Anonymization and Re-Identification

Vitaly Shmatikov

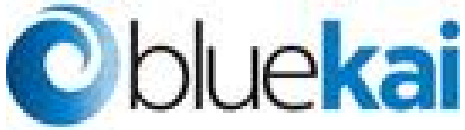

KEEP
CALM
YOU'RE
BEING
WATCHED

Tastes and Purchases

Social Networks

Health Care and Genetics

Web Tracking

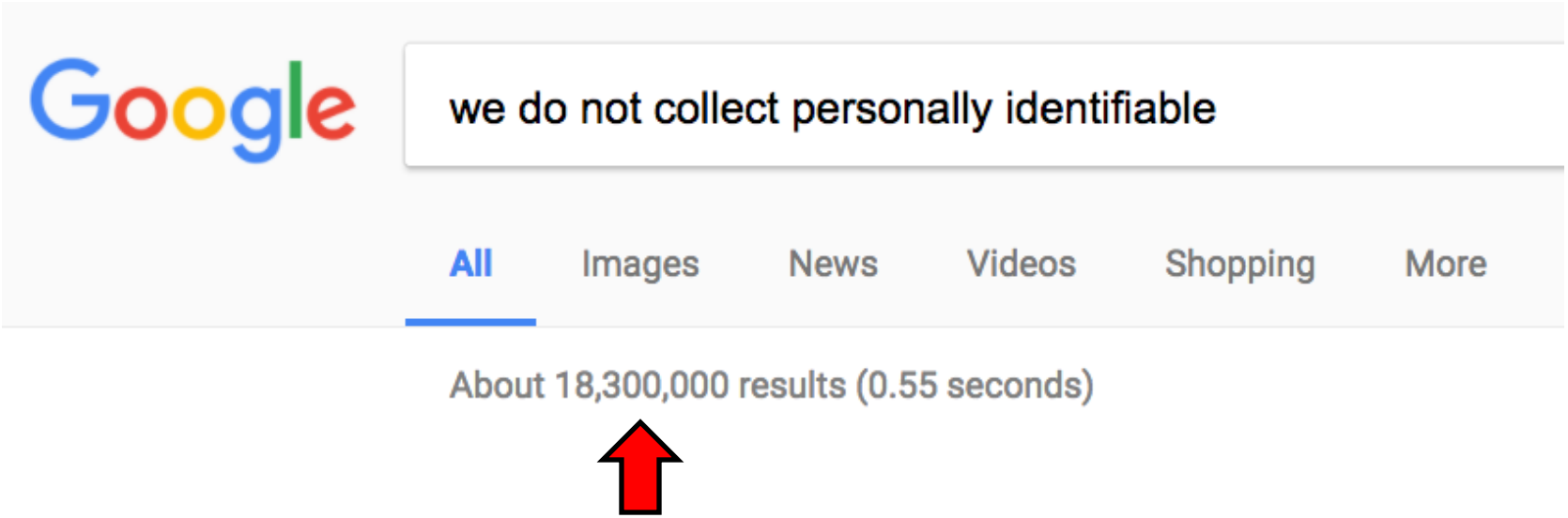


Online-Offline Aggregation

Solution: Anonymity!

"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 identifiable
All
Images
News
Videos
Shopping
More
About 18,300,000 results (0.55 seconds)

National Institute of
Standards and Technology
U.S. Department of Commerce

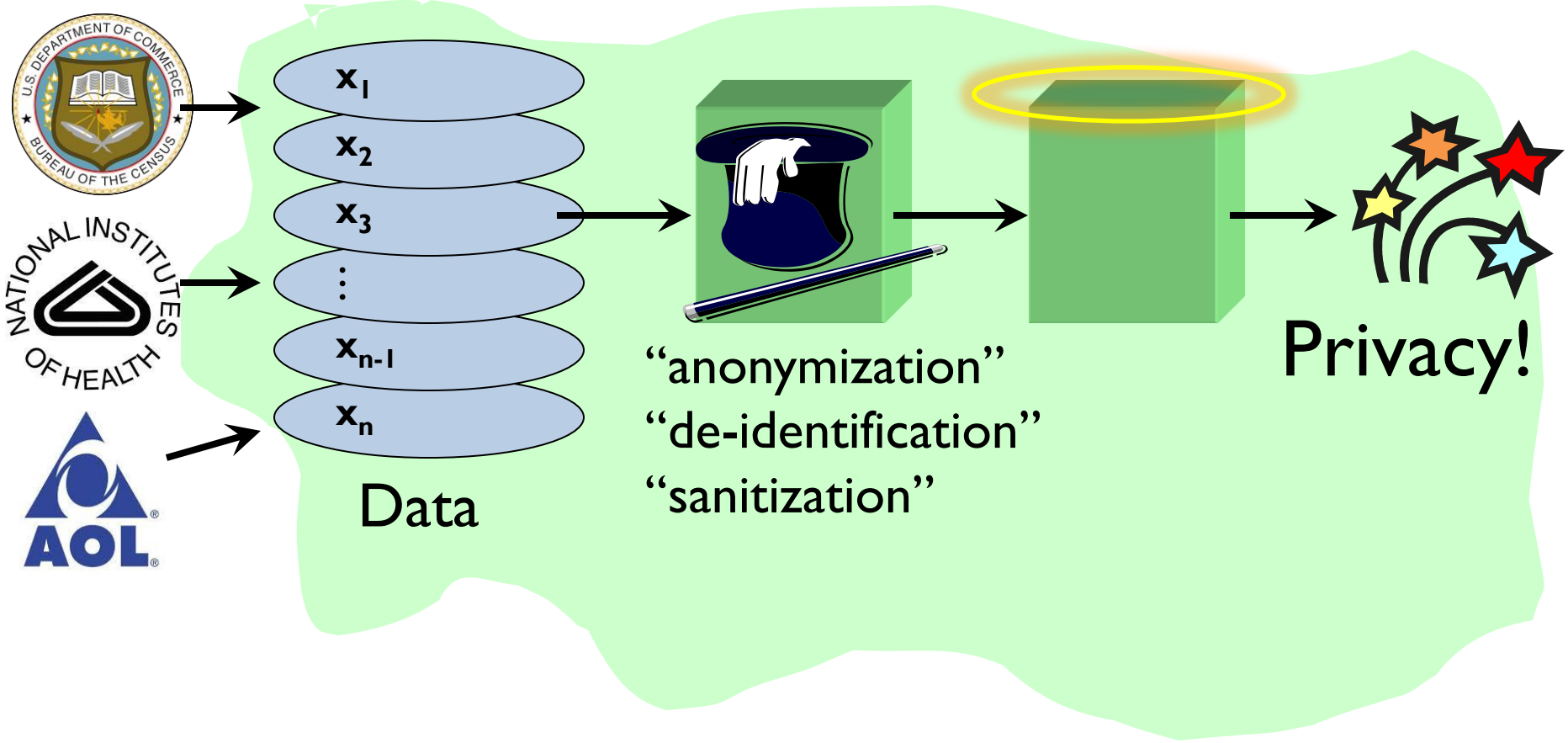
Special Publication 800-122

Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)

Recommendations of the National Institute of Standards and Technology

Erika McCallister
Tim Grance

“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

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

Reading Material

Microdata

ID	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47677	29	F	Ovarian Cancer
Betty	47602	22	F	Ovarian Cancer
Charles	47678	27	M	Prostate Cancer
David	47905	43	M	Flu
Emily	47909	52	F	Heart Disease
Fred	47906	47	M	Heart Disease

Voter registration data

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

Latanya Sweeney's Attack (1997)

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	...nicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
			09/27/64	female	02139	divorced	hypertension
			09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	.....
.....	.....	.....	.....	.....	.....	.....	
.....	.....	.....	.....	.....	.....	.....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	.....
.....	.....	.....	.....	.....	.....	.....	

Figure 1: ...e-identifying anonymous data by linking to external data

Public voter dataset

Quasi-Identifiers

- Key attributes
  - Name, address, phone number - uniquely identifying!
  - Always remove before release
- Quasi-identifiers
  - (5-digit ZIP code, birth date, gender) uniquely identify 87% of the population in the U.S.
  - Can be used for linking anonymized datasets with other datasets

Identifiers vs. Sensitive Attributes

- Sensitive attributes
  - Medical records, salaries, etc.
  - These attributes is what the researchers need, so they are released unmodified

Key Attribute		Quasi-identifier		Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

K-Anonymity: Intuition

- The information for each person contained in the released table cannot be distinguished from at least k-1 individuals whose information also appears in the release
  - Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender.
- Any quasi-identifier present in the released table must appear in at least k records

K-Anonymity Protection Model

- Private table → Released table RT
- Attributes: $A_1, A_2, \ldots, A_n$
- Quasi-identifier subset: $A_i, \ldots, A_j$

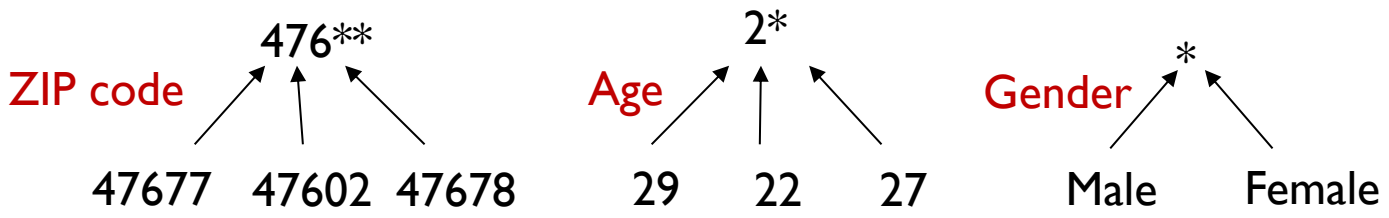
Let RT(A_1,...,A_n) be a table, QI_{RT} =(A_i,..., A_j) be the quasi-identifier associated with RT, A_i,...,A_j $\subseteq$ A_1,...,A_n, and RT satisfy k-anonymity. Then, each sequence of values in RT[A_x] appears with at least k occurrences in RT[QI_{RT}] for x=i,...,j.

Goal: each record is indistinguishable from at least k-1 other records ("equivalence class")

Achieving k-Anonymity

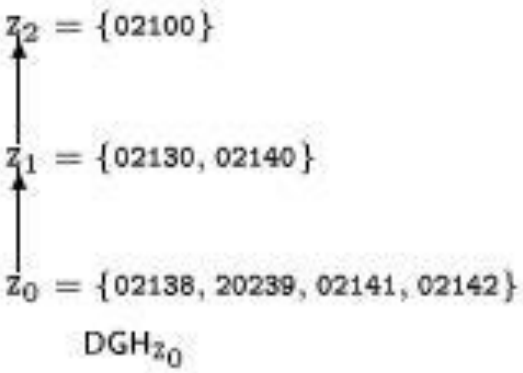
Lots of algorithms in the literature aiming to produce "useful" anonymizations, usually without any clear notion of utility

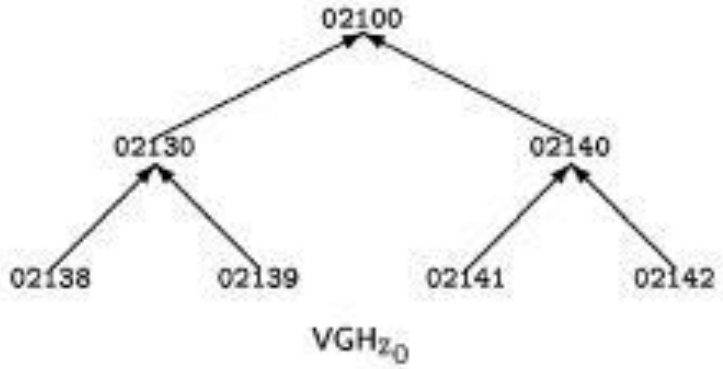
- Generalization
  - Replace quasi-identifiers with less specific but semantically consistent values until get k identical
  - Partition ordered-value domains into intervals

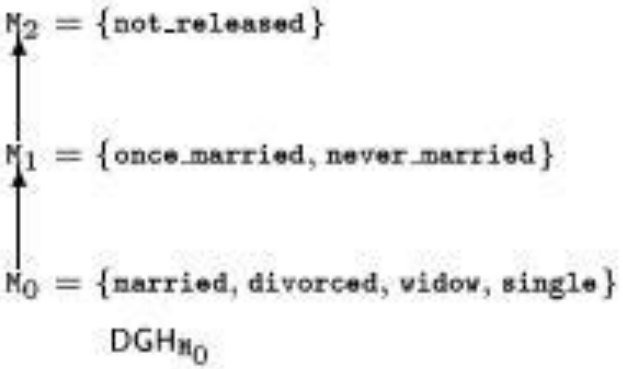


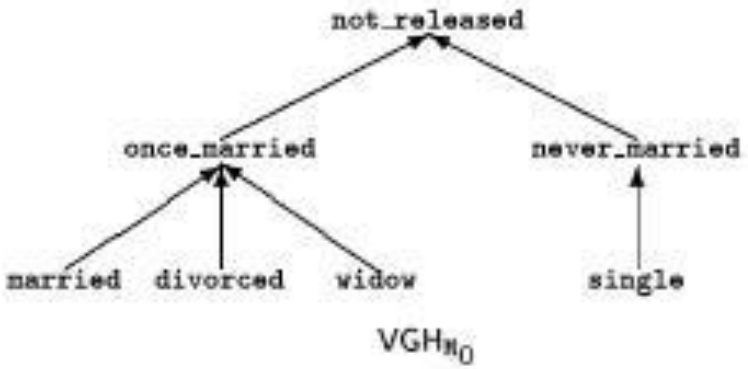
- Suppression
  - When generalization causes too much information loss (this often happens with "outliers")

Generalization in Action









Example of a k-Anonymous Table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2 Example of *k*-anonymity, where *k*=2 and QI={*Race*, *Birth*, *Gender*, *ZIP*}

Example of Generalization (1)

Released table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

External data source

Name	Birth	Gender	ZIP	Race
Andre	1964	m	02135	White
Beth	1964	f	55410	Black
Carol	1964	f	90210	White
Dan	1967	m	02174	White
Ellen	1968	f	02237	White

Figure 2 Example of *k*-anonymity, where *k*=2 and QI={*Race*, *Birth*, *Gender*, *ZIP*}

By linking these two tables, you still don't learn Andre's problem

Example of Generalization (2)

Microdata

QID			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

Generalized table

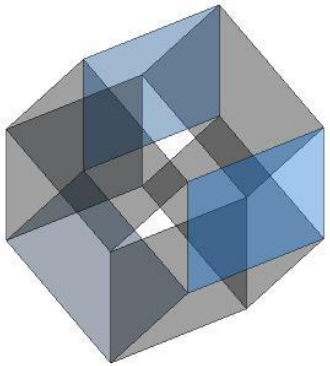
QID			SA
Zipcode	Age	Sex	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

!!

- Released table is 3-anonymous
- If the adversary knows Alice's quasi-identifier (47677, 29, F), he still does not know which of the first 3 records corresponds to Alice's record

Curse of Dimensionality

Aggarwal (VLDB 2005)



- Generalization fundamentally relies on spatial locality
  - Each record must have k close neighbors
- Real-world datasets are very sparse
  - Many attributes (dimensions)
    - Netflix Prize dataset: 17,000 dimensions
    - Amazon customer records: several million dimensions
  - "Nearest neighbor" is very far
- Projection to low dimensions loses all info ⇒ k-anonymized datasets are useless

What Does k-Anonymity Prevent?

- Membership disclosure: Attacker cannot tell that a given person is in the dataset.
- Sensitive attribute disclosure: Attacker cannot tell that a given person has a certain sensitive attribute.
- Identity disclosure: Attacker cannot tell which record corresponds to a given person.

This interpretation is correct, assuming the attacker does not know anything other than quasi-identifiers.
But this does not imply any privacy!

Unsorted Matching Attack

- Problem: records appear in the same order in the released table as in the original table
- Solution: randomize order before releasing

Race	ZIP
Asian	02138
Asian	02139
Asian	02141
Asian	02142
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race	ZIP
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

GT1

Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

GT2

Figure 3 Examples of *k*-anonymity tables based on PT

Complementary Release Attack

Ganta et al. (KDD 2008)

- Different releases of the same private table can be linked to compromise k-anonymity

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	0213*	short of breath
person	1965	female	0213*	hypertension
white	1964	male	0213*	obesity
white	1964	male	0213*	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

GT1

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
white	1960-69	human	02139	hypertension
white	1960-69	human	02139	obesity
white	1960-69	human	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

GT3

Linking Independent Releases

Race	BirthDate	Gender	ZIP	Problem
black	9/20/1965	male	02141	short of breath
black	2/14/1965	male	02141	chest pain
black	10/23/1965	female	02138	painful eye
black	8/24/1965	female	02138	wheezing
black	11/7/1964	female	02138	obesity
black	12/1/1964	female	02138	chest pain
white	10/23/1964	male	02138	short of breath
white	3/15/1965	female	02139	hypertension
white	8/13/1964	male	02139	obesity
white	5/5/1964	male	02139	fever
white	2/13/1967	male	02138	vomiting
white	3/21/1967	male	02138	back pain

PT

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	02138	short of breath
white	1965	female	02139	hypertension
white	1964	male	02139	obesity
white	1964	male	02139	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

LT

Exploiting Distributions

- k-Anonymity does not provide privacy if
  - Sensitive values in an equivalence class lack diversity
  - The attacker has background knowledge

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Carl	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

l-Diversity

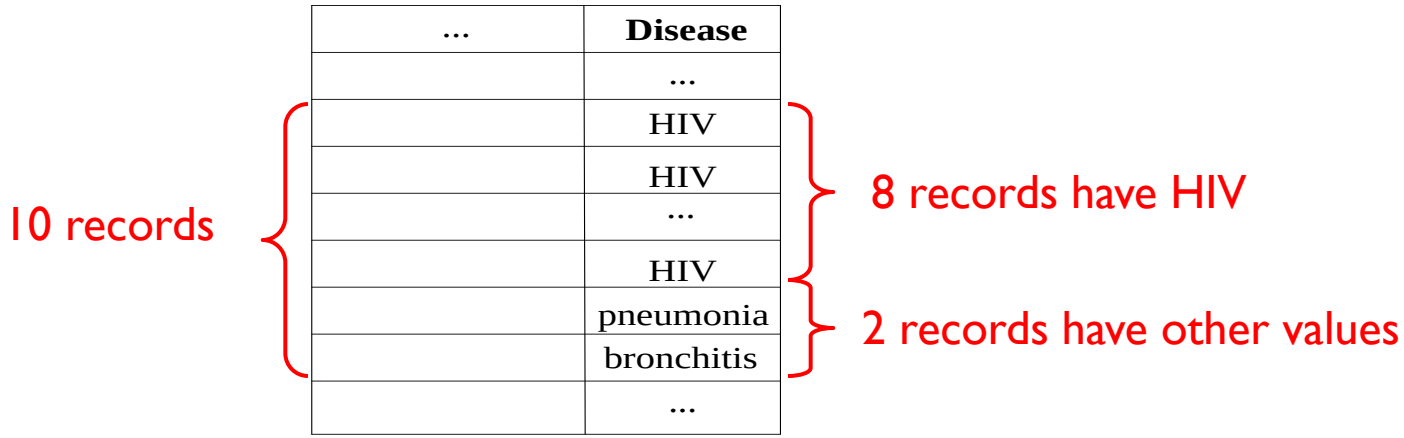
Machanavajjhala et al. (ICDE 2006)

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Sensitive attributes must be "diverse" within each quasi-identifier equivalence class

Distinct l-Diversity

- Each equivalence class has at least l well-represented sensitive values
- Doesn't prevent probabilistic inference attacks



Other Versions of l-Diversity

- Probabilistic l-diversity
  - The frequency of the most frequent value in an equivalence class is bounded by **1/l**
- Entropy l-diversity
  - The entropy of the distribution of sensitive values in each equivalence class is at least **log(l)**
- Recursive (c,l)-diversity
  - $\mathbf{r_1 < c(r_l + r_{l+1} + \ldots + r_m)}$ where r_i is the frequency of the i^{th} most frequent value
    - Most frequent value does not appear too frequently

My Favorite Charts

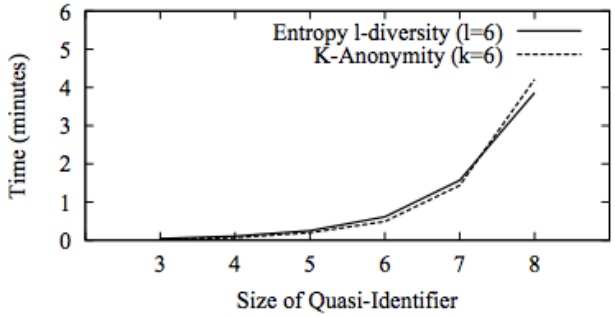


Figure 5. Adults Database

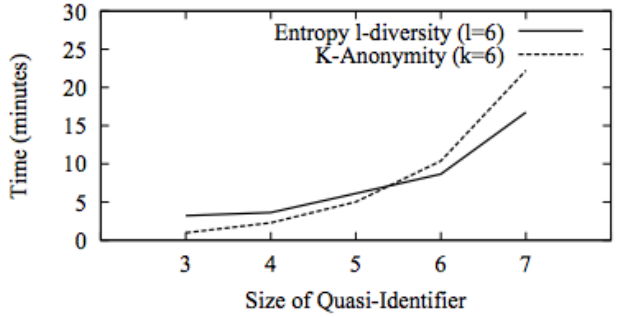


Figure 6. Lands End Database

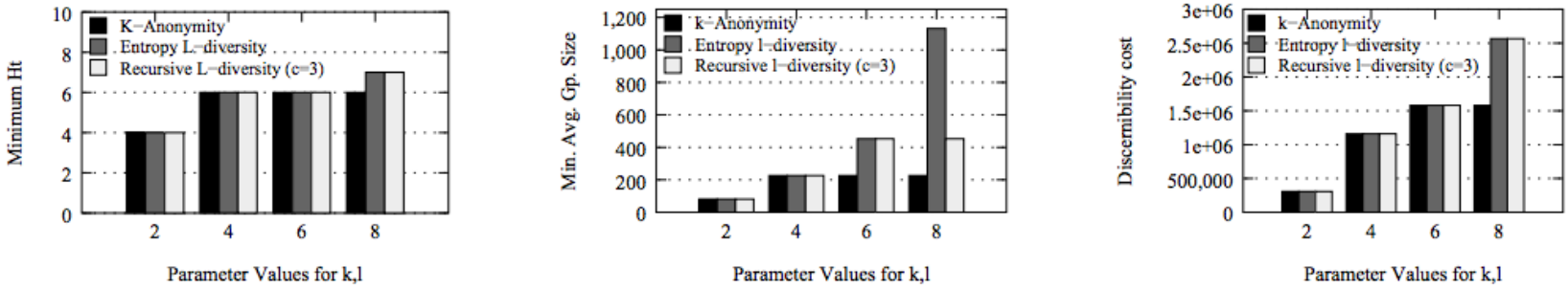


Figure 7. Adults Database. Q = {age, gender, race, marital_status}

Limitations of l-Diversity

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%) – very different sensitivity!
- l-diversity is unnecessary
  - 2-diversity is unnecessary for an equivalence class that contains only HIV- records
- l-diversity is difficult to achieve
  - Suppose there are 10000 records in total
  - To have distinct 2-diversity, there can be at most 10000*1%=100 equivalence classes

Skewness Attack

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
- Consider an equivalence class that contains an equal number of HIV+ and HIV- records
  - Diverse, but potentially violates privacy!
- l-diversity does not differentiate:
  - Equivalence class 1: 49 HIV+ and 1 HIV-
  - Equivalence class 2: 1 HIV+ and 49 HIV-

Does not consider overall distribution of sensitive values!

Sensitive Attribute Disclosure

Similarity attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

1. Bob's salary is in [20k,40k], which is relatively low
2. Bob has some stomach-related disease

Does not consider semantics of sensitive values!

t-Closeness

Li et al. (ICDE 2007)

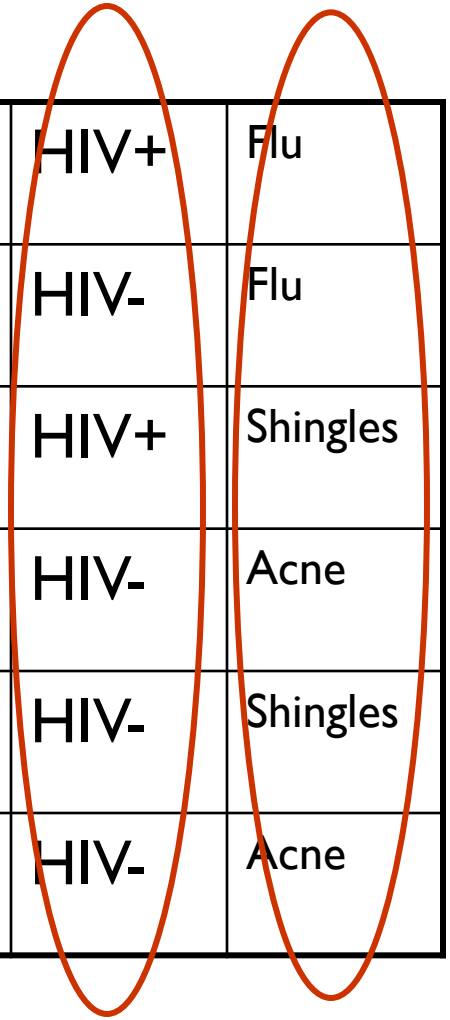
Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original database

Trick question: Why publish quasi-identifiers at all?

Anonymous, "t-Close" Dataset

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
Asian/AfrAm	787XX	HIV+	Shingles
Caucas	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne



This is k-anonymous, l-diverse and t-close…

…so secure, right?

What Does Attacker Know?

Bob is Caucasian and I heard he was admitted to hospital with flu…

This is against the rules! "flu" is not a quasi-identifier

Yes… and this is yet another problem with k-anonymity

Caucas	787X X	HIV+	Flu
Asian/AfrAm	787X	HIV-	Flu
		HIV+	Shingles
	787X	HIV-	Acne
		HIV-	Shingles
Caucas	787X X	HIV-	Acne

HIPAA Privacy Rule

"Under the safe harbor method, covered entities must remove all of a list of 18 enumerated identifiers and have no actual knowledge that the information remaining could be used, alone or in combination, to identify a subject of the information."

“The identifiers that must be removed include direct identifiers, such as name, street address, social security number, as well as other identifiers, such as birth date, admission and discharge dates, and five-digit zip code. The safe harbor requires removal of geographic subdivisions smaller than a State, except for the initial three digits of a zip code if the geographic unit formed by combining all zip codes with the same initial three digits contains more than 20,000 people. In addition, age, if less than 90, gender, ethnicity, and other demographic information not listed may remain in the information. The safe harbor is intended to provide covered entities with a simple, definitive method that does not require much judgment by the covered entity to determine if the information is adequately de-identified."

AOL Search Logs

- In August 2006, AOL released anonymized search query logs
  - 657K users, 20M queries over 3 months
- Opposing goals
  - Analyze data for research purposes, provide better services for users and advertisers
  - Protect privacy of AOL users
    - Government laws and regulations
    - Search queries may reveal income, evaluations, intentions to acquire goods and services, etc.

AOL User 4417749

- AOL query logs have the form

  <AnonID, Query, QueryTime, ItemRank, ClickURL (truncated URL)>

- Sample queries of user with AnonID 4417749:
  - "numb fingers", "60 single men", "dog that urinates on everything", "landscapers in Lilburn, GA", several people with the last name Arnold
    - Only 14 citizens with the last name Arnold near Lilburn
  - NYT contacted the 14 citizens, found out AOL User 4417749 is 62-year-old Thelma Arnold

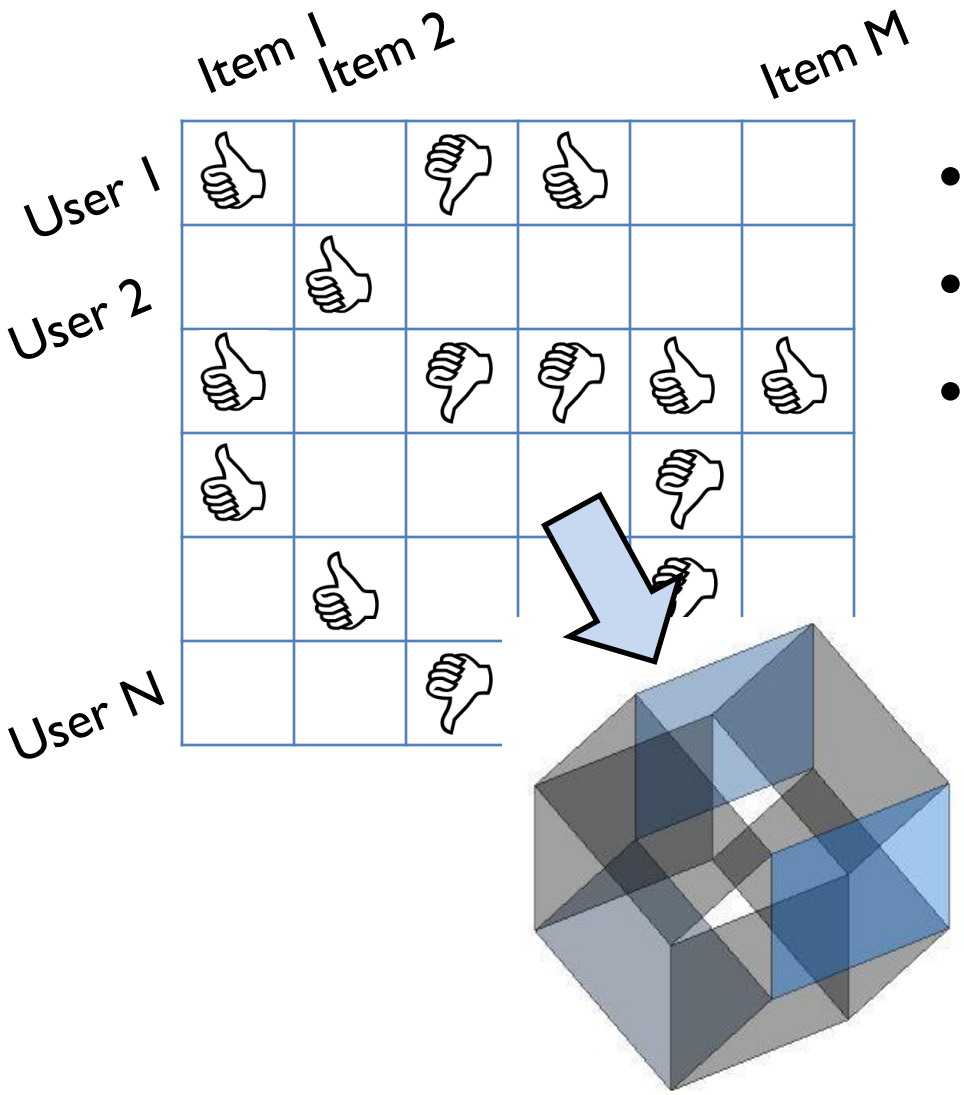
Anonymization Considered Harmful

- Syntactic
  - Focuses on data transformation, not on what can be learned from the anonymized dataset
  - Anonymized dataset can leak sensitive info
- "Quasi-identifier" fallacy
  - Assumes a priori that attacker will not know certain information about his target
- Relies on locality
  - Destroys utility of many real-world datasets

The Myth of the PII

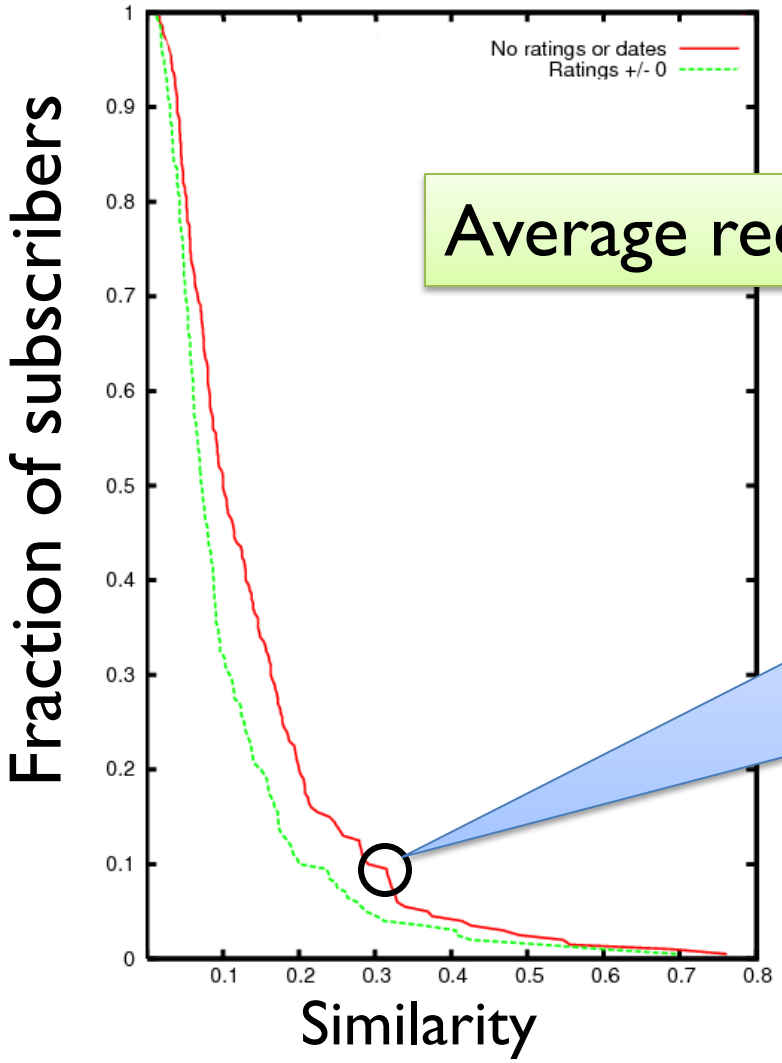
- Data are "anonymized" by removing personally identifying information (PII)
  - Name, Social Security number, phone number, email, address… what else?
- Problem: PII has no technical meaning
  - Defined in disclosure notification laws (if certain information is lost, consumer must be notified)
  - In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

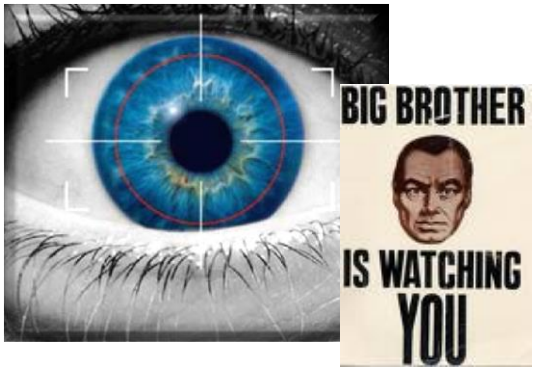


Average record has no "similar" records

Netflix Prize dataset:

Considering just movie names, for 90% of records there isn't a single other record which is more than 30% similar

Privacy Threats



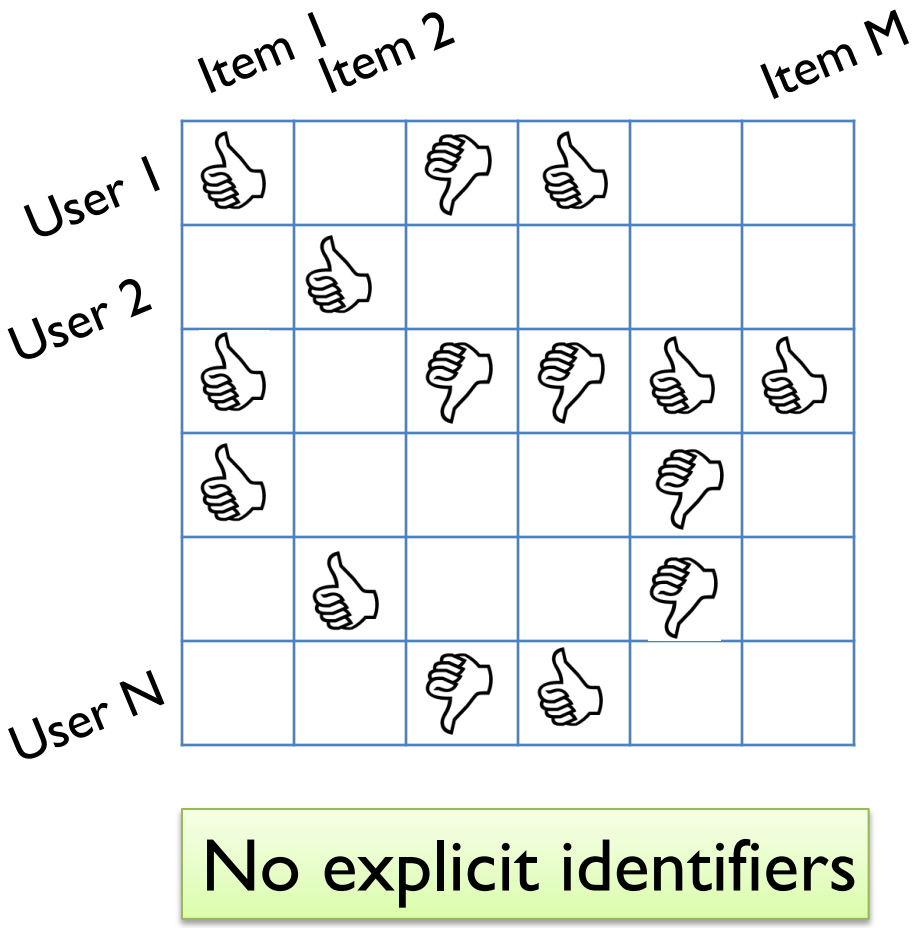
Global surveillance

Spammers
Abusive advertisers and marketers

Phishing

Employers, insurers,
stalkers, nosy friends

It's All About the Aux

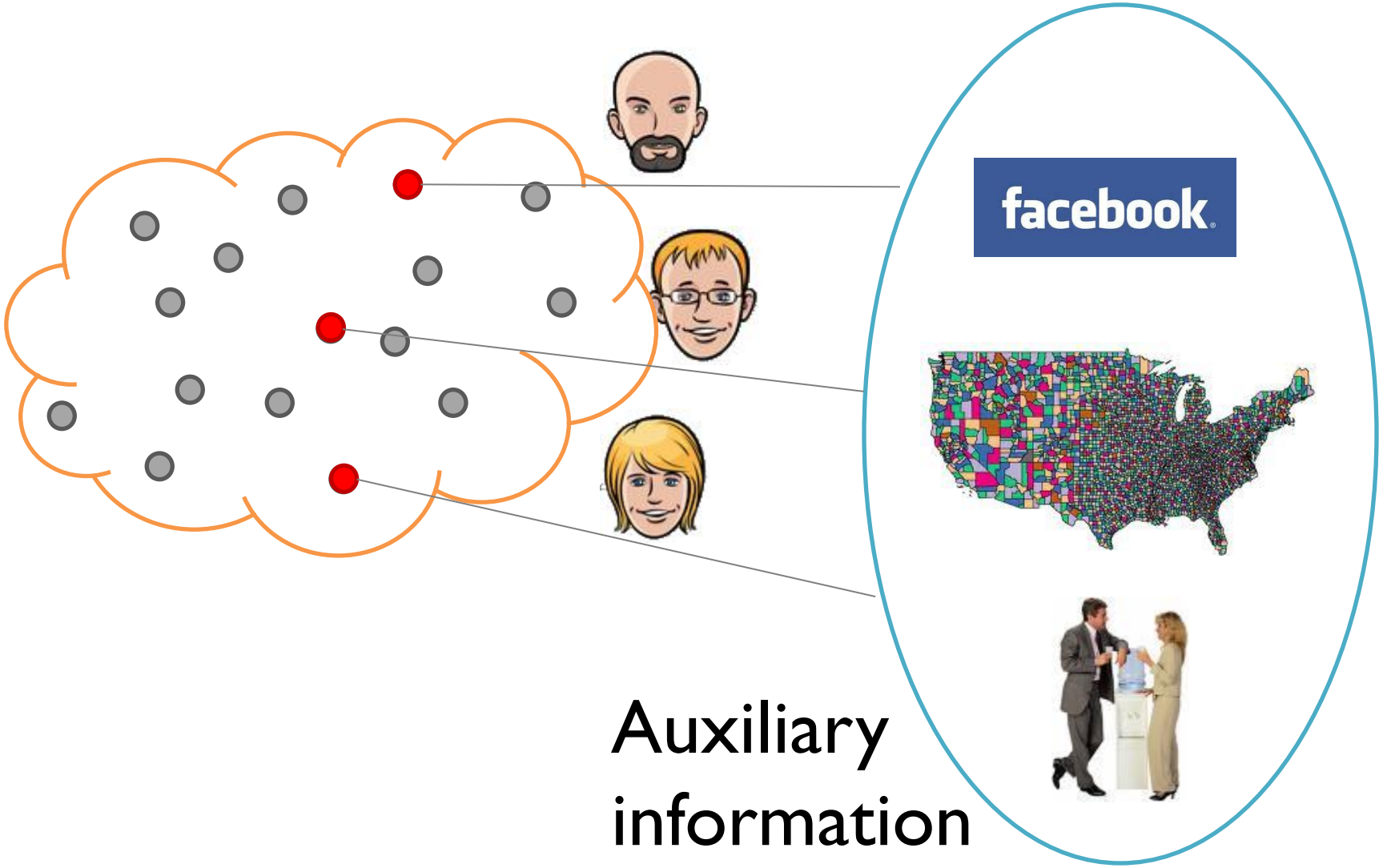


No explicit identifiers

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



Auxiliary information

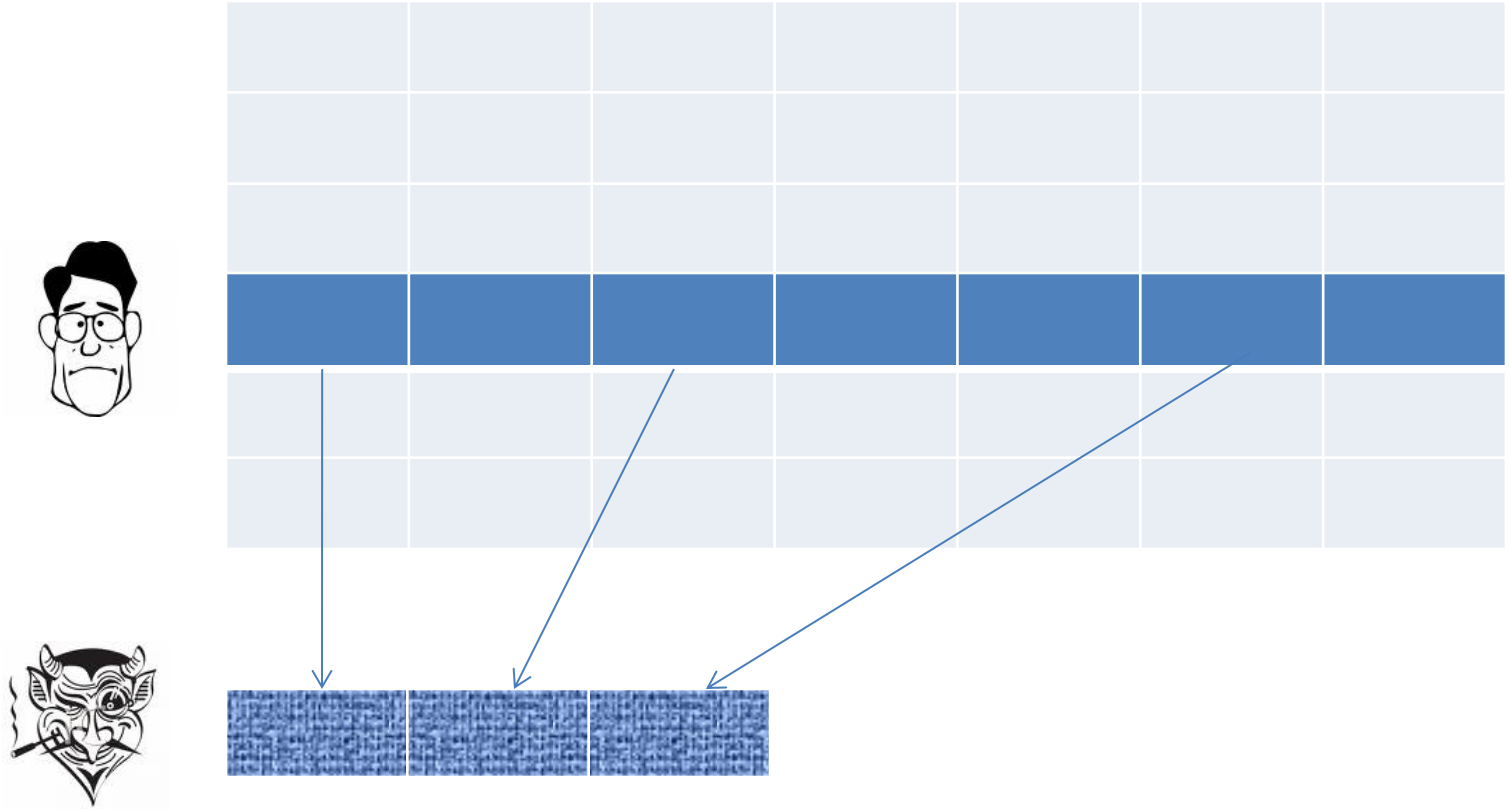
De-anonymization Objectives

- 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 Prize: Home

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

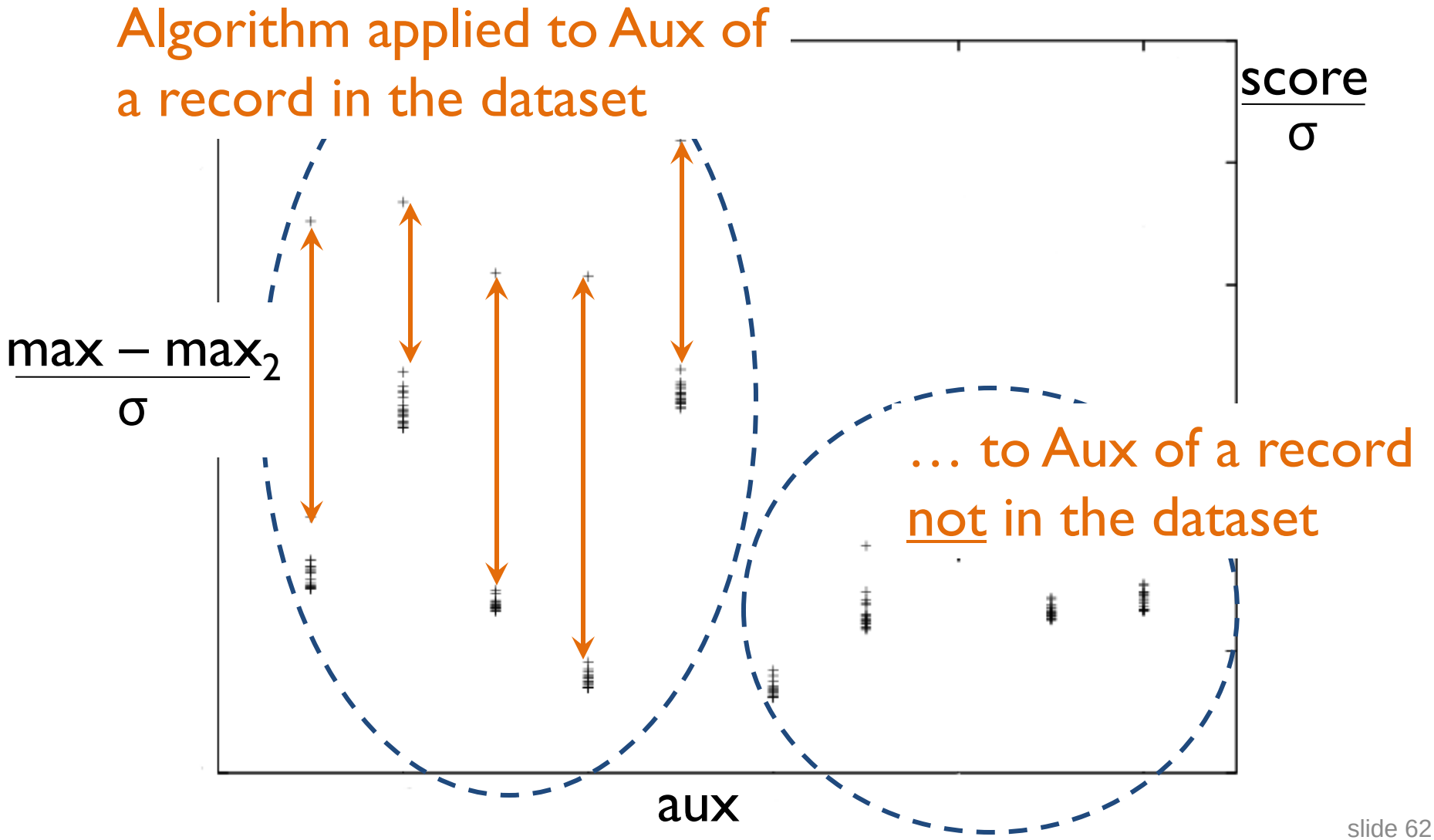
Self-testing

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: $(\text{max-max}_2) / \sigma \geq \phi$

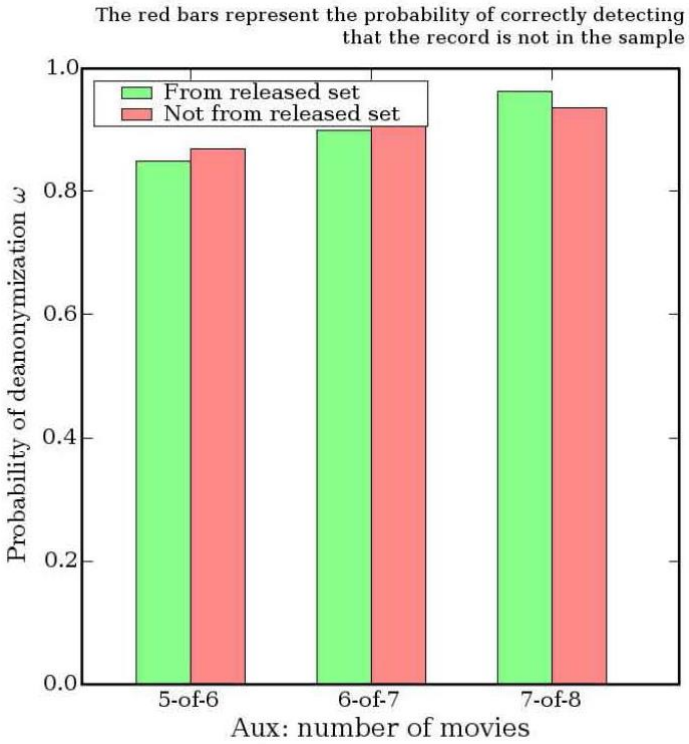
Best score — **Second-best score** — **Eccentricity threshold**

Eccentricity in the Netflix Dataset



Self-testing: Experimental Results

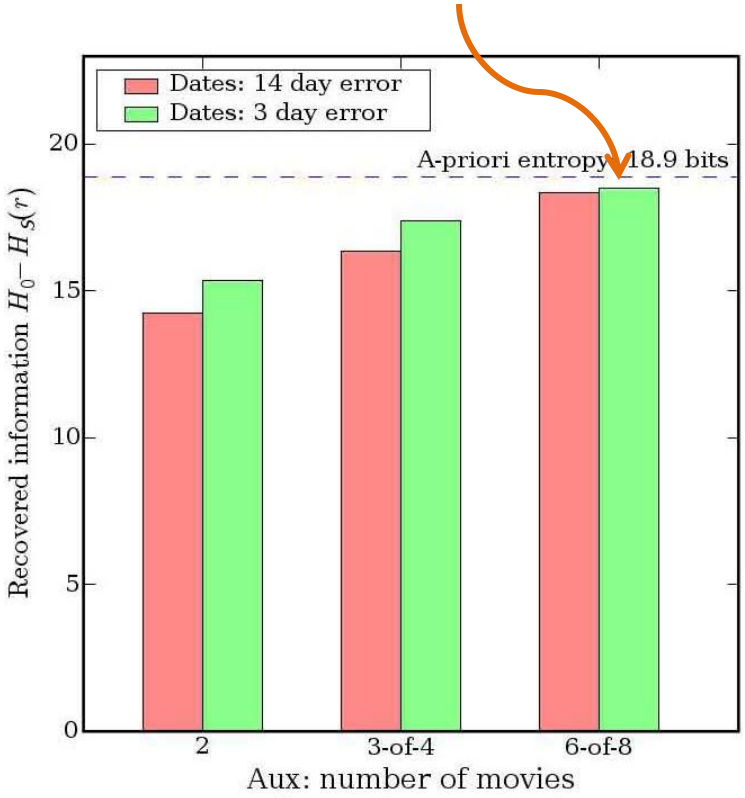
- After algorithm finds a match, remove the found record and re-run
- With very high probability, the algorithm now declares that there is no match



Robustness

- 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 huge amount of noise and perturbation

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



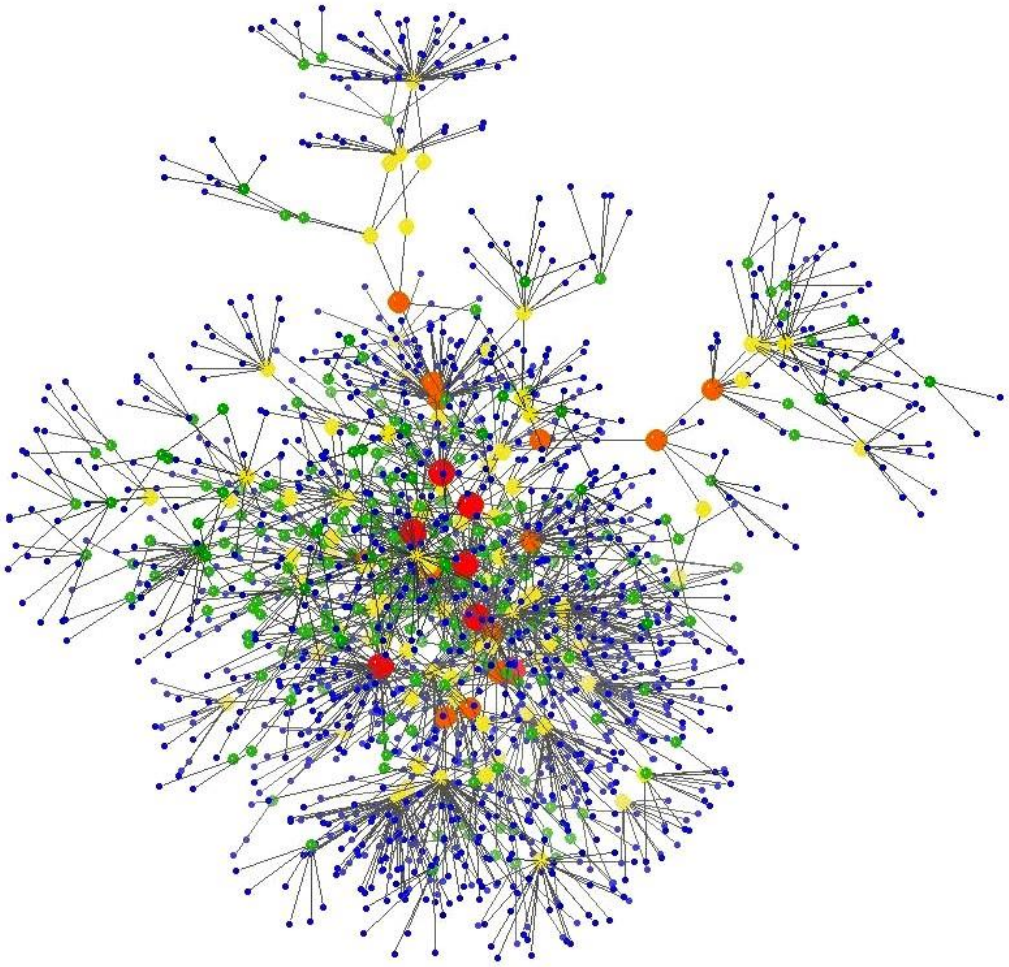
Main Themes

- 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
- 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

Collaborative filtering systems

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

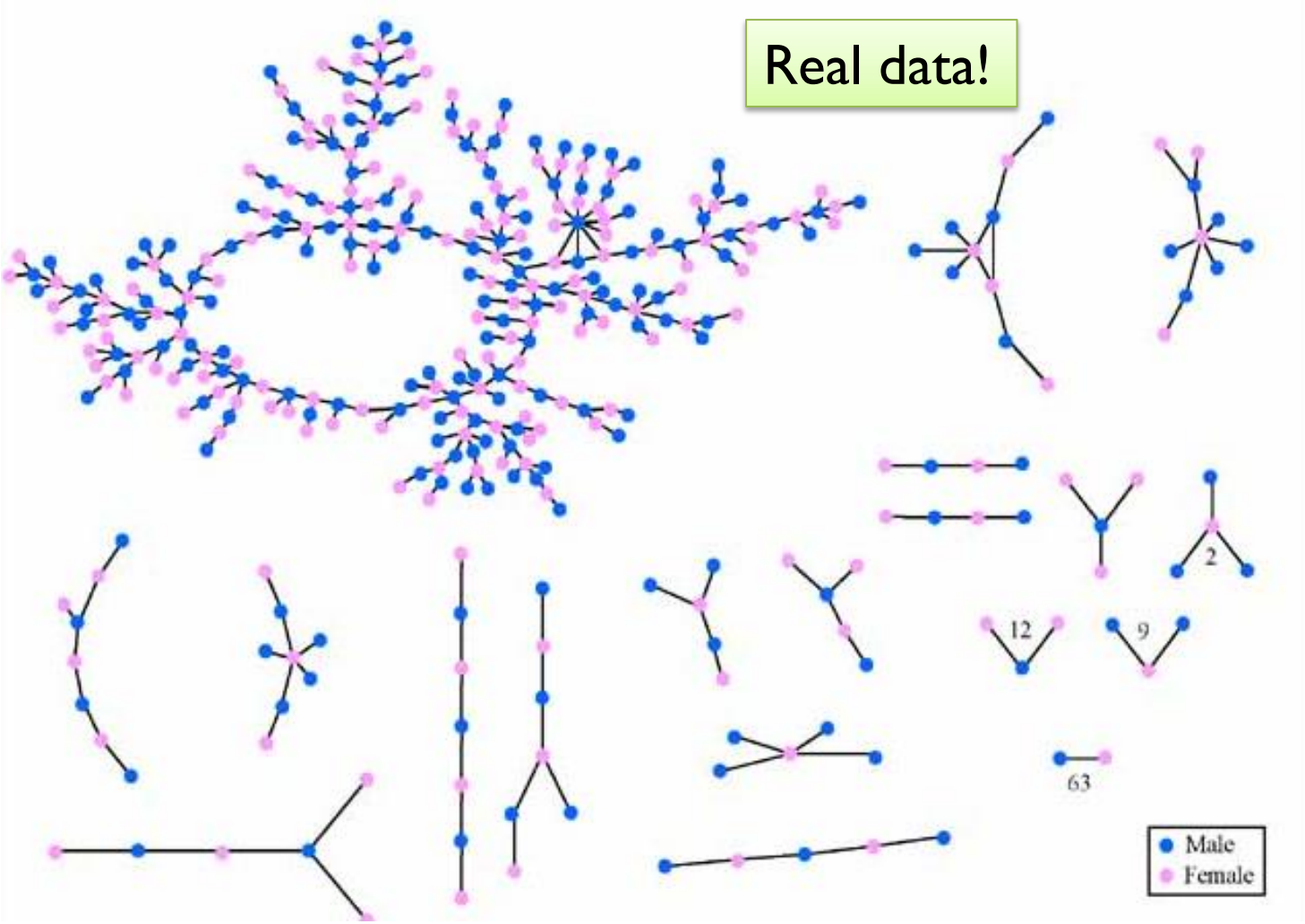
Oakland 2009

Narayanan, Shi, Rubinstein

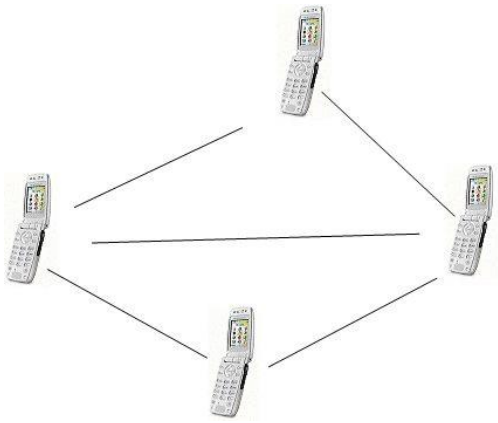
Link Prediction by De-anonymization:
How We Won the Kaggle Social Network Challenge

IJCNN 2011

"Jefferson High": Romantic and Sexual Network



Phone Call Graphs

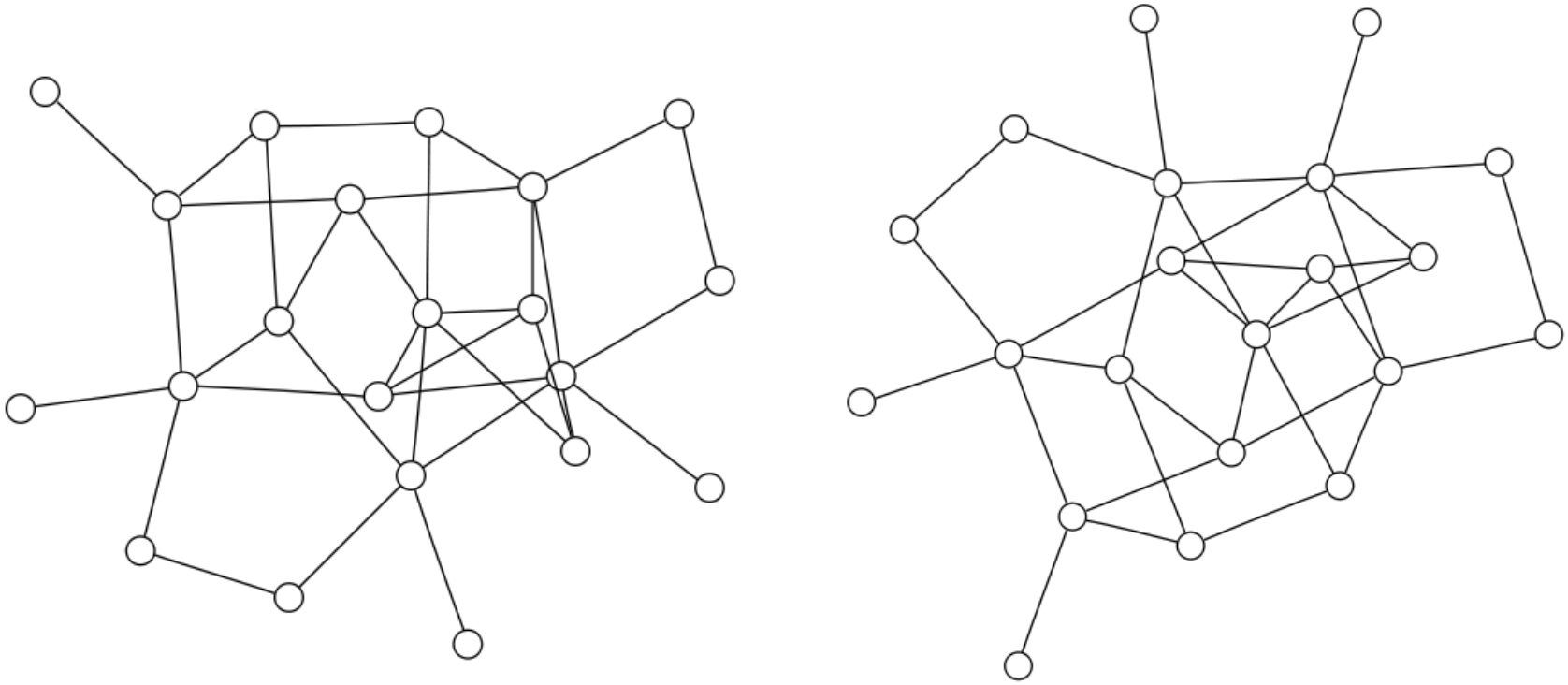


2 **trillion** edges

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

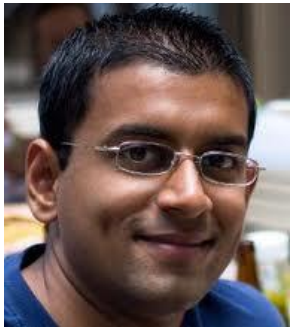
Structural De-anonymization



Goal: structural mapping between two graphs

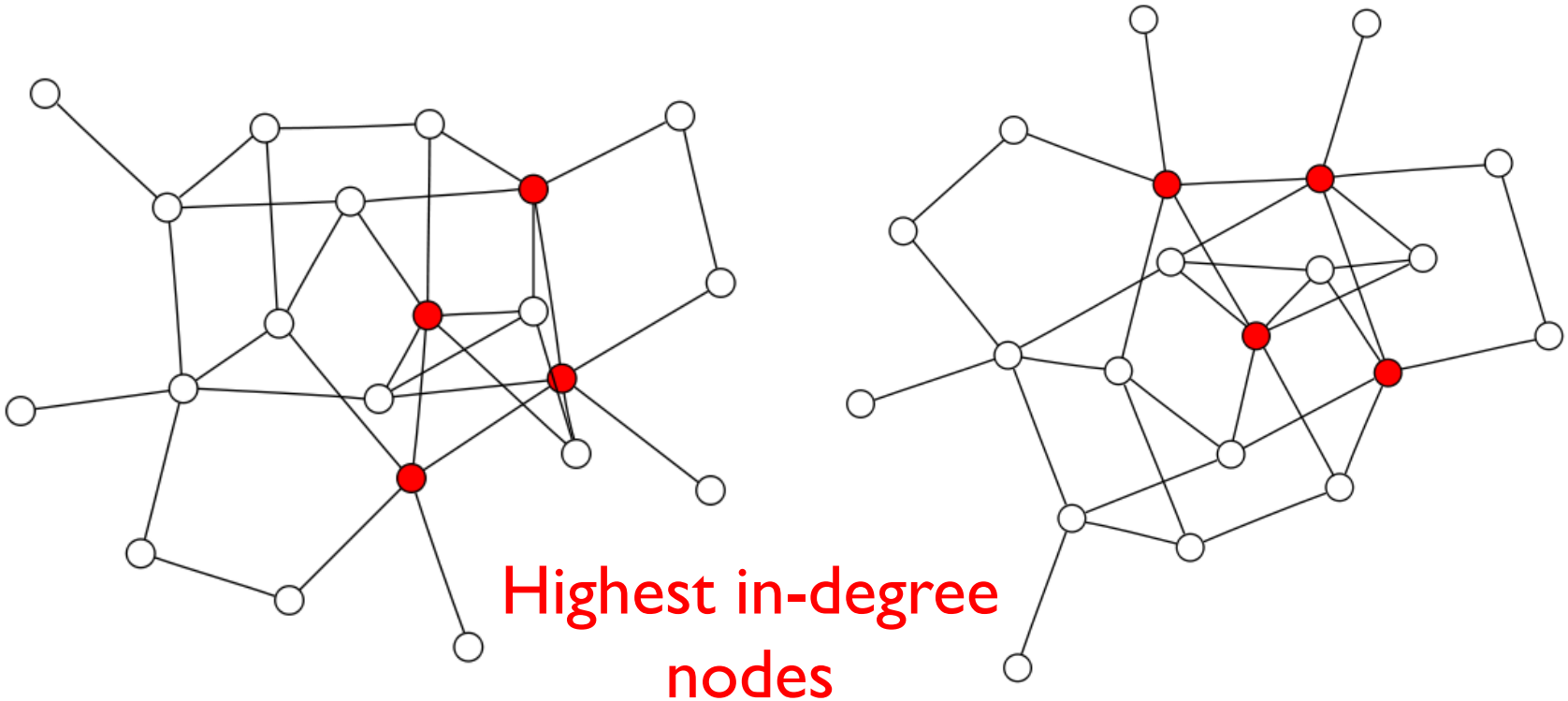
For example, Facebook vs. anonymized phone call graph

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")

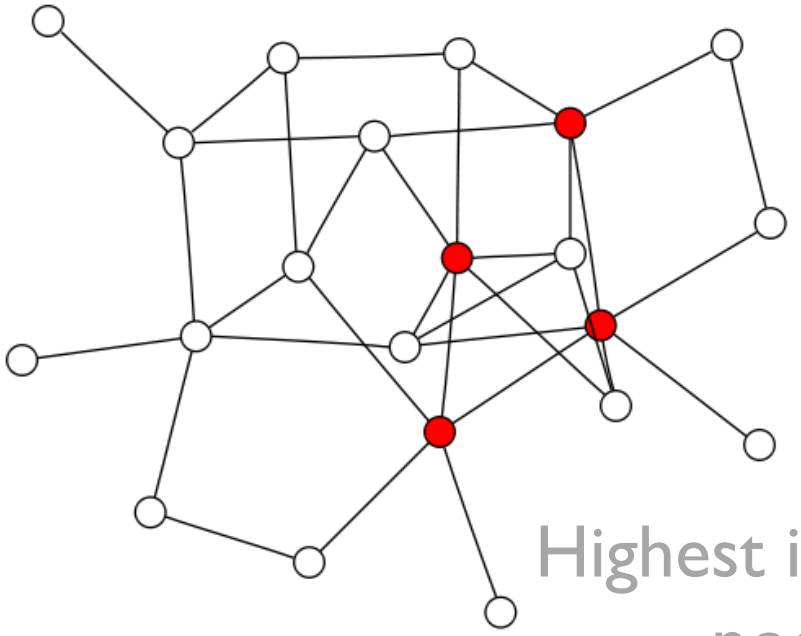
Where To Start?



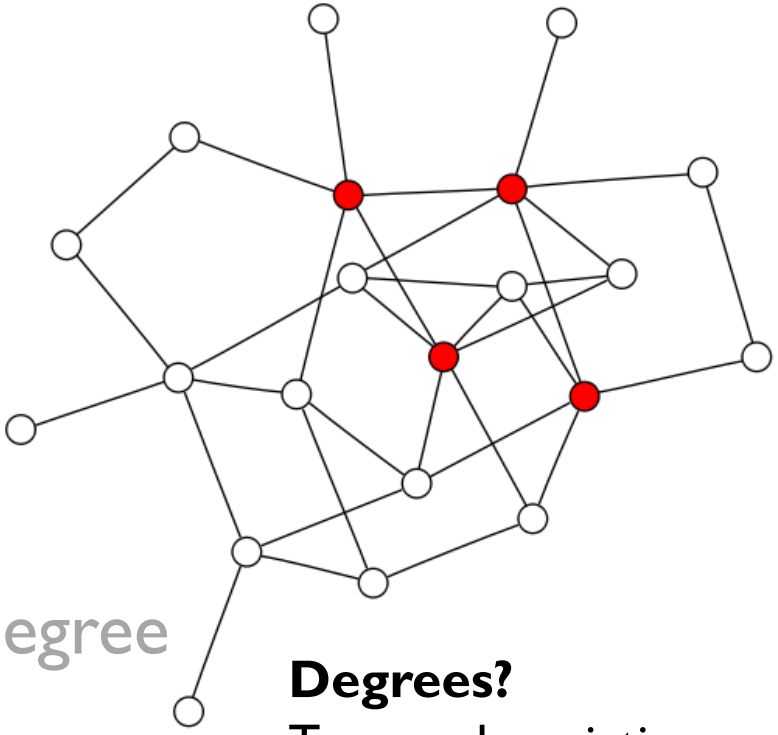
Highest in-degree nodes

Only a subset of nodes and edges in common

How To Match?



Highest in-degree nodes



Degrees?
Too much variation

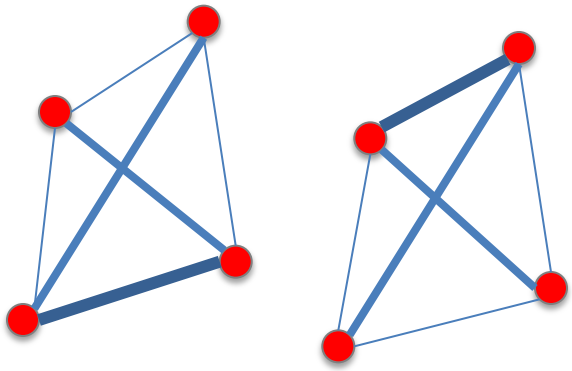
Subgraph structure?
Too sparse

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

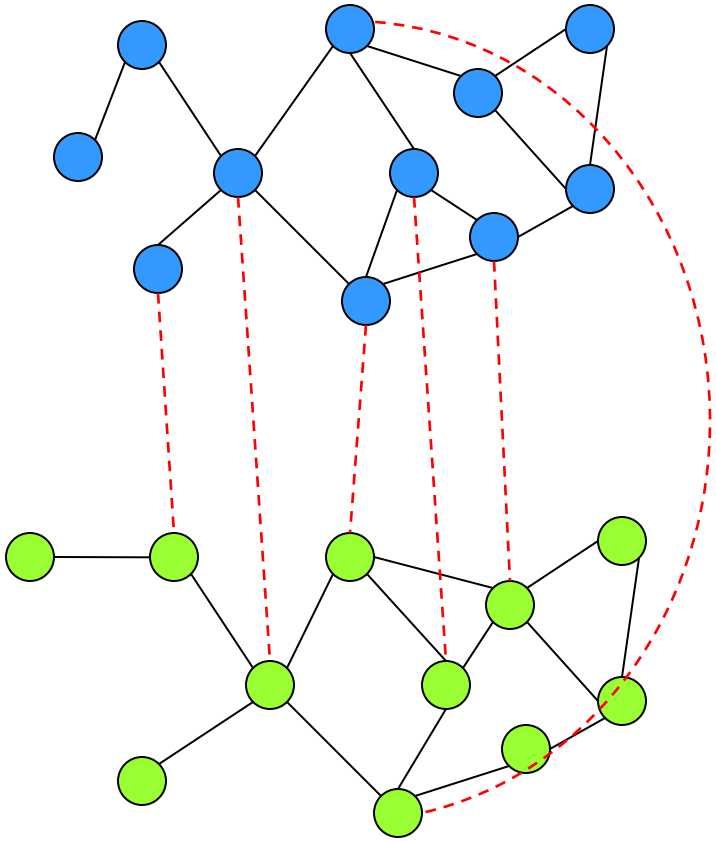
Seed Matching as Combinatorial Optimization

- 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

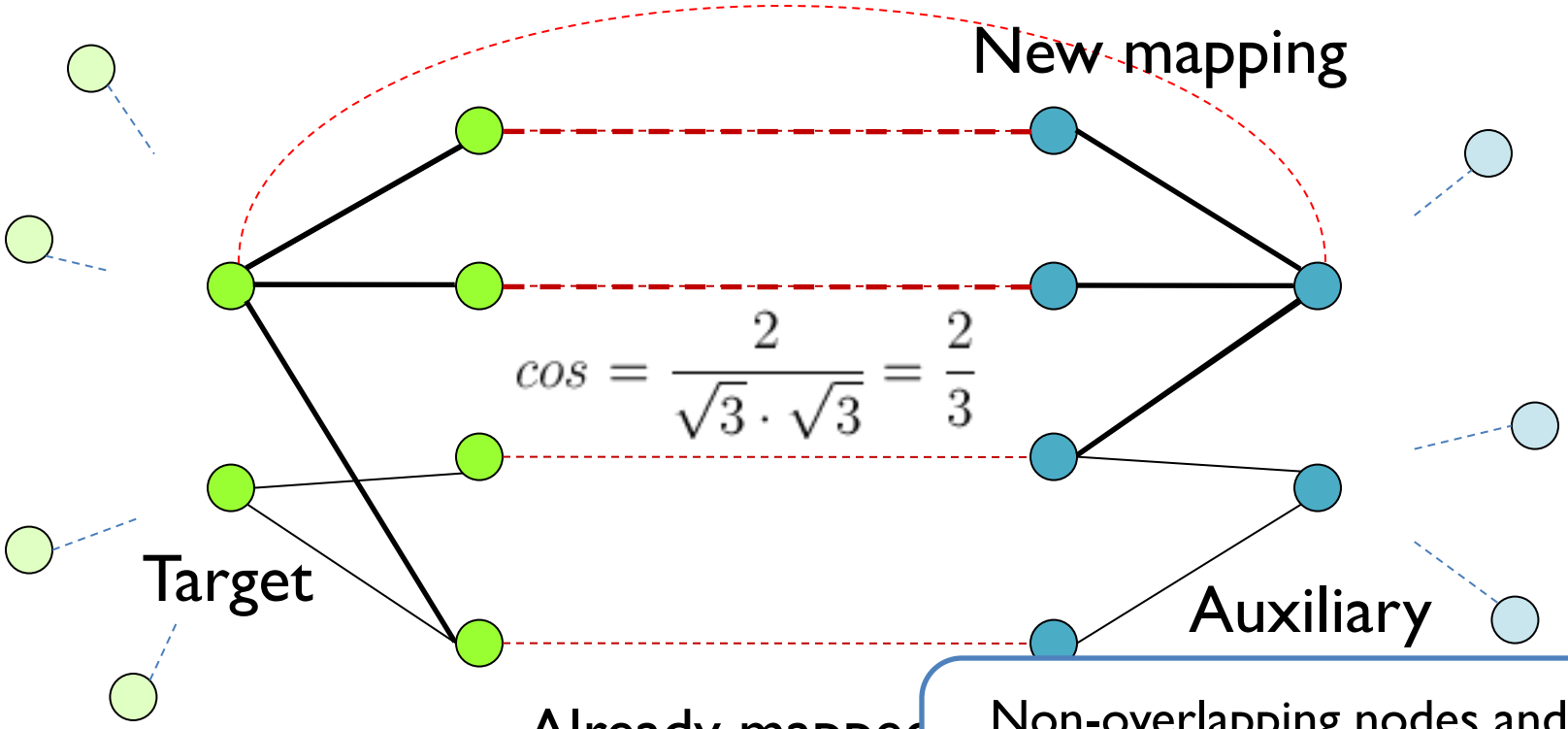


Iterative Propagation



"Seed" mapping

Propagation: Measuring Similarity



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

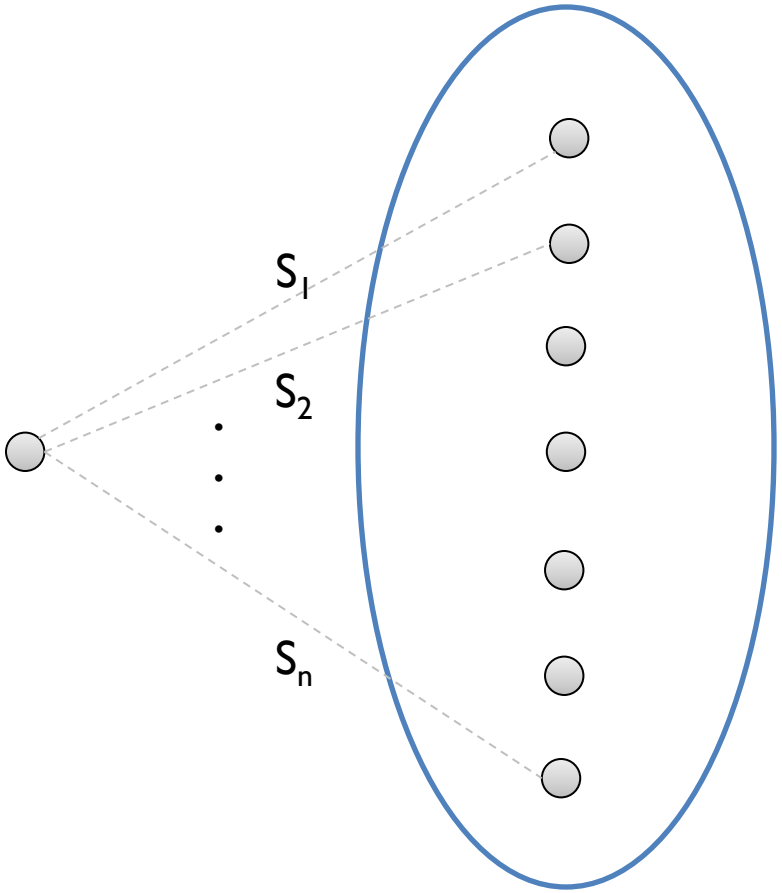
Self-correction

Eccentricity

Non-bijective

Deletion

Eccentricity



If true positive:

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

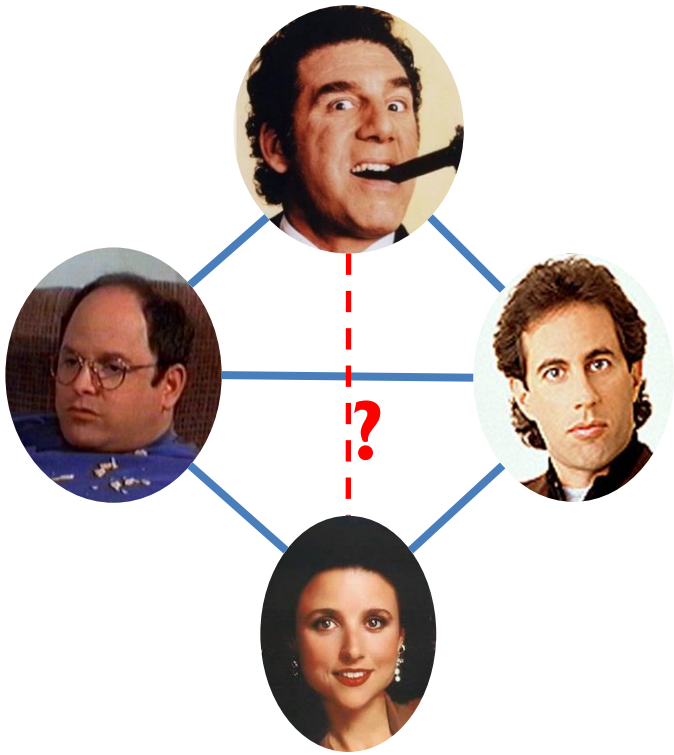
If false positive:

- $s_{max} - s_{max2}$ 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!

HIPAA
Health Insurance Portability & Accountability Act

Narayanan, Shmatikov
CACM column, 2010

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