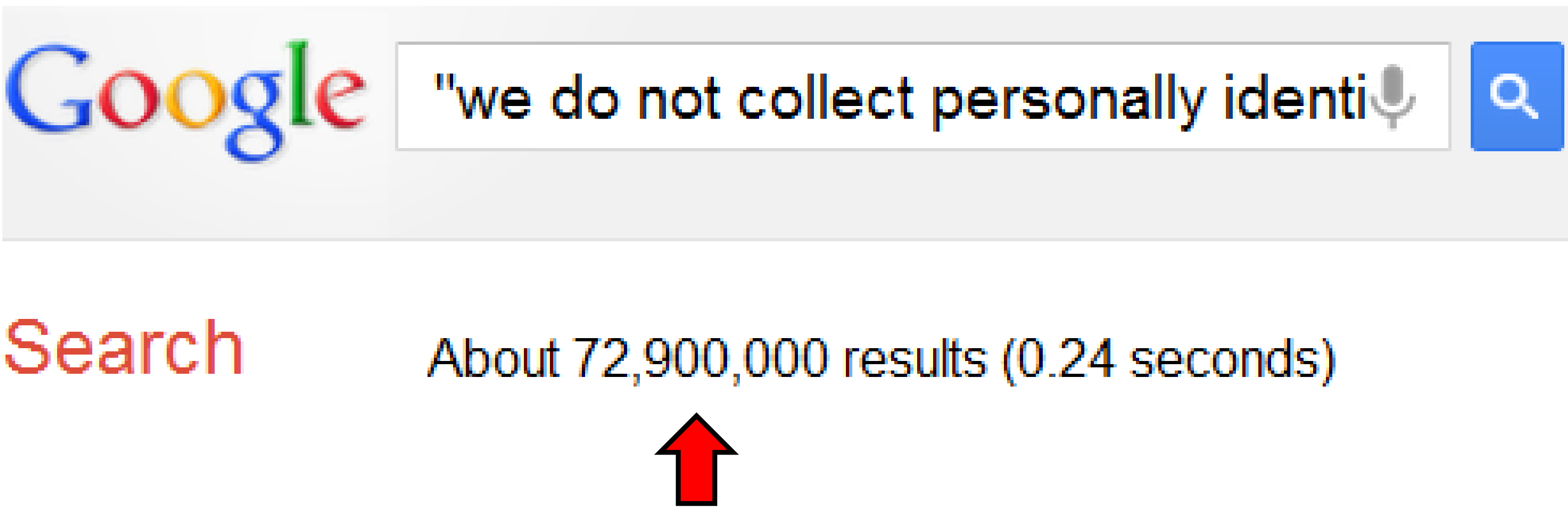
ACXIOM®

Solution: Anonymity!

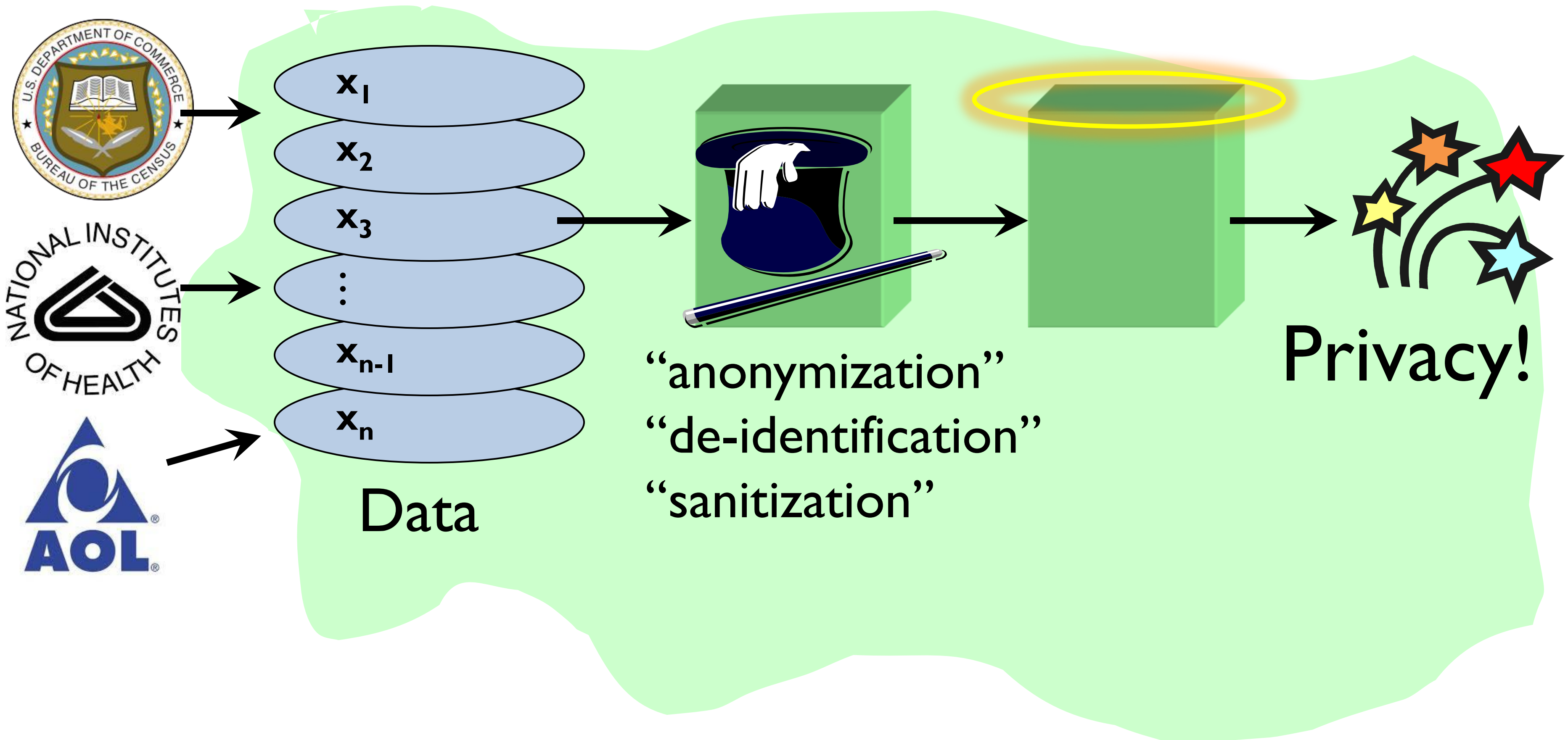
"… breakthrough technology that uses social graph data to dramatically improve online marketing … "Social Engagement Data" consists of anonymous information regarding the relationships between people"

"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



Whose Data Is It, Anyway?

"Everyone owns and should control their personal data"

- Social networks
  - Information about relationships is shared
- Genome
  - Shared with all blood relatives
- Recommender systems
  - Complex algorithms make it impossible to trace origin of data

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

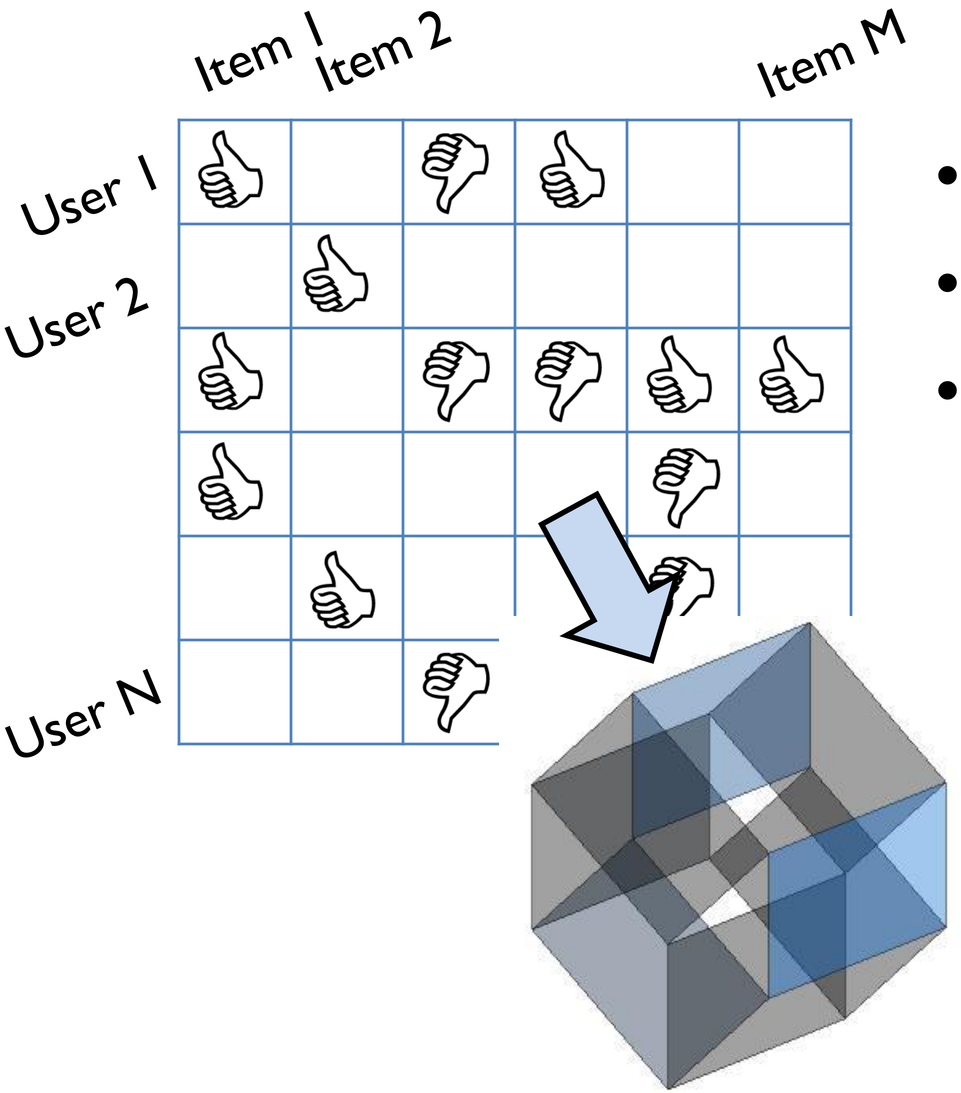
Reading Material

Sweeney
Weaving Technology and Policy Together to Maintain Confidentiality
JLME 1997

Narayanan and Shmatikov
Robust De-anonymization of Large Sparse Datasets
Oakland 2008

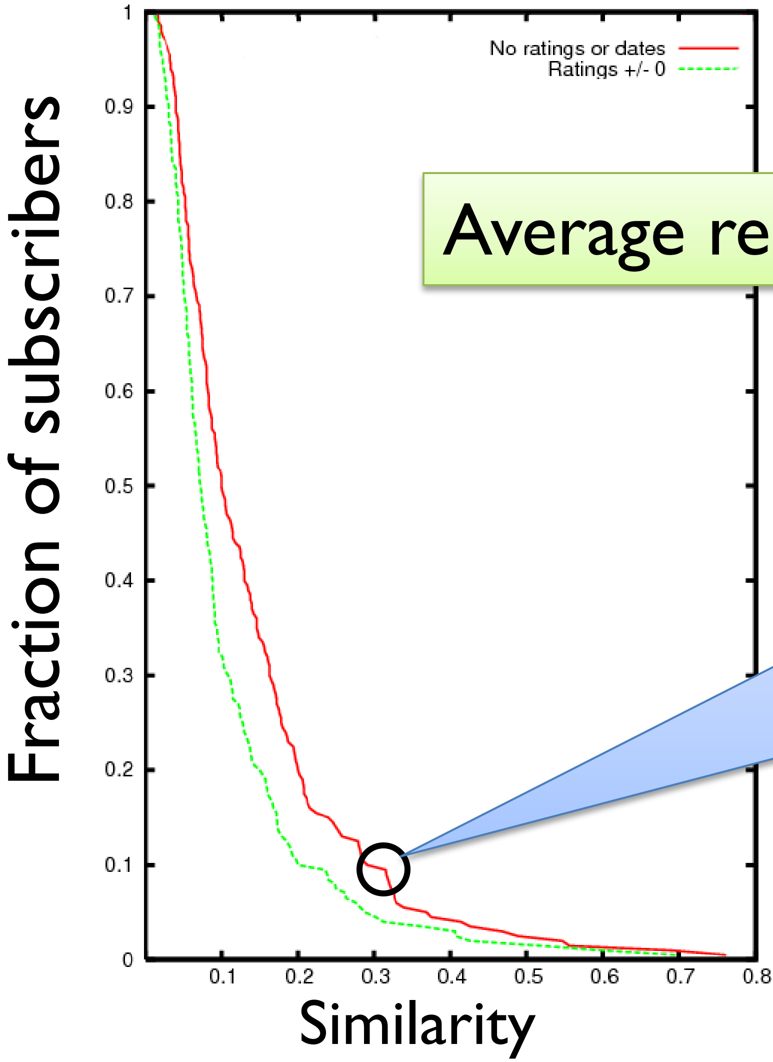
Homer et al.
Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays
PLoS Genetics 2008

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"

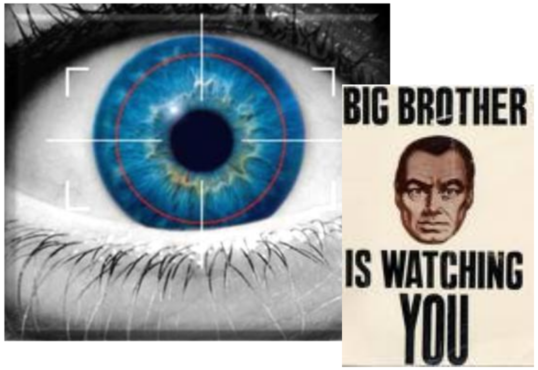


Average record has no "similar" records

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



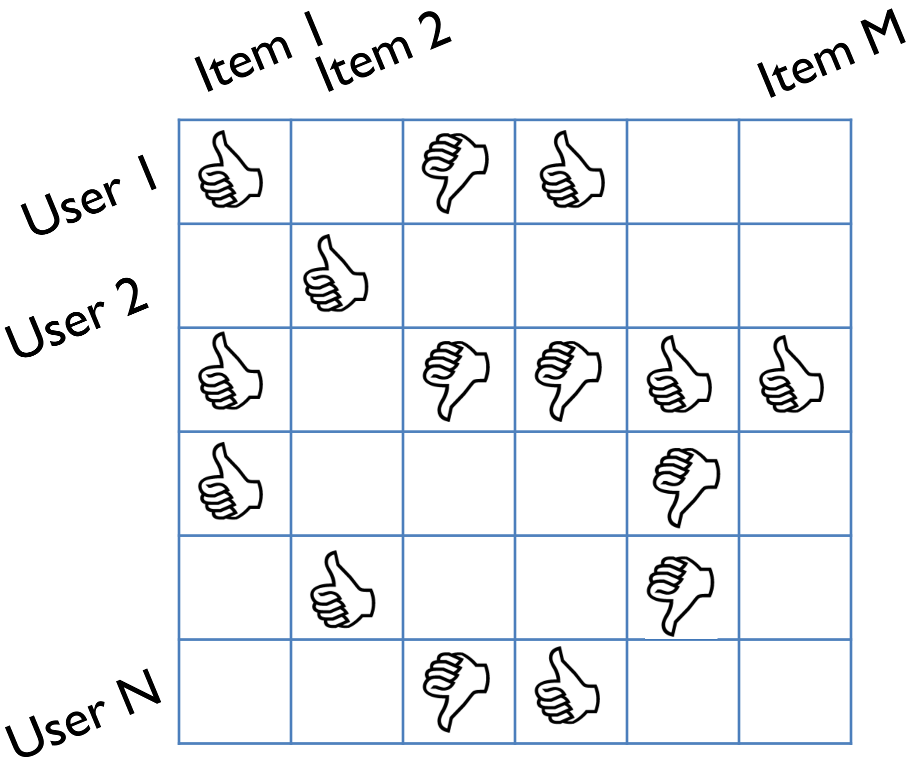
Global surveillance

Spammers
Abusive advertisers and marketers

Phishing

Employers, insurers, stalkers, nosy friends

It's All About the Aux

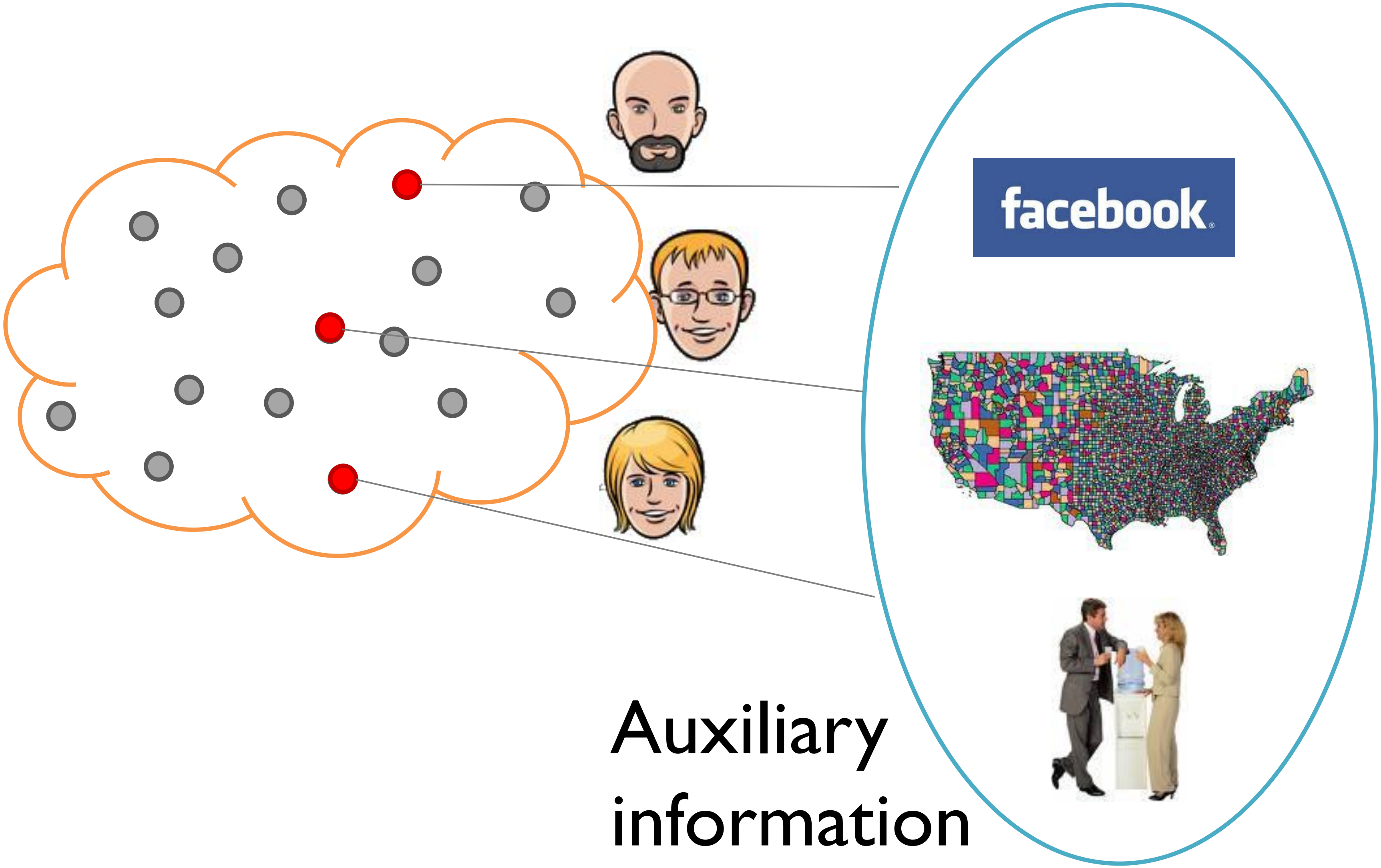


No explicit identifiers

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



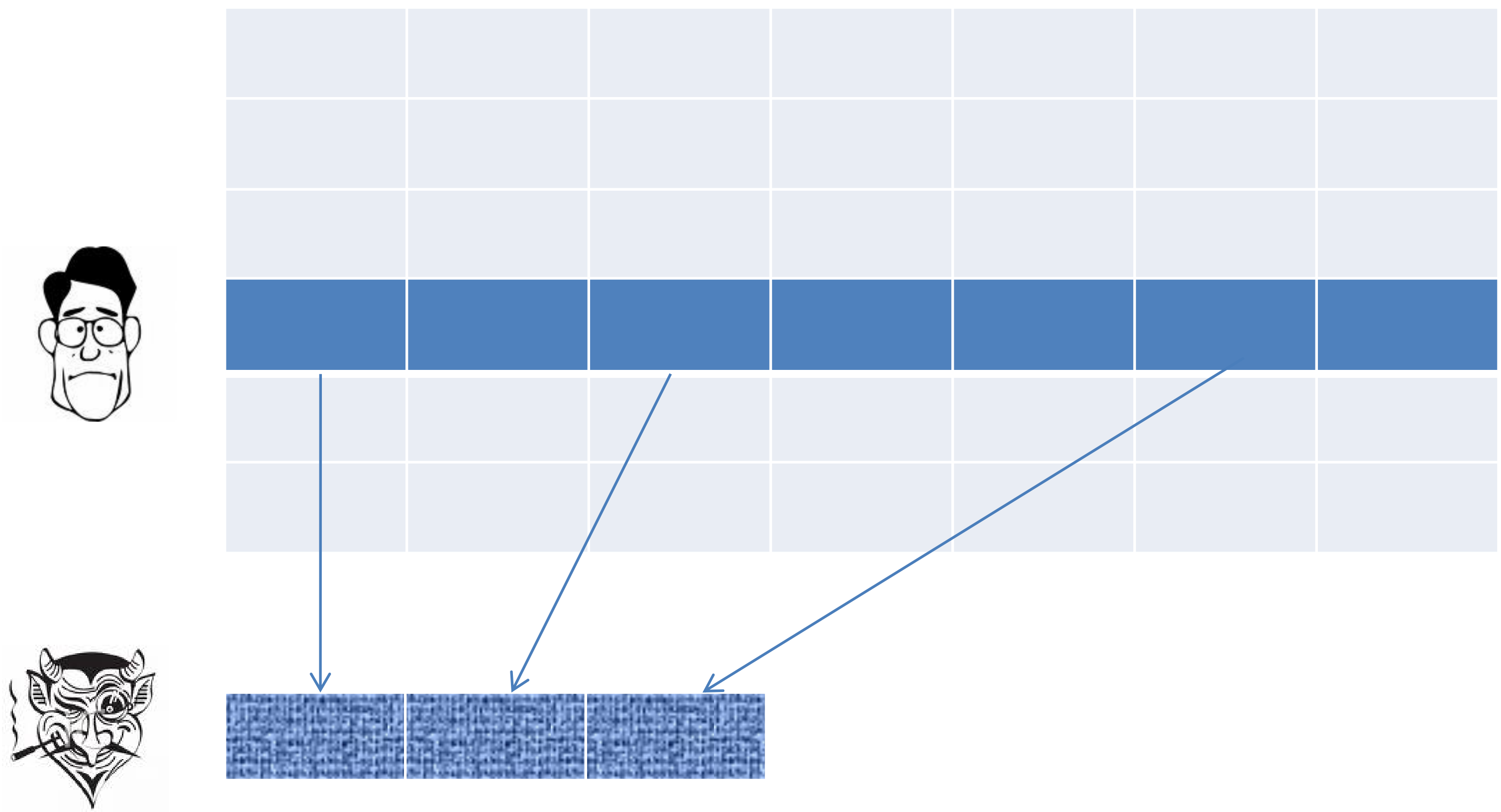
De-anonymization Objectives

- 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

De-anonymization Challenges

- Auxiliary information is noisy
  - Can't use standard information retrieval techniques
- Released records may be perturbed
- Only a sample of records has been released
- False matches
  - No oracle to confirm success!

Aux as Noisy Projection



What De-anonymization Is Not

- Not linkage (statistics, Census studies)
- Not search (information retrieval)
- Not classification (machine learning)
- Not fingerprinting (forensics)

"Scoreboard" Algorithm

- Scoring function
  - Assigns a score to each record in the released sample based on how well it matches Aux
    - $\Sigma_{i \in supp(aux)}$ Similarity(aux_i, r_i) / log(|support(i)|) gives higher weight to rarer attributes

Intuition: weight is a measure of entropy

- Record selection
  - Use "eccentricity" of the match to separate true and spurious matches

Extremely versatile paradigm

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

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

Self-testing

Methodological question: how does the attacker know the matches aren't spurious?

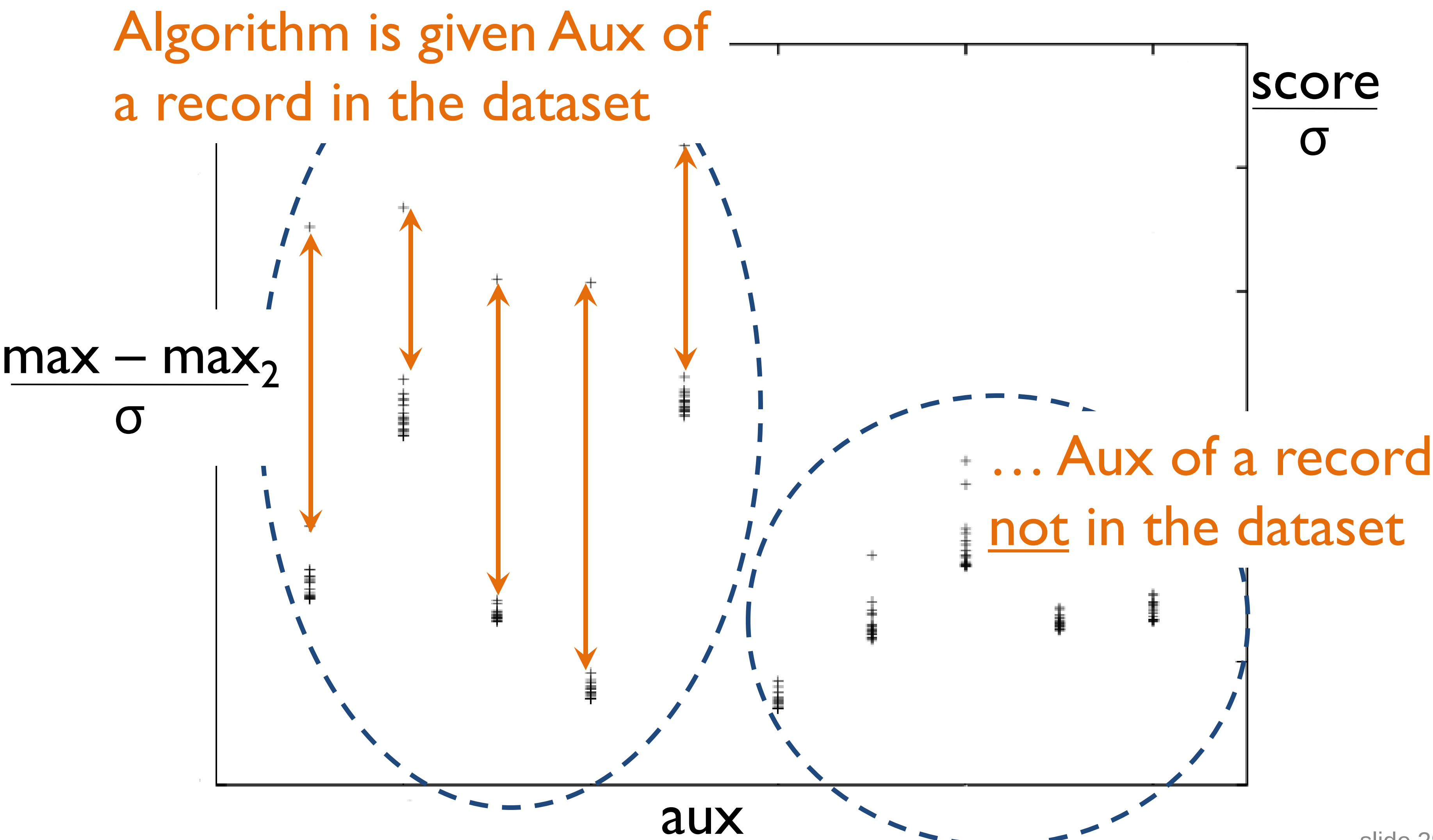
- No de-anonymization oracle or "ground truth"
- Compute a score for each record: how well does it match the auxiliary information?
- Heuristic: $(\max - \max_2) / \sigma \geq \phi$

Best score

Second-best score

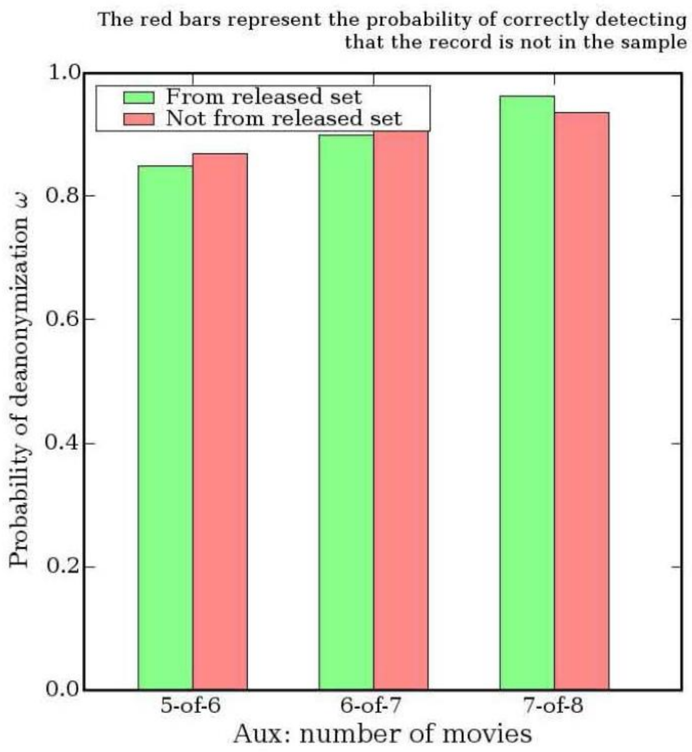
Eccentricity threshold

Eccentricity in the Netflix Dataset



Self-testing: Experimental Results

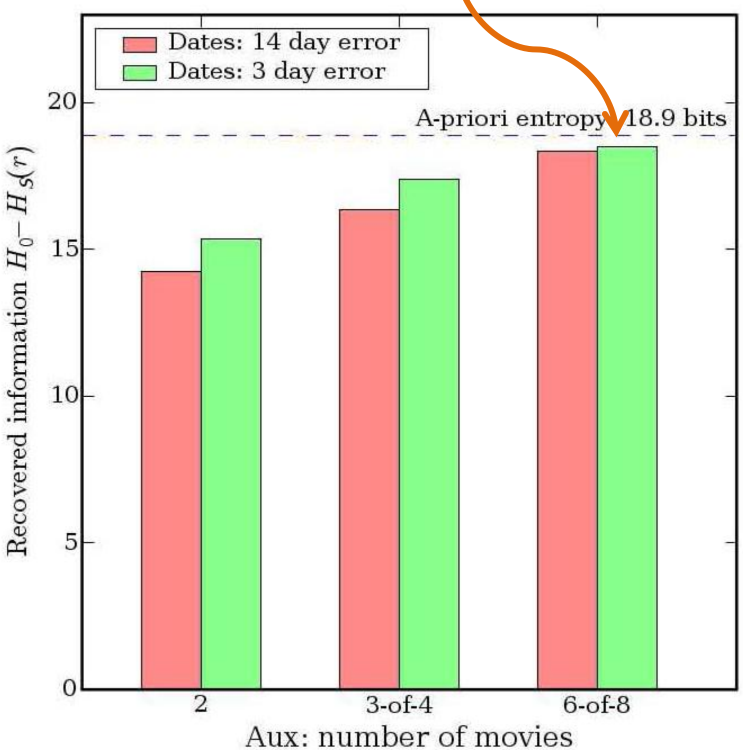
- After algorithm finds a match, remove the found record and re-run
- With very high probability, the algorithm now declares that there is no match



Robustness

With 6 approximately correct & 2 completely wrong ratings, recover all entropy

- Algorithm is robust to errors in attacker's Aux
  - Dates and ratings may be known imprecisely, some may be completely wrong
  - Perturbation = noise in the data = doesn't matter!
  - Nearest neighbor is so far, can tolerate <u>huge</u> amount of noise and perturbation



Main Themes

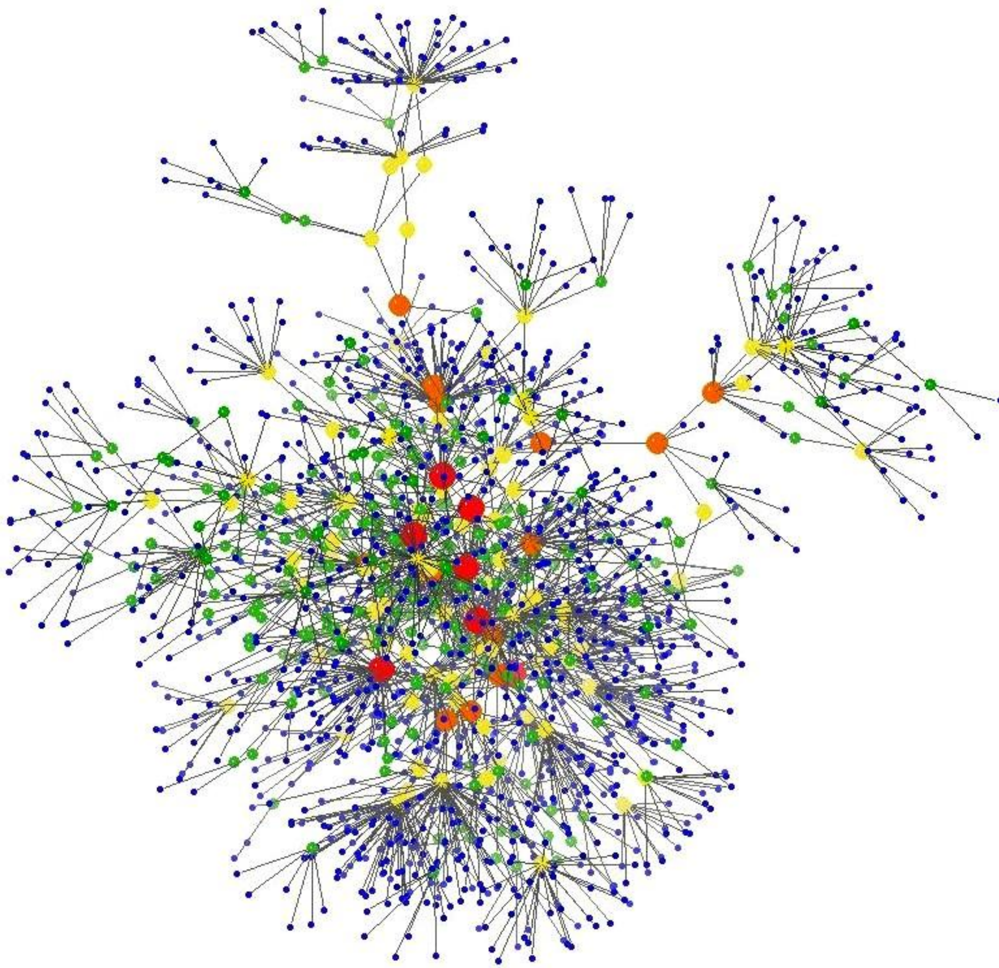
- Conceptual
  - Datasets are sparse
    - No "nearest neighbors"
  - Aux is logarithmic in number of records, linear in noise
  - "Personally identifiable" is meaningless
  - Distinction between aggregate and individual data unclear

Recommender systems

- Methodological
  - Scoring function to match records
  - Self-testing to avoid false matches
  - Self-correction leads to ever more accurate re-identification
  - Simple heuristics improve accuracy

Social networks

Exploiting Data Structure



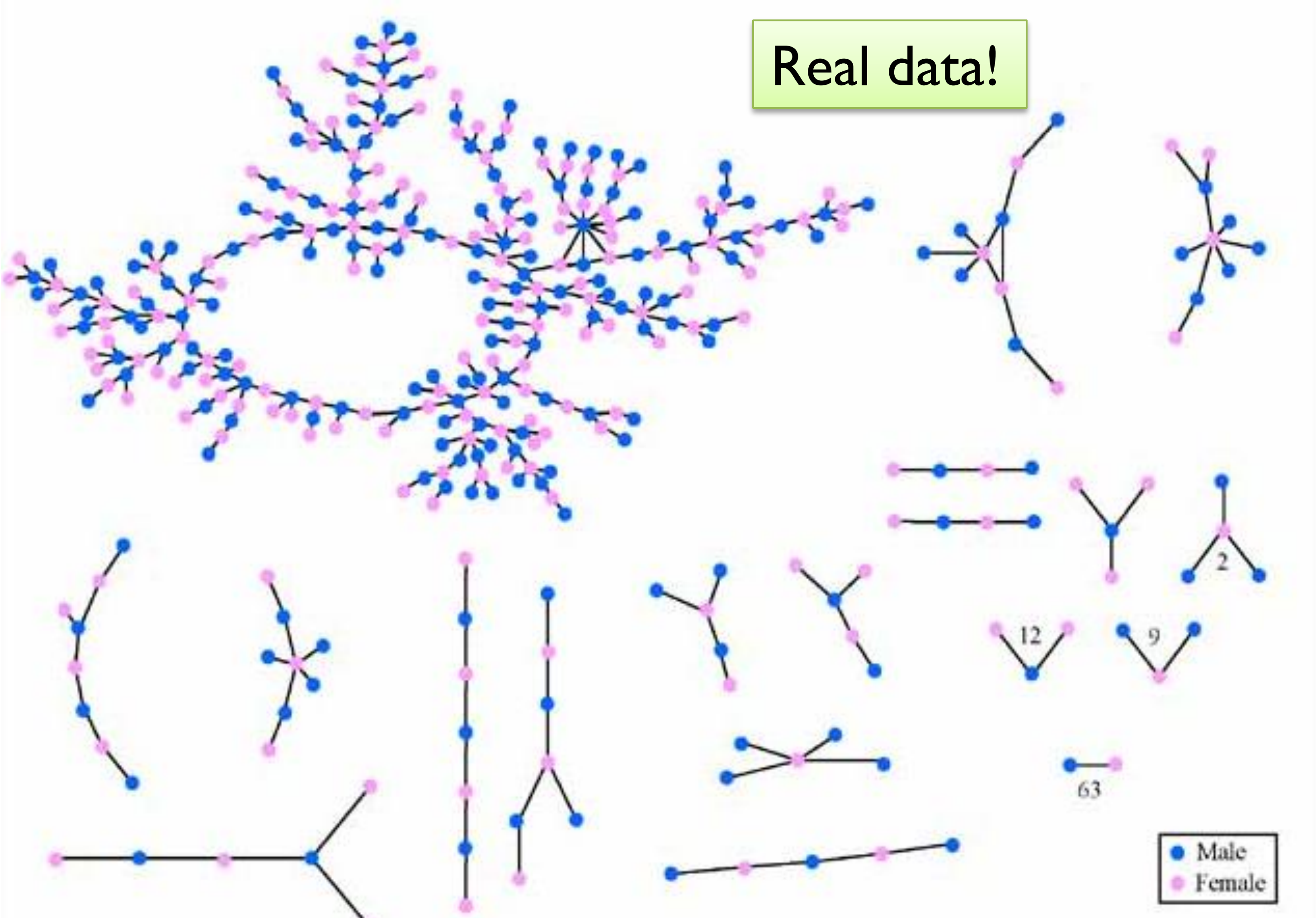
Reading Material

Backstrom, Dwork, Kleinberg
Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography
WWW 2007 and CACM 2011

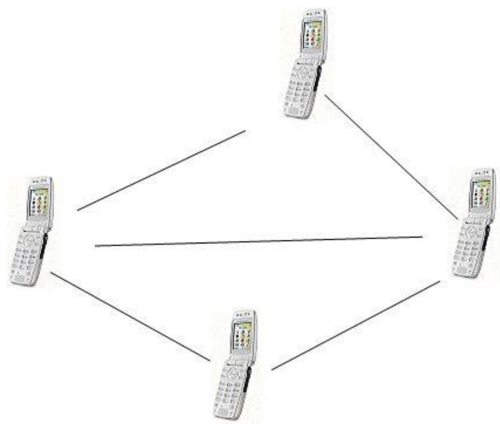
Narayanan and Shmatikov
De-anonymizing Social Networks
Oakland 2009

Narayanan, Shi, Rubinstein
Link Prediction by De-anonymization:
How We Won the Kaggle Social Network Challenge
IJCNN 2011

"Jefferson High": Romantic and Sexual Network



Phone Call Graphs

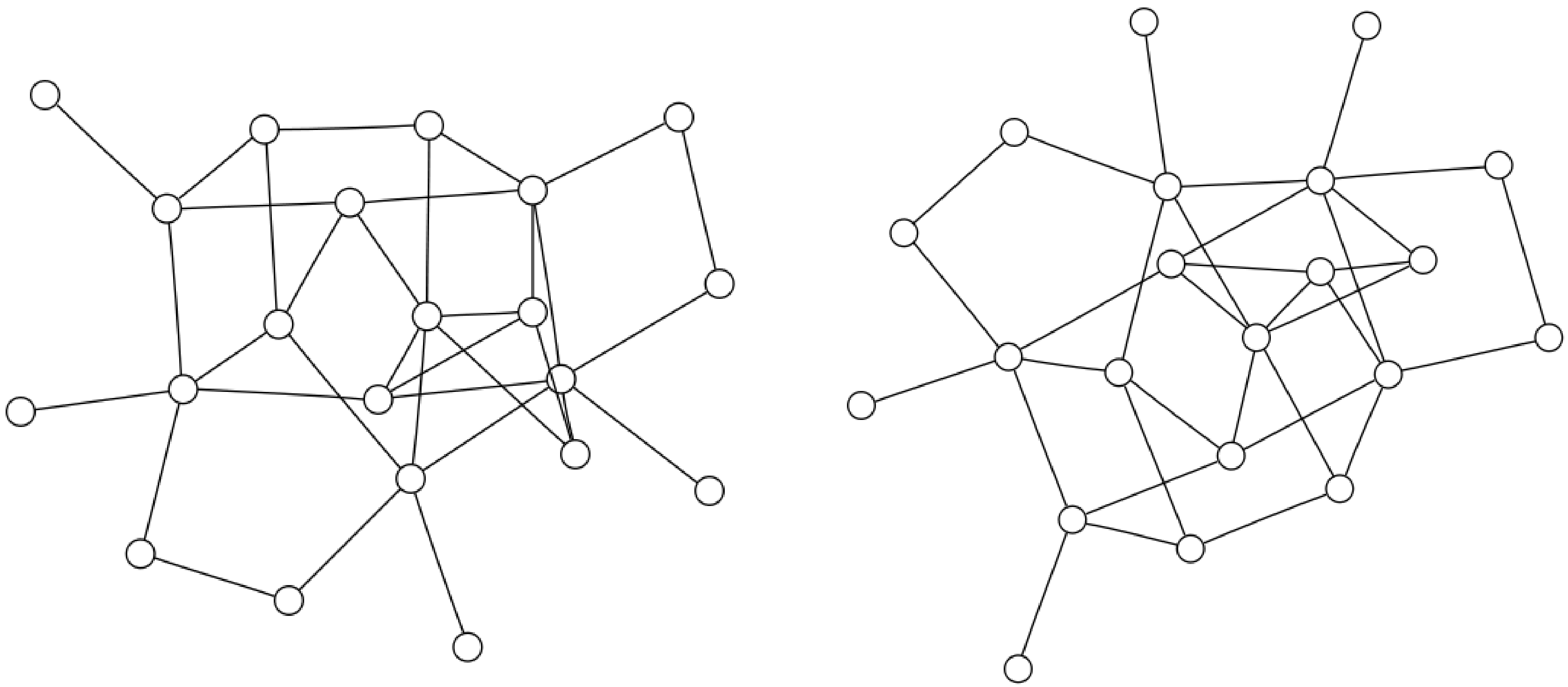


2 **trillion** edges

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



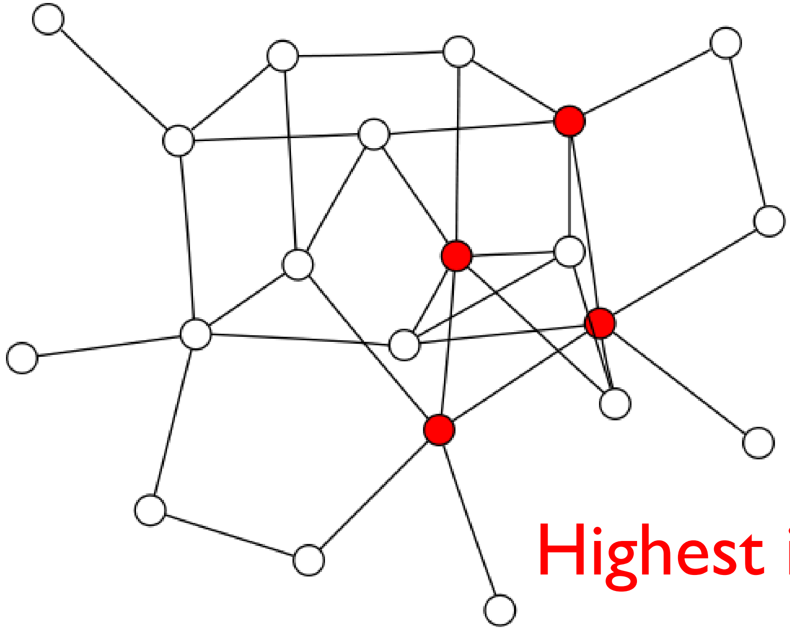
Goal: structural mapping between two graphs

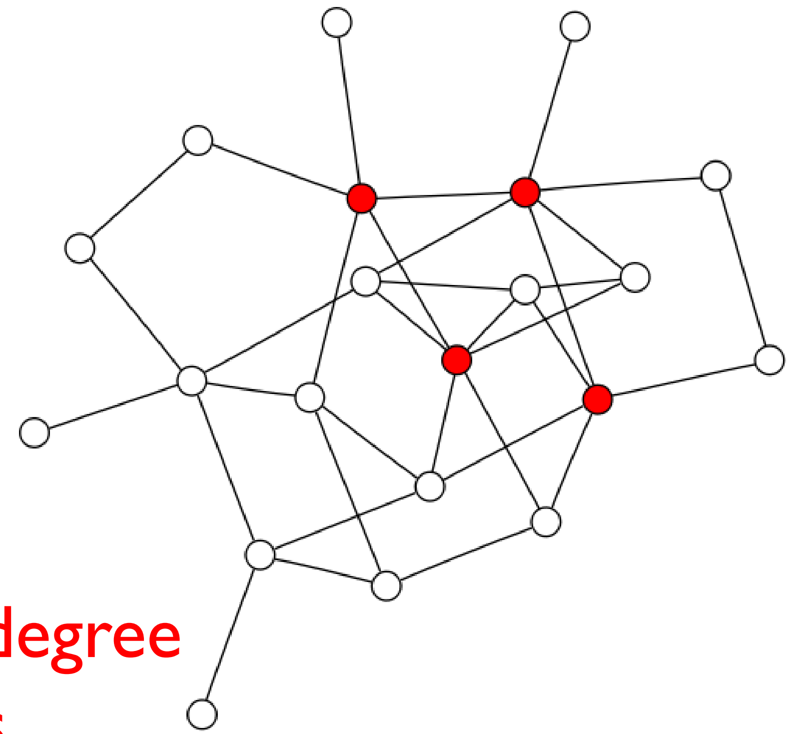
For example, Facebook vs. anonymized phone call graph

Two-Stage Paradigm

- Seed matching
  - Detailed knowledge about a small number of nodes
  - Used to create initial "seed" mapping between auxiliary information and anonymized graph
- Propagation
  - Iteratively extend the mapping using already mapped nodes
  - Self-reinforcing (similar to "spread of epidemic")

Where To Start?

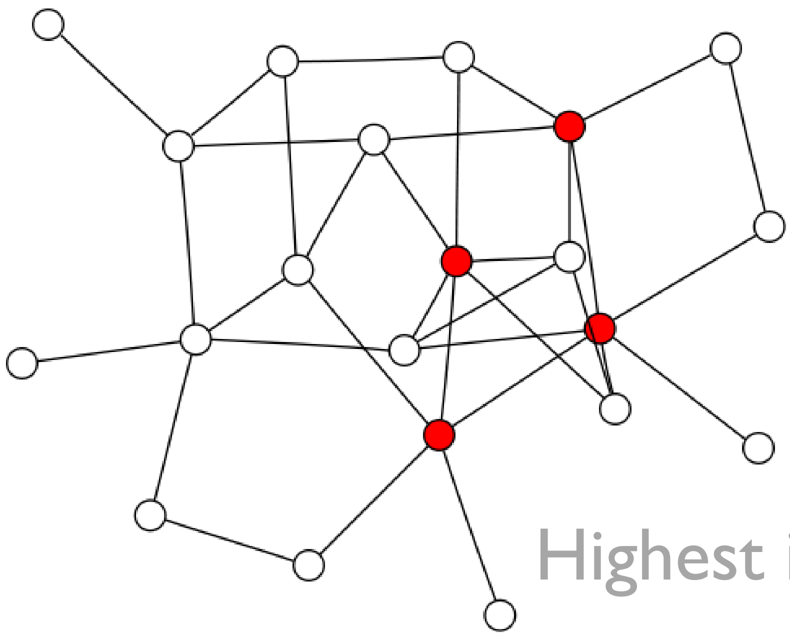




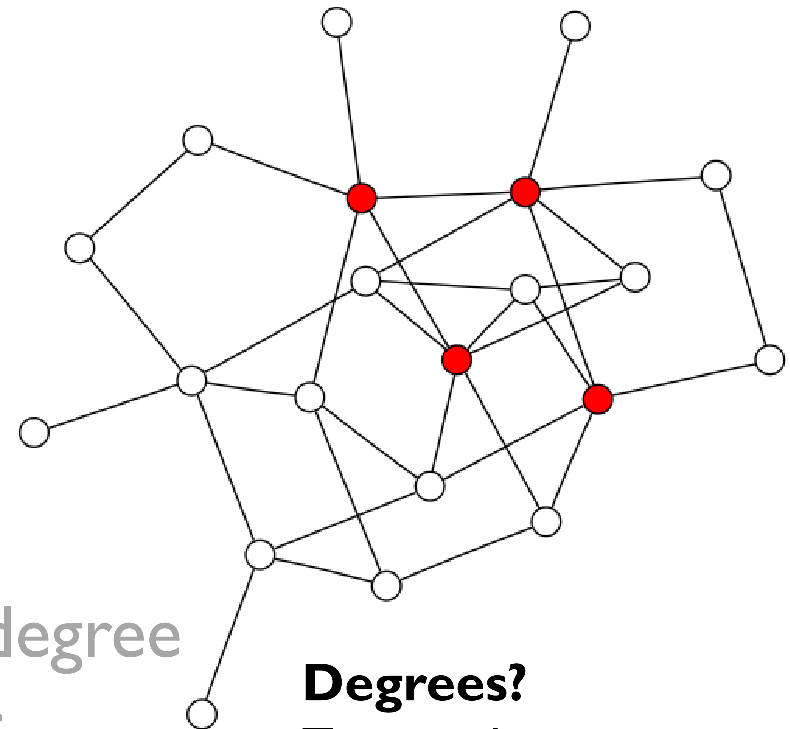
Highest in-degree nodes

Only a subset of nodes and edges in common

How To Match?



Highest in-degree nodes



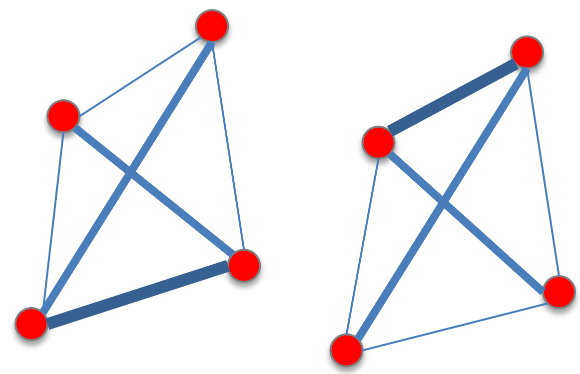
Degrees?
Too much variation

Subgraph structure?
Too sparse

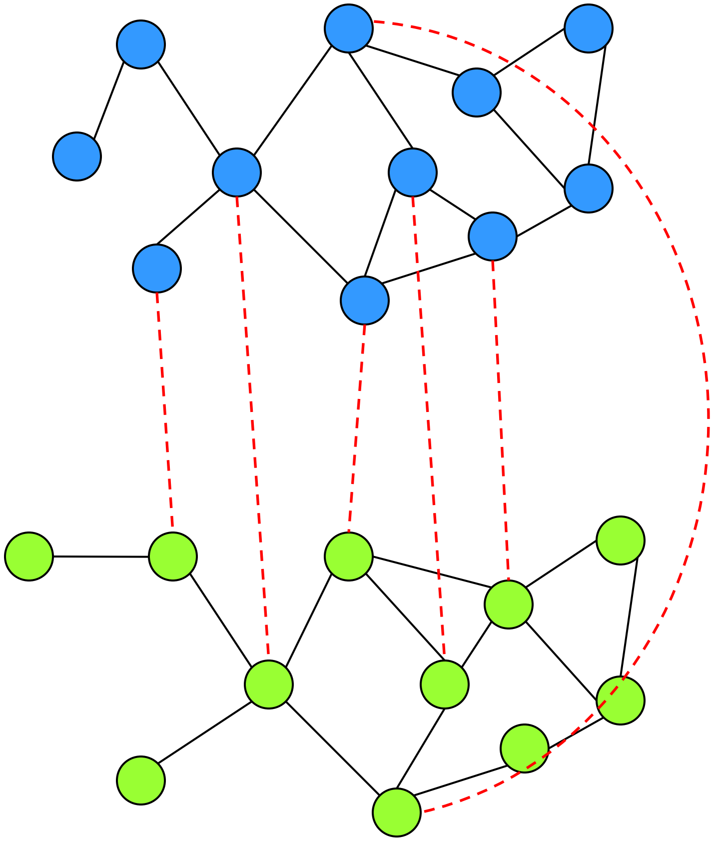
Number of **common neighbors** between each pair of nodes

Seed Matching as Combinatorial Optimization

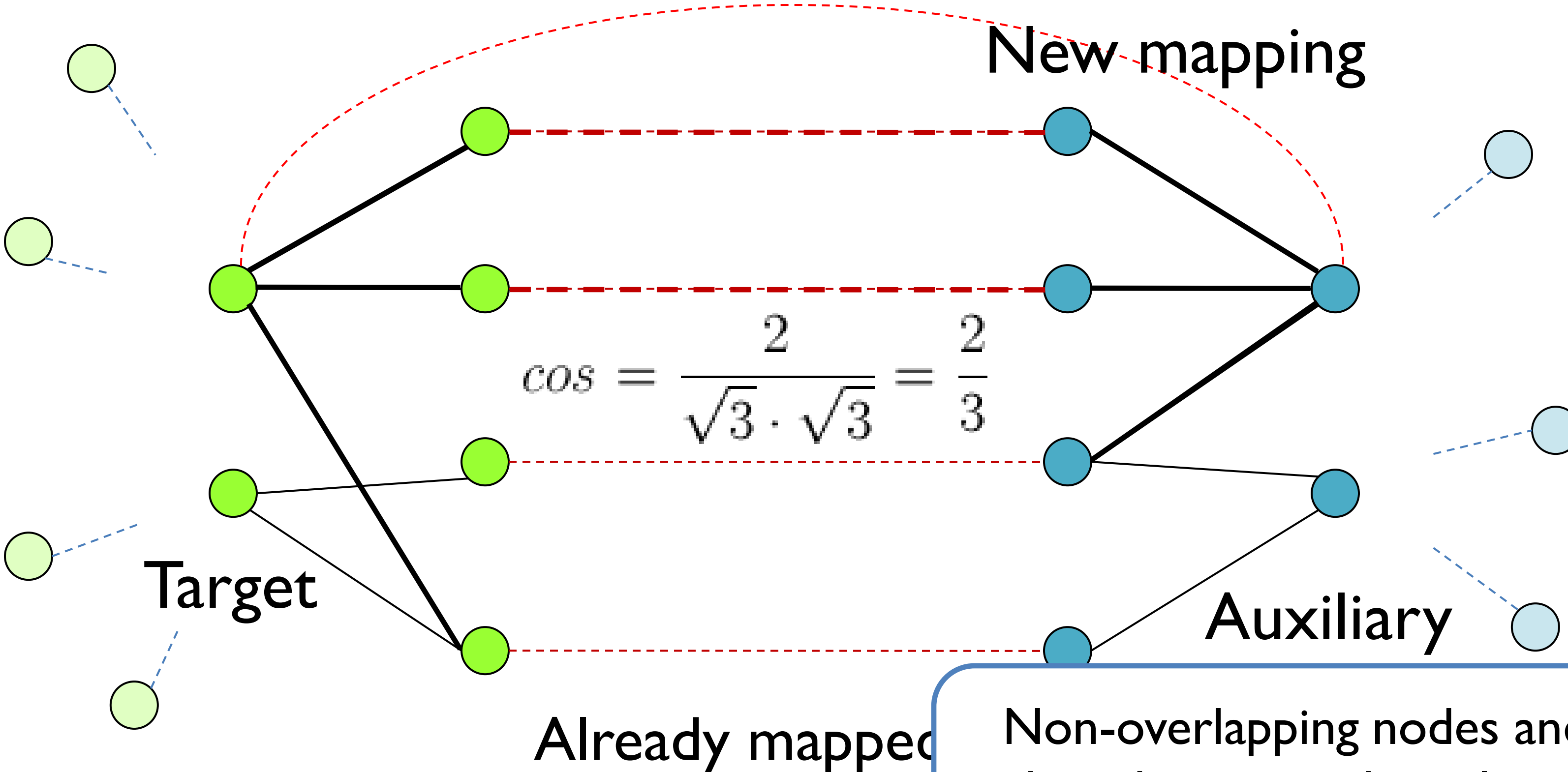
- Complete graphs on 20 – 100 “seed” nodes
- Edge weights = common neighbor coefficients (cosines)
- Reduced to known problem: weighted graph matching – use simulated annealing
- Now we have a mapping between seed nodes



Iterative Propagation



Propagation: Measuring Similarity



$$cos = \frac{2}{\sqrt{3} \cdot \sqrt{3}} = \frac{2}{3}$$

Non-overlapping nodes and edges due to graph evolution, data perturbation, etc.

Problem: dealing with noise

Adaptations To Handle Noise

Reverse map

Edge directionality

Edge weights

Node weights

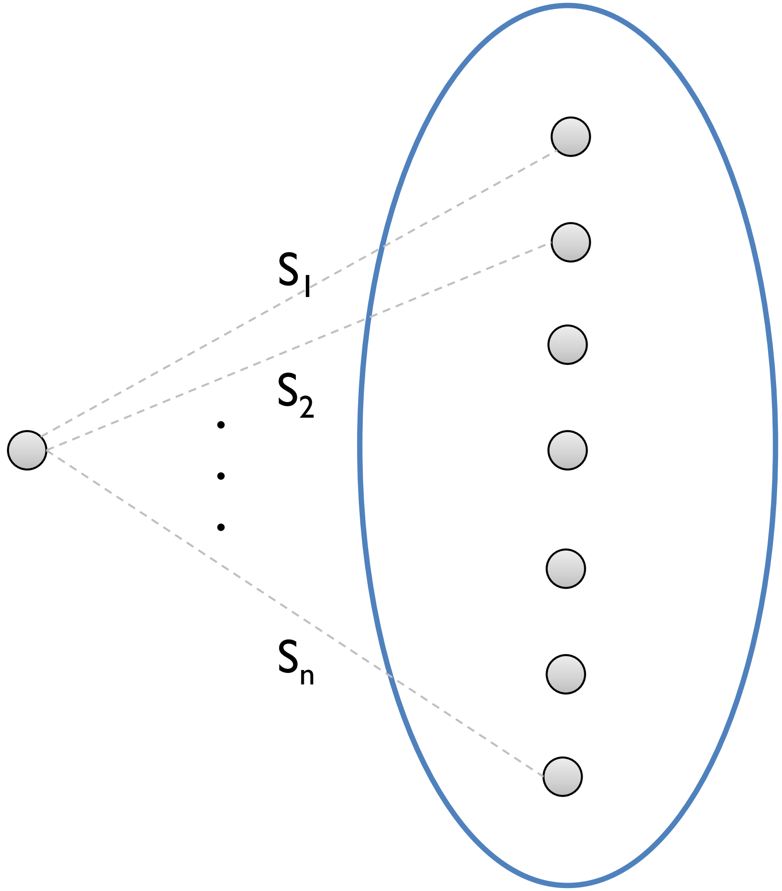
Self-correction

Eccentricity

Non-bijective

Deletion

Eccentricity



If true positive:

- $s_{max} - s_{max2}$ is large

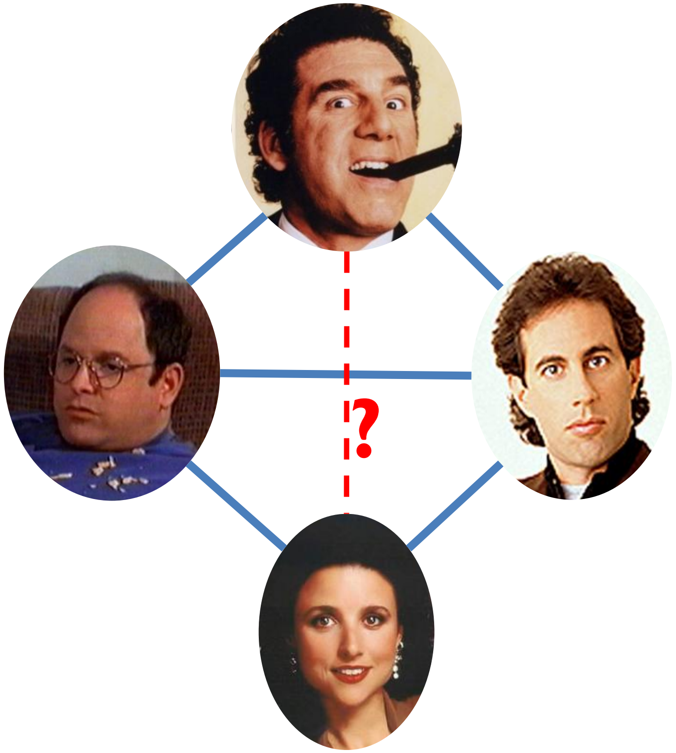
If false positive:

- $s_{max} - s_{max2}$ is small

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



Other De-anonymization Results

- Social networks – again and again
- Location data
- Stylometry (writing style)

…

- Genetic data
  - Same general approach
  - Different data models, algorithms, scaling challenges

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”