

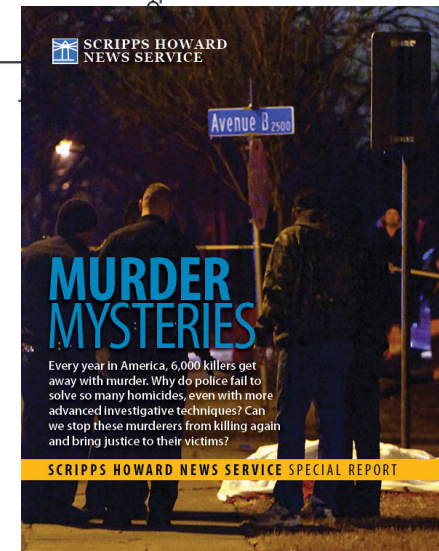
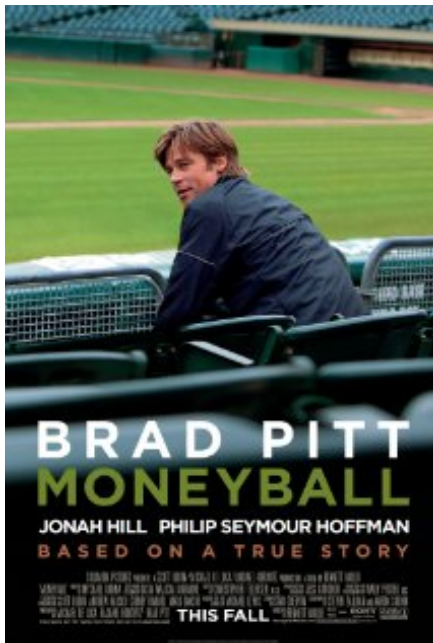
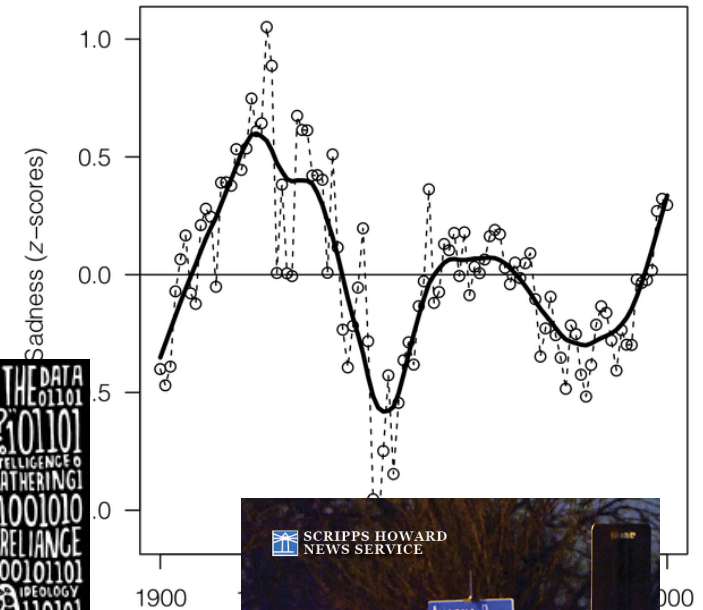
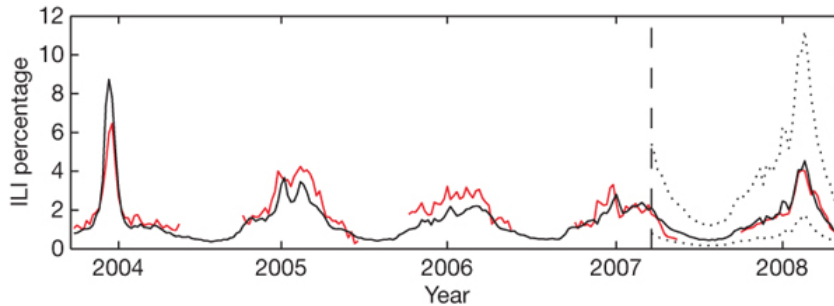
# Privacy in Data Analysis

Everything Data  
CompSci 216 Spring 2015



**DUKE**  
COMPUTER SCIENCE

# Data and \_\_\_\_\_ your favorite subject



# Where is all this data coming from?



# Where is all this data coming from?

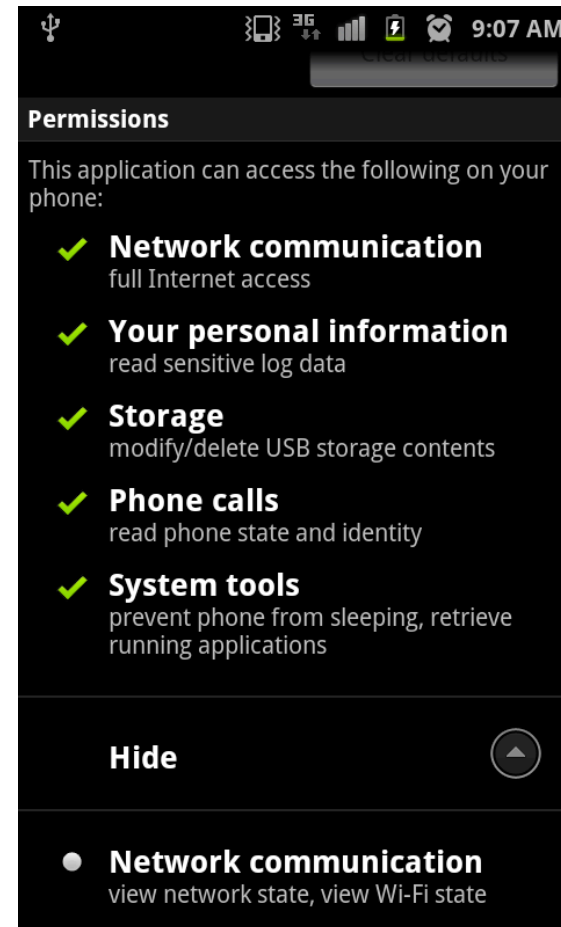
- Census surveys
- IRS Records
- Medical records
- Insurance records
- Search logs
- Browse logs
- Shopping histories
- Photos
- Videos
- Smart phone Sensors
- Mobility trajectories
- ...

**Very sensitive information ...**

# Sometimes users can know and control who sees their information

Who Can View My Full Profile	
<input type="radio"/>	My Friends Only
<input checked="" type="radio"/>	Public
<input type="radio"/>	Only Users <i>Over 18</i>

Privacy Settings	
<input type="checkbox"/>	Friend Requests - Require email or last name
<input type="checkbox"/>	Comments - approve before posting
<input type="checkbox"/>	Hide Online Now
<input type="checkbox"/>	Show My Birthday to my Friends 🎂
<input type="checkbox"/>	Photos - No Forwarding
<input type="checkbox"/>	Blog Comments - Friends Only
<input type="checkbox"/>	Friend Requests - No Bands
<input type="checkbox"/>	Block Users Under 18 From Contacting Me



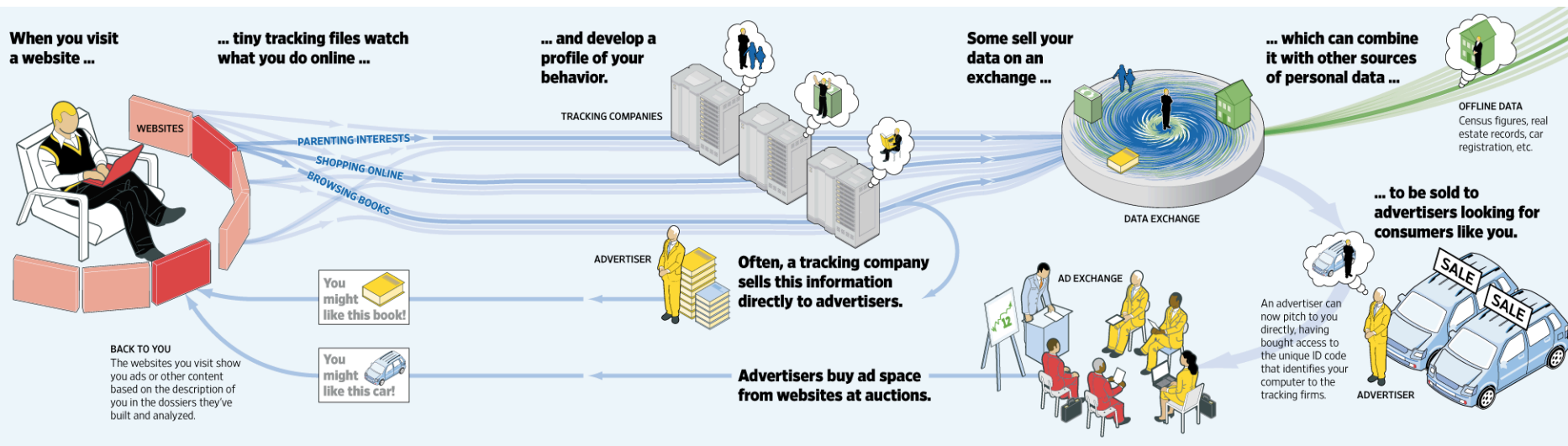
# ... but not always !!

The image is a composite graphic illustrating a security warning. It features three main elements:

- Twitter Post:** A screenshot of a tweet from the user '4sq.cc' that reads: "Got a table to myself this time. Feel free to stop by and sit in it. (@ Starbucks)". The tweet includes a link to 'http://4sq.cc' and a timestamp of '12:57 PM Nov 18th via foursquare'.
- Cartoon Burglar:** A cartoon illustration of a burglar with a beard, wearing a mask and a striped shirt, carrying a large green sack with a dollar sign on it.
- Facebook Post:** A screenshot of a Facebook post titled "Raising awareness about over-sharing". The post includes the text "Check out our [guest blog post](#) on the CDT website." and shows engagement metrics: "Like", "Send", and "29,523 people like this." Below the post, it says "Check your own Twitter timeline for checkins".

Overlaid on the Facebook post is the large, bold, red text "PLEASE ROB ME". To the right of this text are two red location pin icons, each with a white 'X' inside, set against a background of a stylized map.

# Example: Targeted Advertising



<http://graphicsweb.wsj.com/documents/divSlider/media/ecosystem100730.png>



# What websites track your behavior?

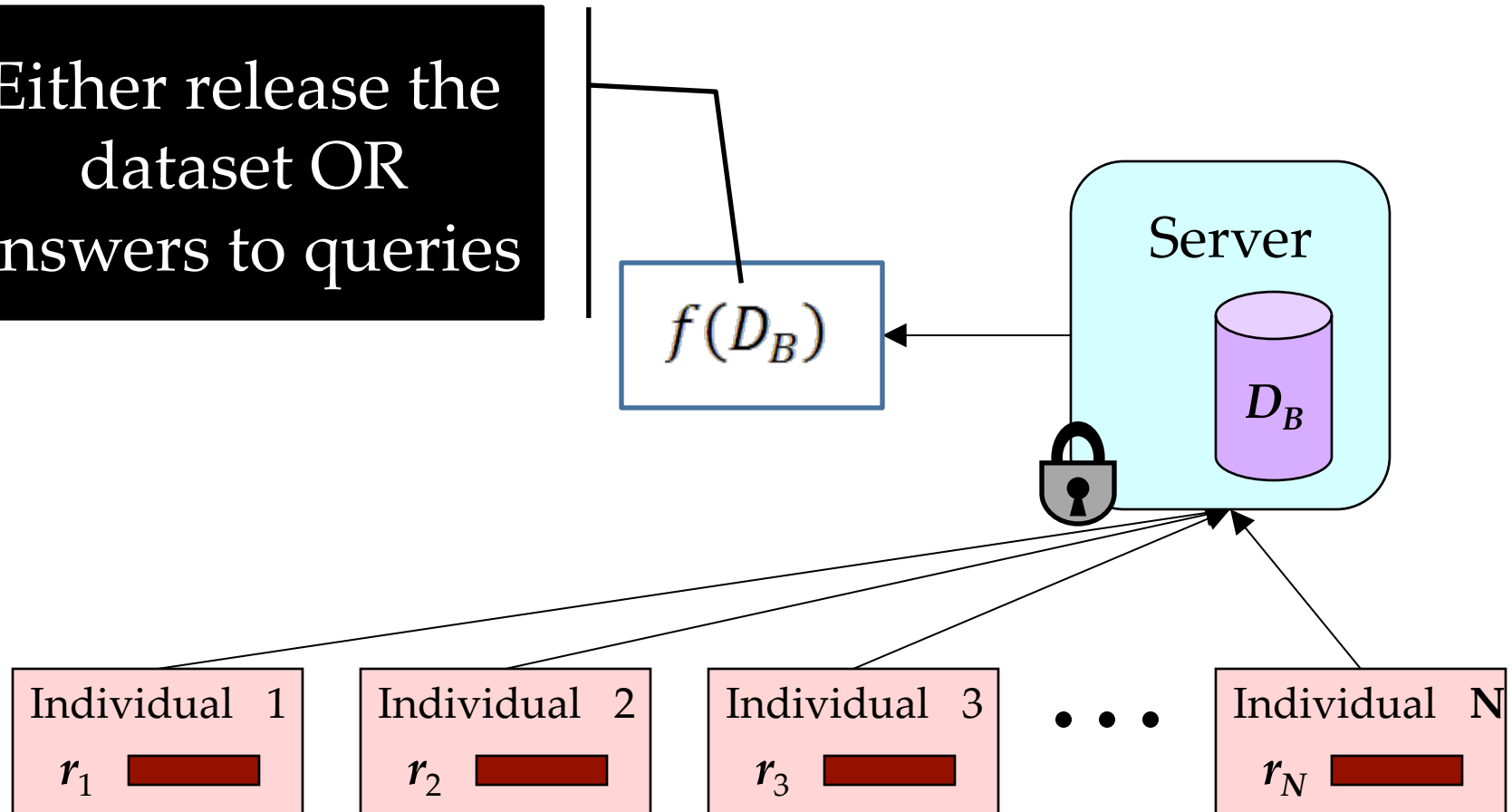
Site	Exposure Index	Trackers
dictionary.com	Very High	234
merriam-webster.com	High	131
comcast.net	High	151
careerbuilder.com	High	118
photobucket.com	High	127
msn.com	High	207
answers.com	Medium	120
yp.com	Medium	89
msnbc.com	Medium	117
yahoo.com	Medium	106
aol.com	Medium	133
wiki.answers.com	Medium	72
cnn.com	Medium	72
about.com	Medium	83
cnet.com	Medium	81
verizonwireless.com	Medium	90
imdb.com	Medium	55
live.com	Medium	115
att.com	Medium	58
walmart.com	Medium	66
bbc.co.uk	Medium	45
ebay.com	Medium	42
ehow.com	Medium	55

<http://blogs.wsj.com/wtk/>



# Servers track your information ... so what?

Either release the  
dataset OR  
answers to queries



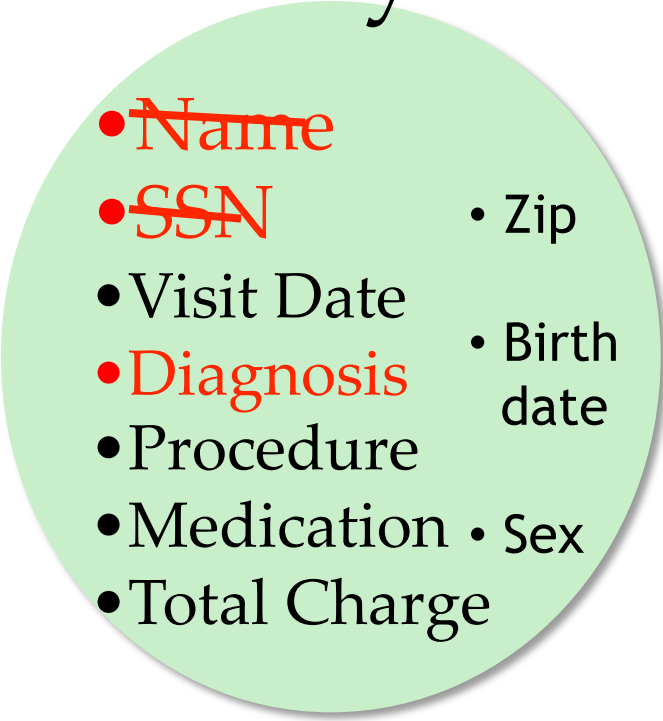
# Does it matter ... I am anonymous, right?



Source (<http://xkcd.org/834/>)

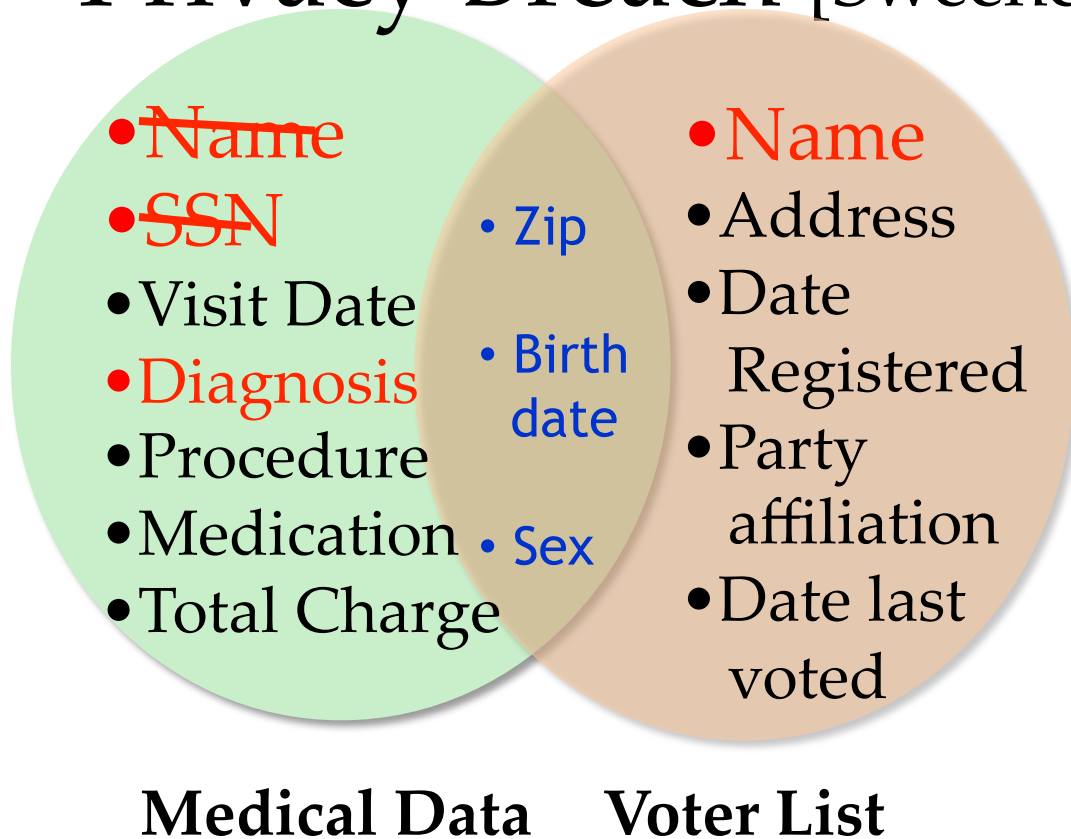
*What if we ensure our names and other identifiers are never released?*

# The Massachusetts Governor Privacy Breach [Sweeney IJUFKS 2002]

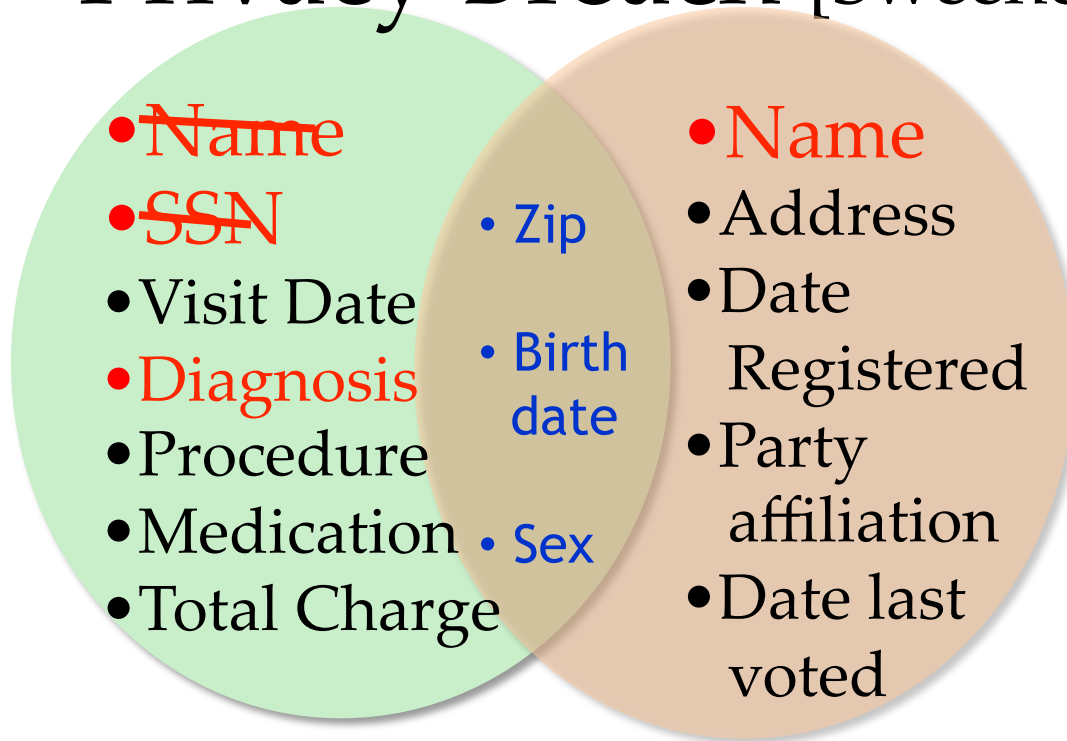
- 
- ~~Name~~
  - ~~SSN~~
  - Visit Date
  - ~~Diagnosis~~
  - Procedure
  - Medication
  - Total Charge
  - Zip
  - Birth date
  - Sex

**Medical Data**

# The Massachusetts Governor Privacy Breach [Sweeney IJUFKS 2002]



# The Massachusetts Governor Privacy Breach [Sweeney IJUFKS 2002]

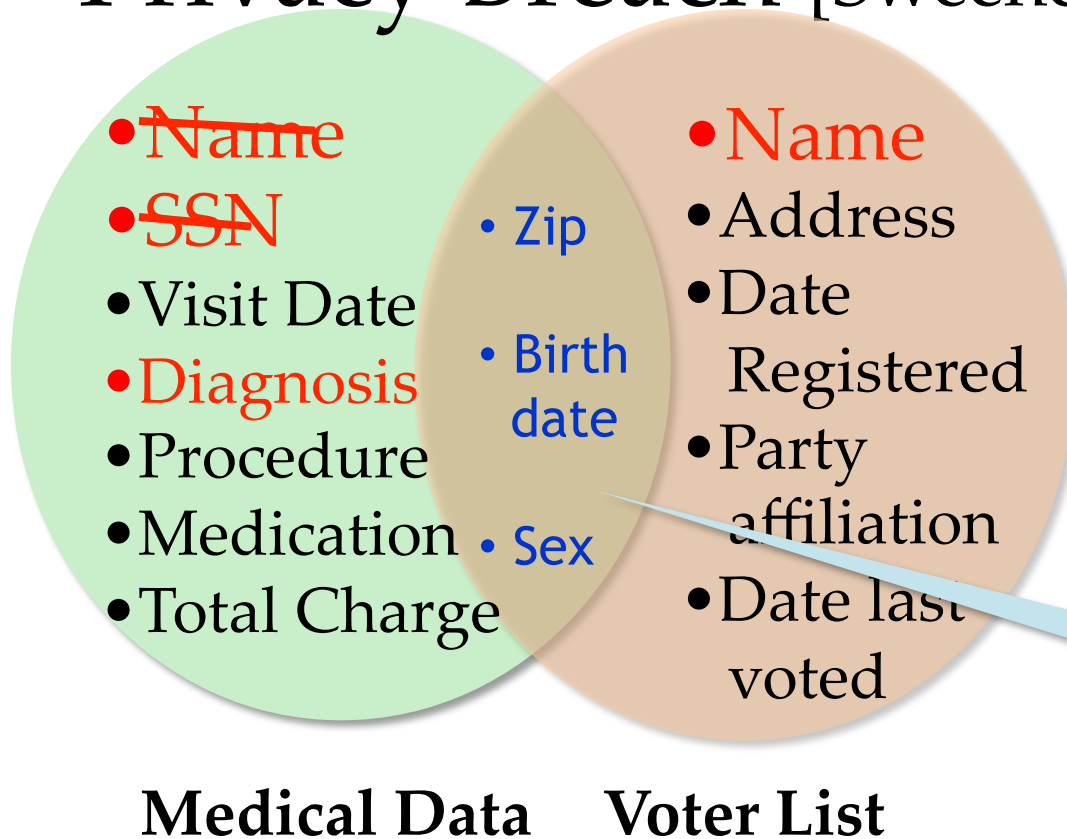


Medical Data      Voter List

- Governor of MA uniquely identified using ZipCode, Birth Date, and Sex.

**Name linked to Diagnosis**

# The Massachusetts Governor Privacy Breach [Sweeney IJUFKS 2002]



- 87 % of US population **uniquely identified** using ZipCode, Birth Date, and Sex.

**Quasi Identifier**

# AOL data publishing fiasco



— IN SOLIDARITY WITH THE MANY AOL USERS WHOSE OFTEN EMBARRASSING WEB SEARCHES WERE RELEASED TO THE PUBLIC, I OFFER A SAMPLE OF MY OWN SEARCH HISTORY:



[Web](#) [Images](#) [Video](#) <sup>New!</sup> [News](#) [Maps](#) [more »](#)

[Advanced Search](#)  
[Preferences](#)  
[Language Tools](#)

velociraptors  
site:imdb.com "jurassic park"  
raptors  
dromaeosaurids  
utahraptor  
"home depot" deadbolts  
security home improvement  
surviving a raptor attack  
robert bakker paleontologist  
robert bakker "possible raptor sympathizer"  
site:en.wikipedia.org surviving a raptor attack  
learning from mistakes in jurassic park  
big-game rifles  
tire irons  
treating raptor wounds  
do raptors fear fire  
how to make a molotov cocktail  
do raptors fear death  
can raptors pick locks  
how to tell if my neighbors are raptors



# AOL data publishing fiasco ...

Ashwin222

Uefa cup

Ashwin222

Uefa champions league

Ashwin222

Champions league final

Ashwin222

Champions league final 2013

Jun156

exchangeability

Jun156

Proof of deFinetti's theorem

Brett12345

Zombie games

Brett12345

Warcraft

Brett12345

Beatles anthology

Brett12345

Ubuntu breeze

Austin222

Python in thought

Austin222

Enthought Canopy

# User IDs replaced with random numbers

865712345

Uefa cup

865712345

Uefa champions league

865712345

Champions league final

865712345

Champions league final 2013

236712909

exchangeability

236712909

Proof of deFinetti's theorem

112765410

Zombie games

112765410

Warcraft

112765410

Beatles anthology

112765410

Ubuntu breeze

865712345

Python in thought

865712345

Enthought Canopy


# Privacy Breach

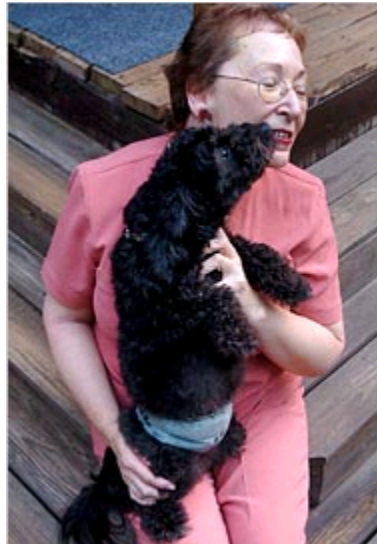
[NYTimes 2006]

## A Face Is Exposed for AOL Searcher No. 4417749

By [MICHAEL BARBARO](#) and [TOM ZELLER Jr.](#)

Published: August 9, 2006

 [SIGN IN TO E-  
THIS](#)



# Privacy violations from Facebook



[http://article.wn.com/view/2012/08/28/  
Facebooks\\_new\\_app\\_bazaar\\_violates\\_punters\\_privacy\\_lobbyists/](http://article.wn.com/view/2012/08/28/Facebooks_new_app_bazaar_violates_punters_privacy_lobbyists/)

# Inference from Impressions: Sexual Orientation

[Korolova JPC 2011]

Facebook Profile

Number of  
Impressions

25

+ Who are  
interested in  
**Men**

0

+ Who are  
interested in  
**Women**



+

Online Data



- who live in the **United States**
- who live within 50 miles of **Staten Island, NY**
- between the ages of **23 and 27** inclusive
- who are **female**
- who are connected to **DogAnd PonyShow**
- in one of the categories: **Pop Culture, Science Fiction/Fantasy, Alternative, Rock, Classic Rock or iPhone**



Facebook uses private information to predict match to ad

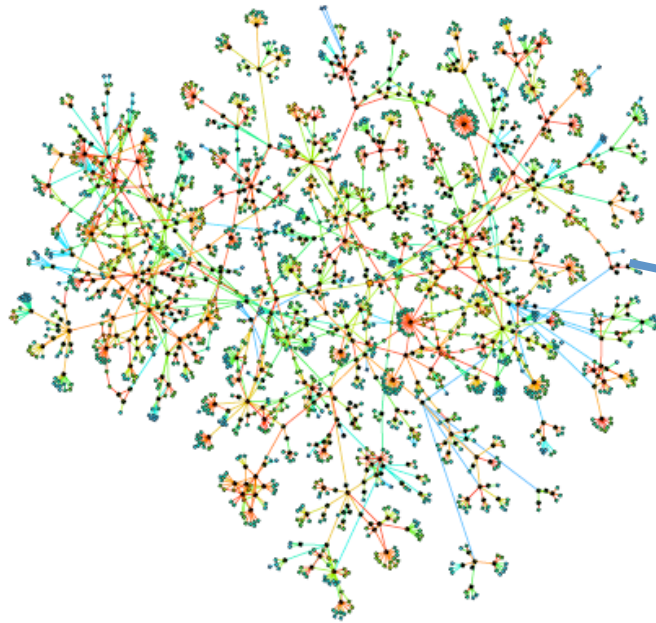
# Reason for Privacy Breach

- Anyone can run a campaign with strict targeting criteria
  - Zip, birthdate and sex uniquely identify 87% of US population
- “Private” and “Friends only” profile info used to determine match
- Default privacy settings lead to users having many publicly visible features
  - Default privacy setting for Likes, location, work place, etc. is **public**

# Can Facebook release its graph ?

- Suppose we release just release the nodes and edges in the Facebook graph ...



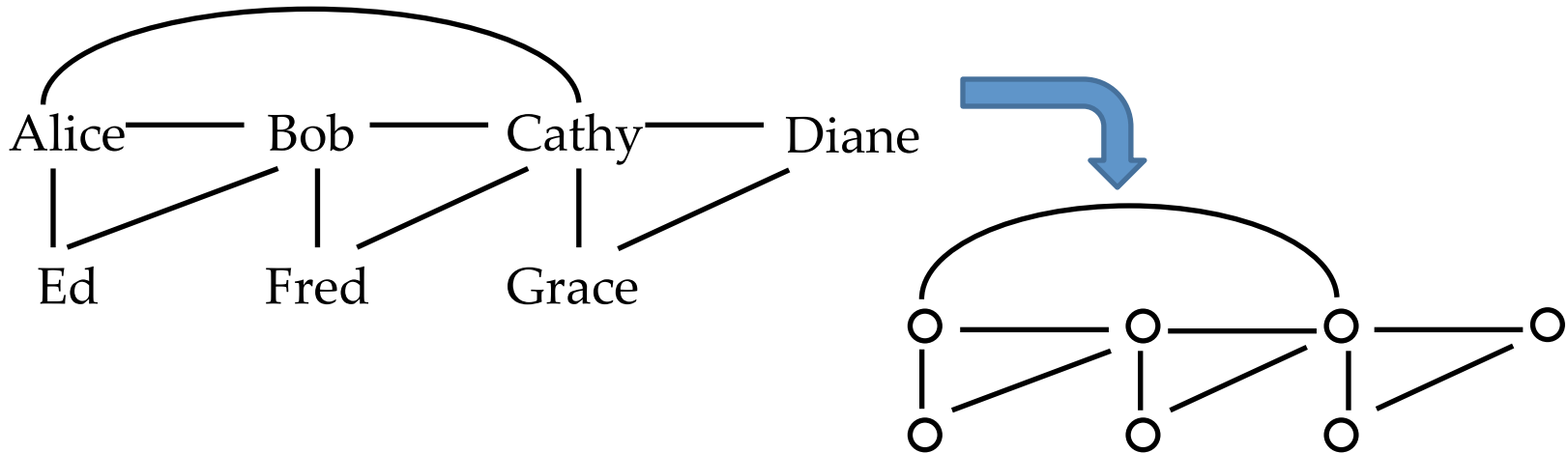


Mobile communication  
networks  
[J. Onnela et al. PNAS 07]

Sexual & Injection Drug  
Partners  
[Potterat et al. STI 02]

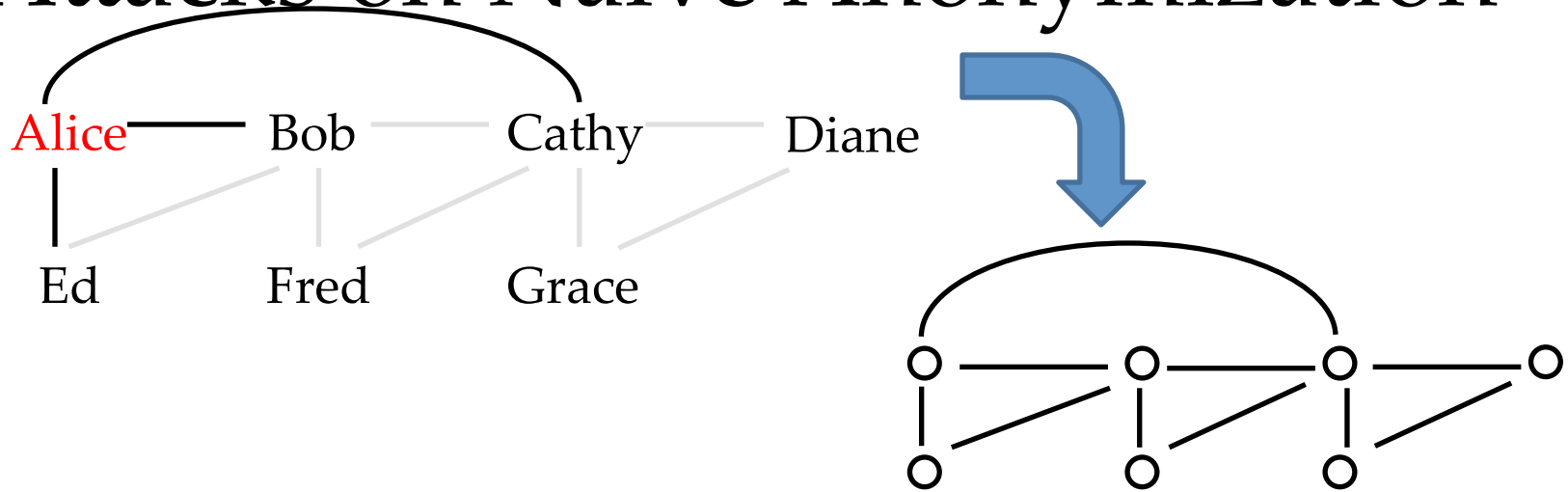


# Naïve anonymization



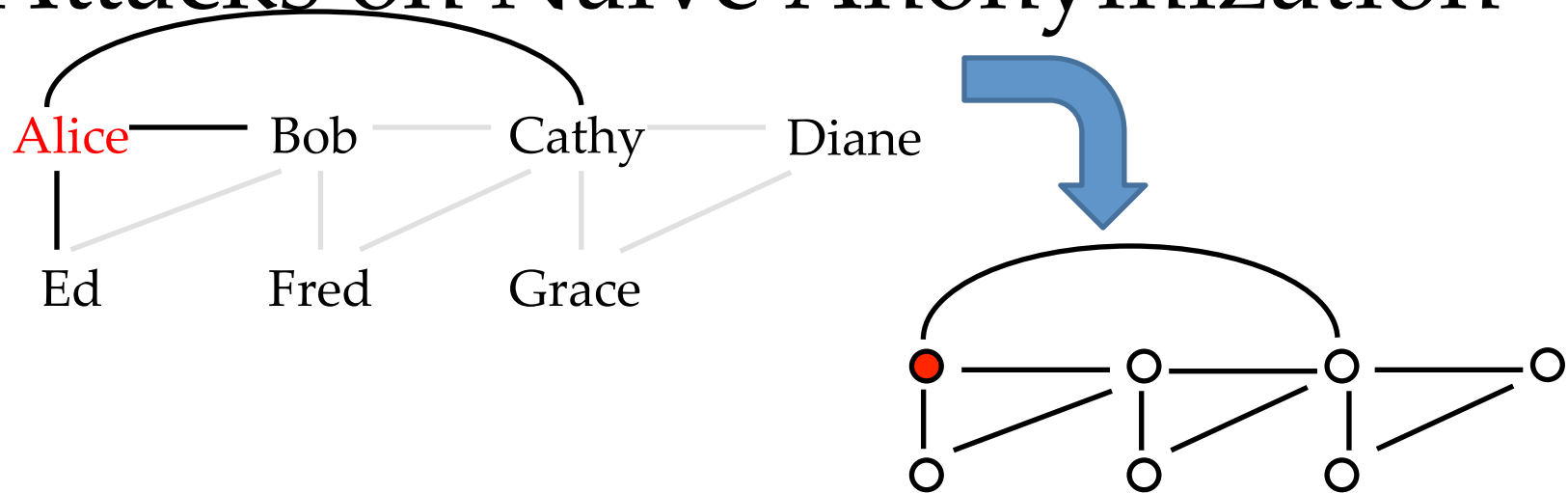
- Consider the above email communication graph
  - Each node represents an individual
  - Each edge between two individuals indicates that they have exchanged emails
- Replace node identifiers with random numbers.

# Attacks on Naïve Anonymization



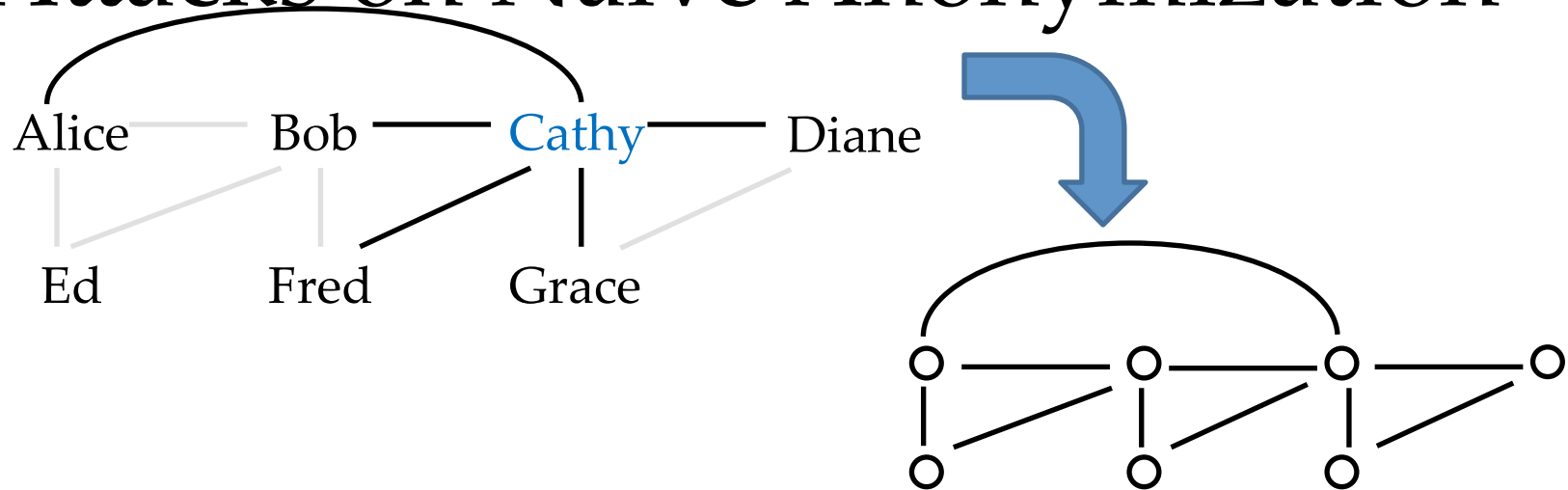
- Alice has sent emails to three individuals only

# Attacks on Naïve Anonymization



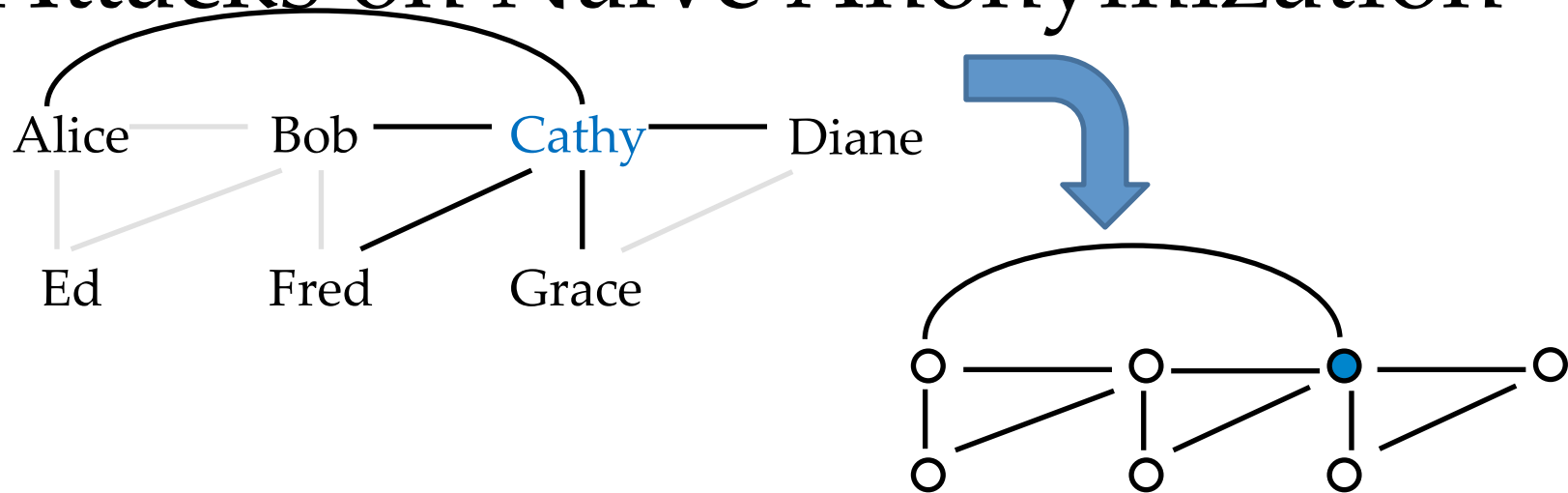
- Alice has sent emails to three individuals only
- Only one node in the anonymized network has a degree three
- Hence, Alice can re-identify herself

# Attacks on Naïve Anonymization



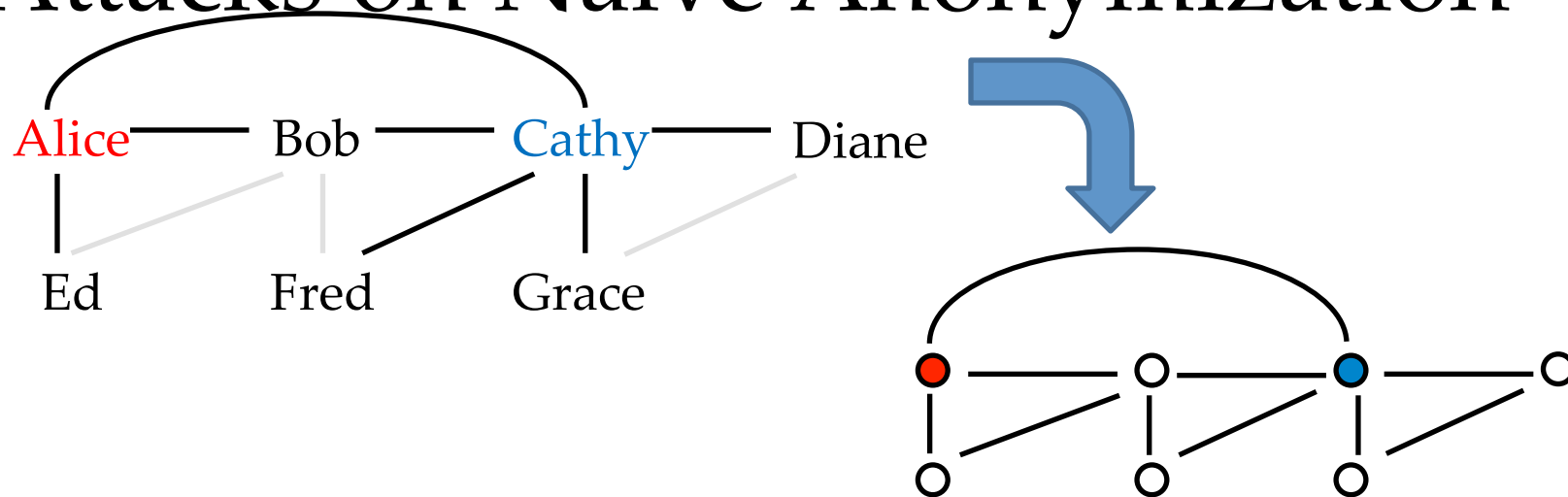
- Cathy has sent emails to five individuals

# Attacks on Naïve Anonymization



- Cathy has sent emails to five individuals
- Only one node has a degree five
- Hence, Cathy can re-identify herself

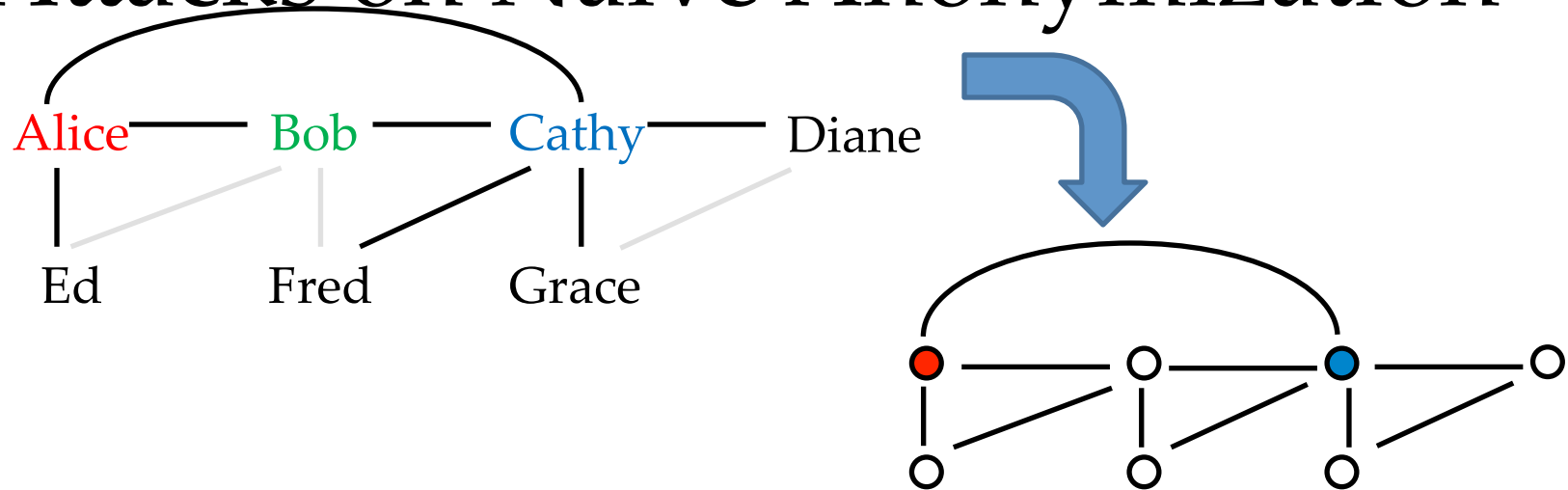
# Attacks on Naïve Anonymization



- Now consider that Alice and Cathy share their knowledge about the anonymized network
- What can they learn about the other individuals?

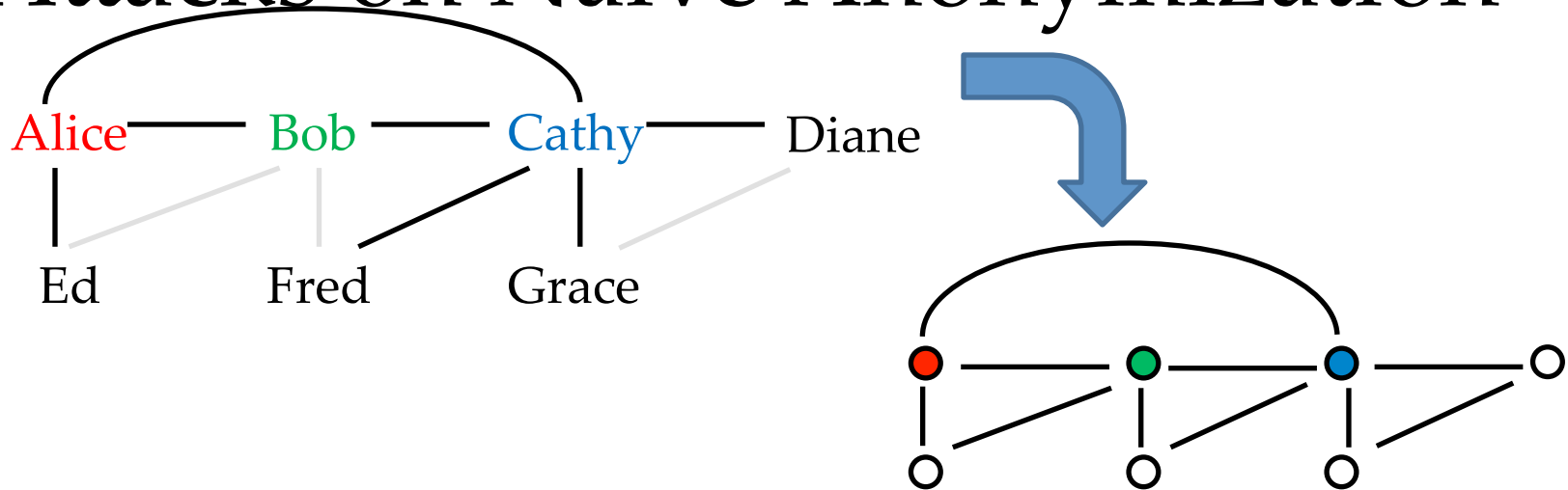


# Attacks on Naïve Anonymization



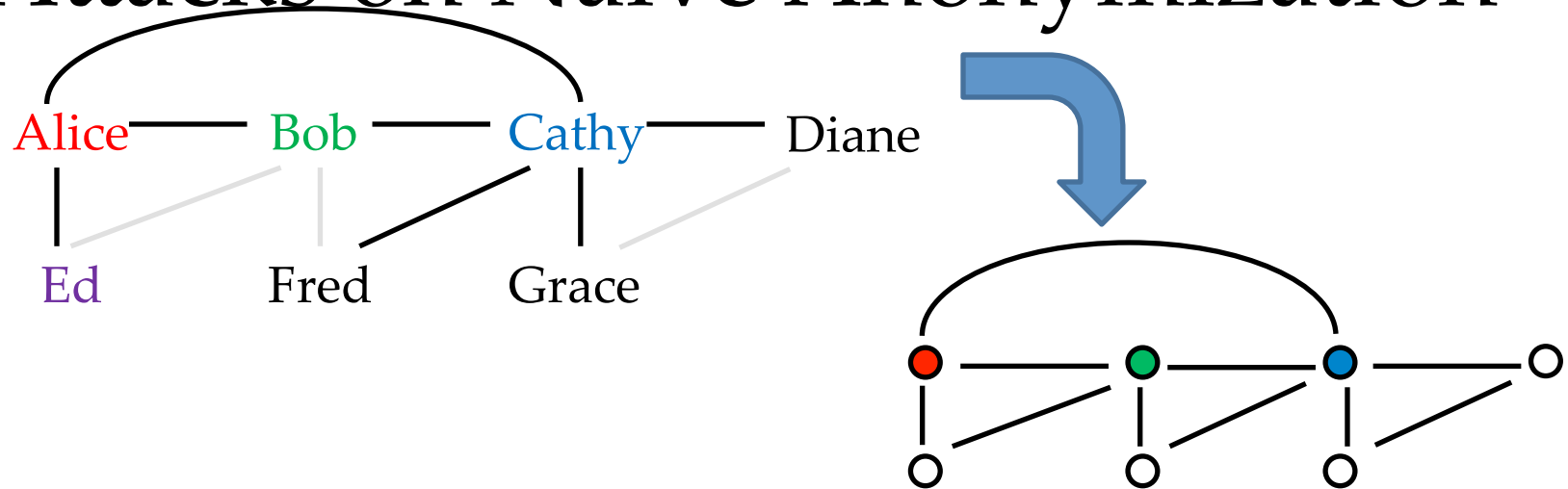
- First, Alice and Cathy know that only Bob have sent emails to both of them

# Attacks on Naïve Anonymization



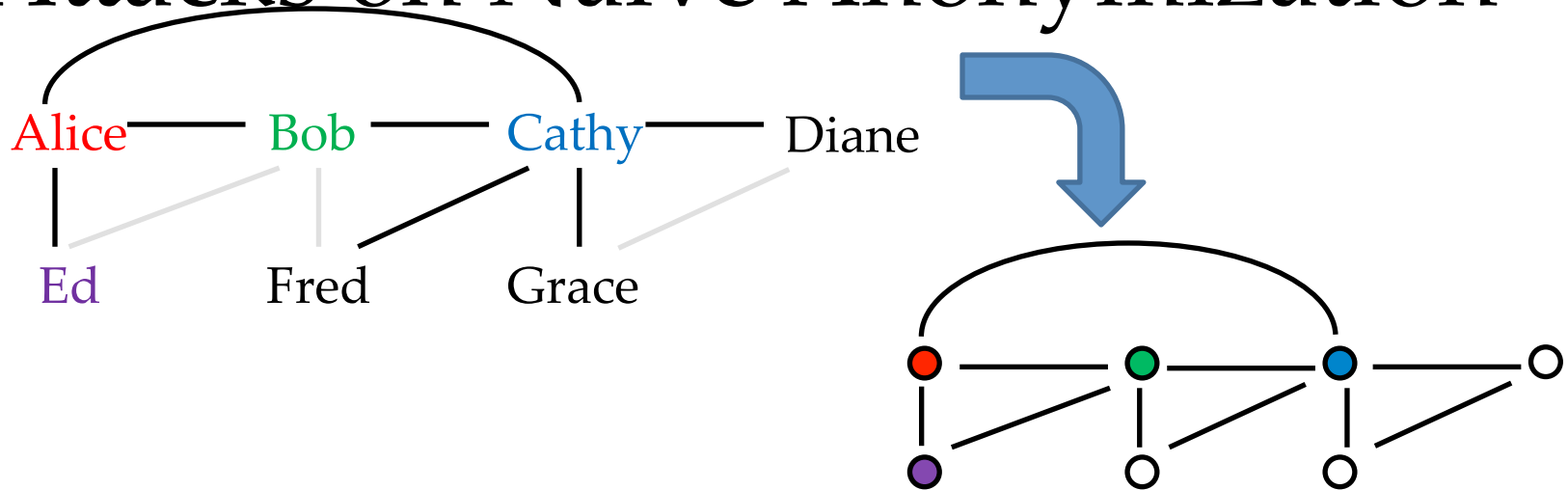
- First, Alice and Cathy know that only Bob have sent emails to both of them
- Bob can be identified

# Attacks on Naïve Anonymization



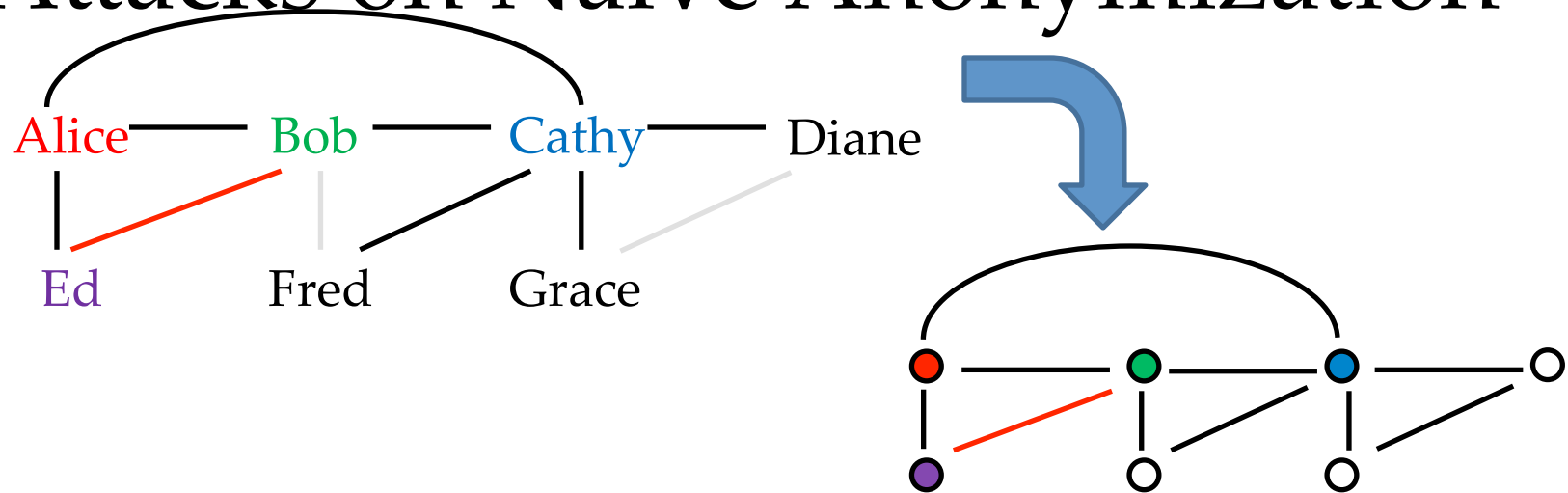
- Alice has sent emails to Bob, Cathy, and Ed only

# Attacks on Naïve Anonymization



- Alice has sent emails to Bob, Cathy, and Ed only
- Ed can be identified

# Attacks on Naïve Anonymization

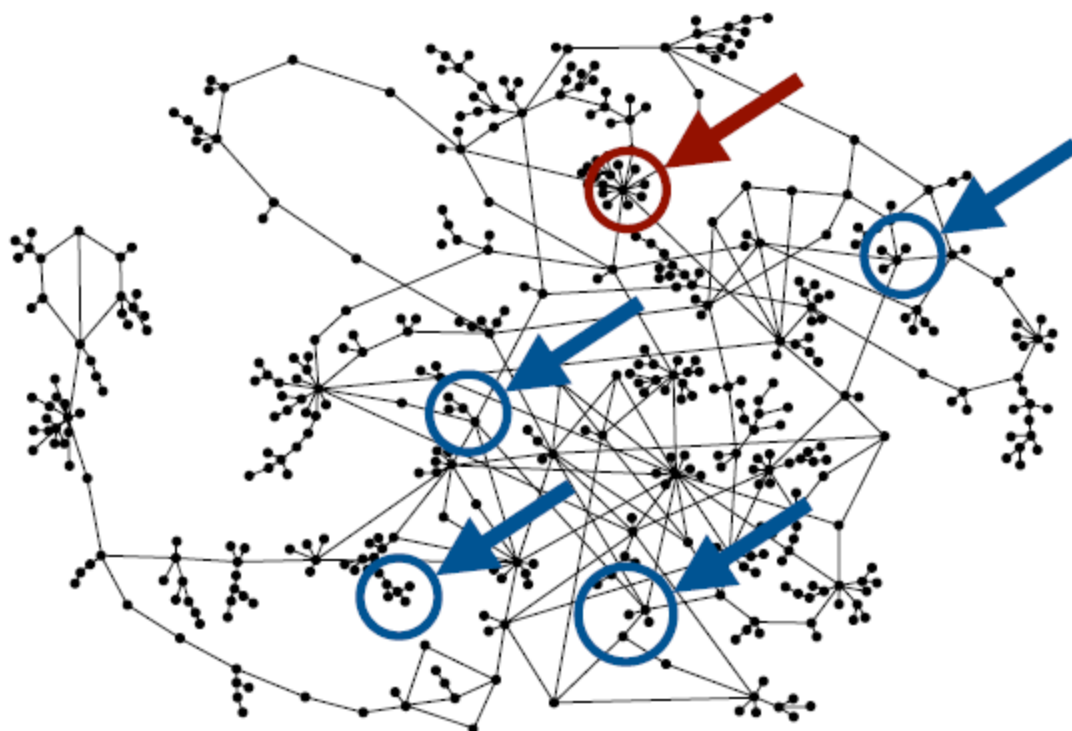


- Alice and Cathy can learn that Bob and Ed are connected

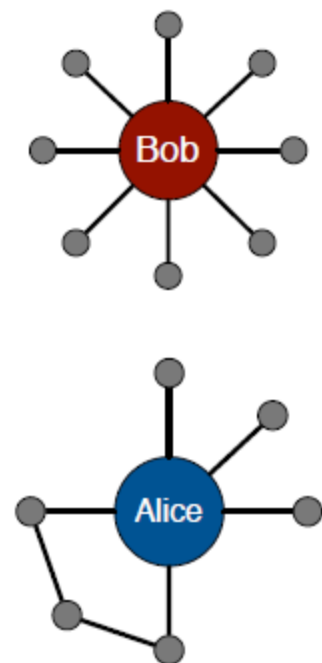
# Attacks

**Matching attack:** the adversary matches external information to a naively anonymized network.

**unique or partial  
node re-identification**



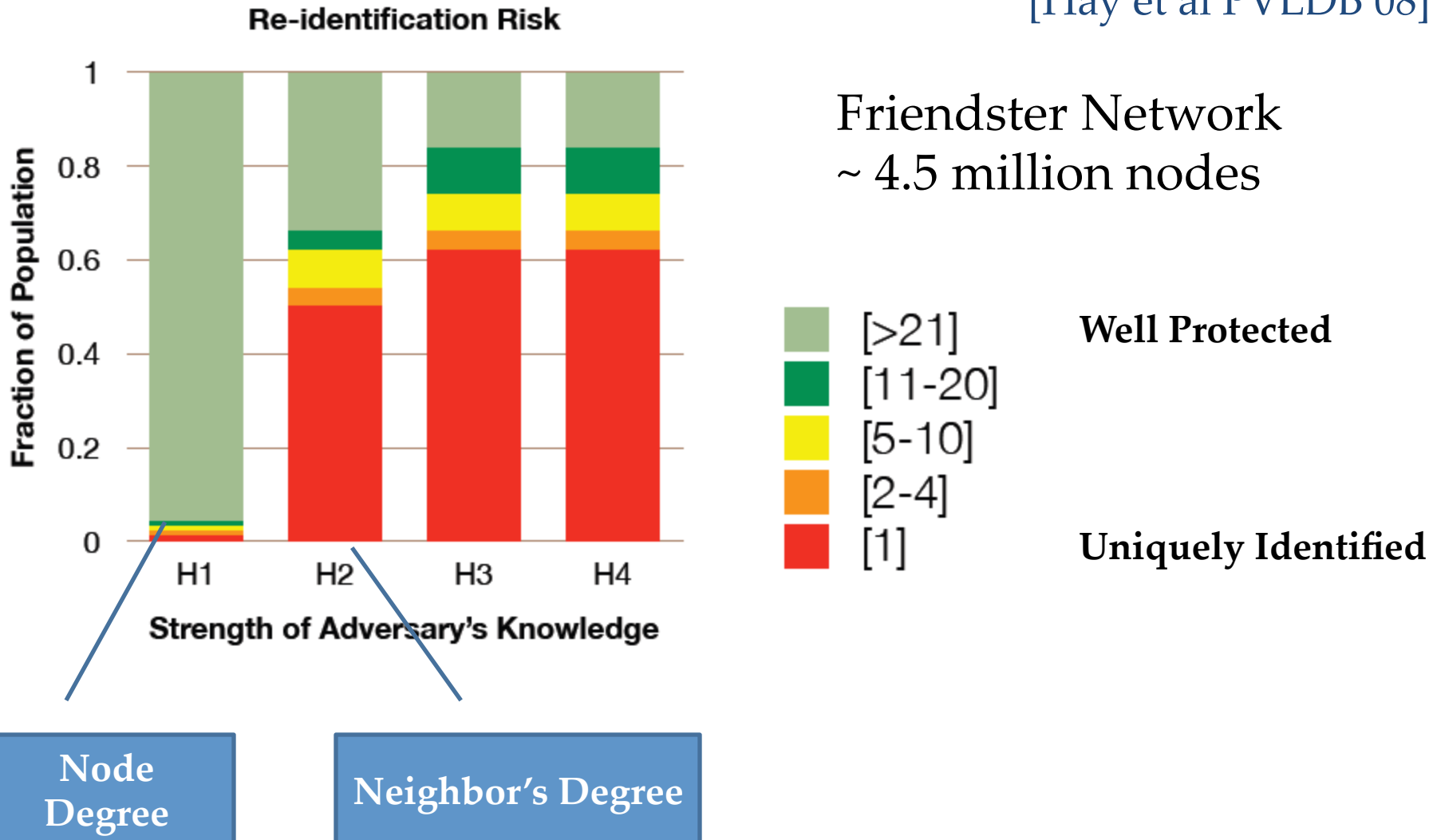
Naively Anonymized Network



External information

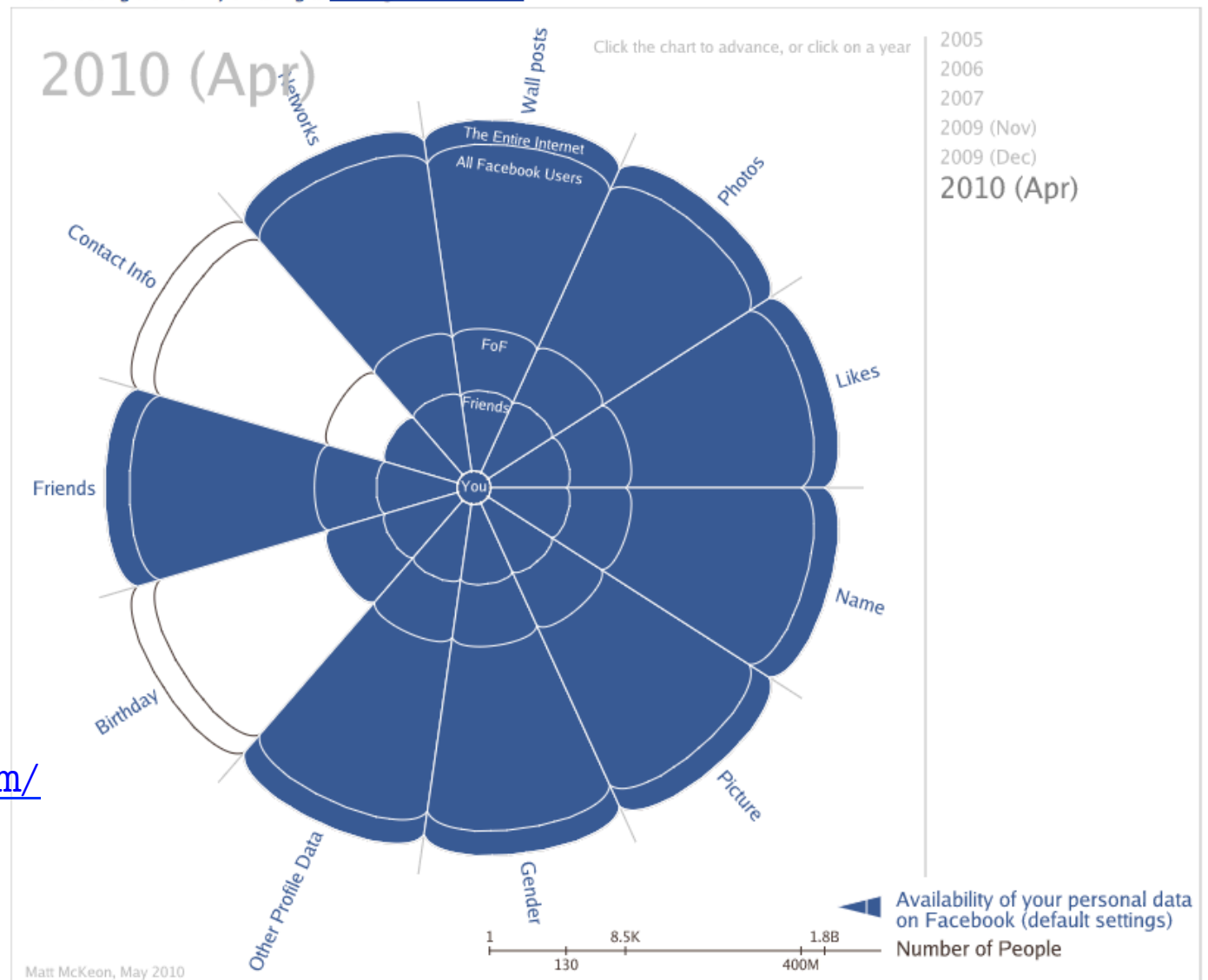
# Local structure is highly identifying

[Hay et al PVLDB 08]





# Sensitive values in social networks



[http://mattmckeon.com/  
facebook-privacy/](http://mattmckeon.com/facebook-privacy/)

# Sensitive values in social networks

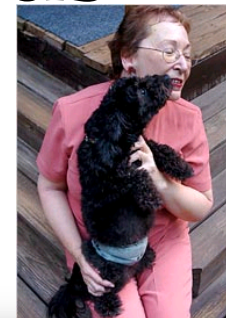
- Some people are privacy conscious (like you)
- Most people are lazy and keep the default privacy settings (i.e., no privacy)
- Can infer your sensitive attributes based on the sensitive attribute of public individuals ...

# Servers track your information ... and you are not anonymous

## A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.  
Published: August 9, 2006

✉ SIGN IN TO E  
THIS



## Why 'Anonymous' Data Sometimes Isn't

By Bruce Schneier ✉ 12.13.07

Last year, Netflix published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using.

The New York Times

Business Day  
Technology

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HE

## Marketers Can Glean Private Data on Facebook

A graphic for Facebook Ads featuring a blue speech bubble with quotation marks and silhouettes of three people.

**Facebook Ads**  
Reach the exact audience you want with relevant targeted ads.



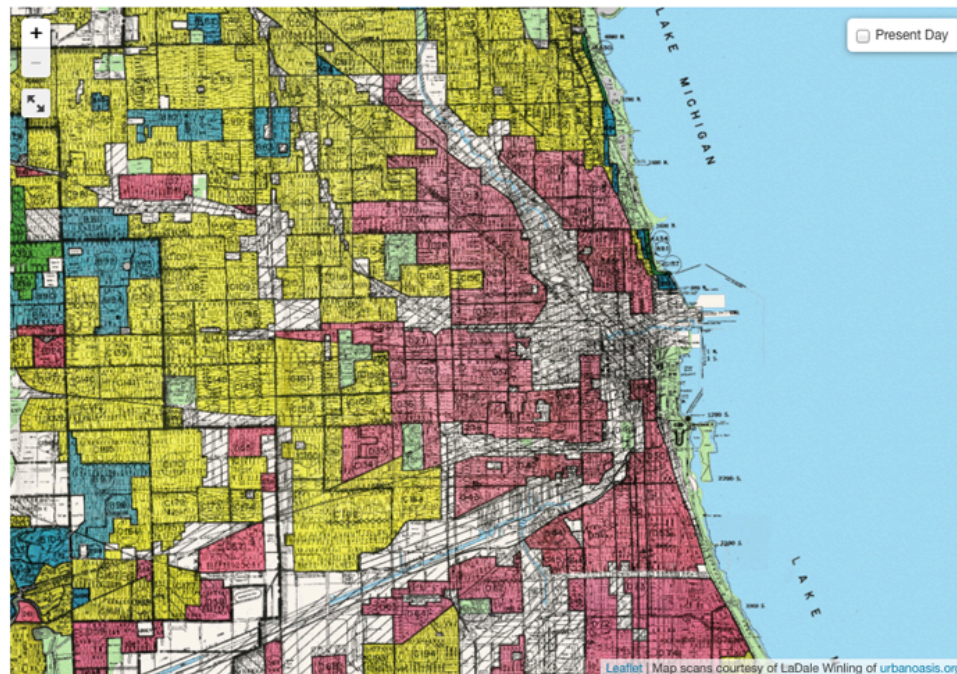
TECH | 2/16/2012 @ 11:02AM | 837,678 views

## How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

# Why care about privacy?

- **Redlining**: the practice of denying, or charging more for, services such as banking, insurance, access to health care, or even supermarkets, or denying jobs to residents in particular, often racially determined, areas.

## Explore Redlining in Chicago



A 1939 Home Owners' Loan Corporation "Residential Security Map" of Chicago shows discrimination against low-income and minority neighborhoods. The residents of the areas marked in red (representing "hazardous" real-estate markets) were denied FHA-backed mortgages. (Map development by Frankie Dintino)

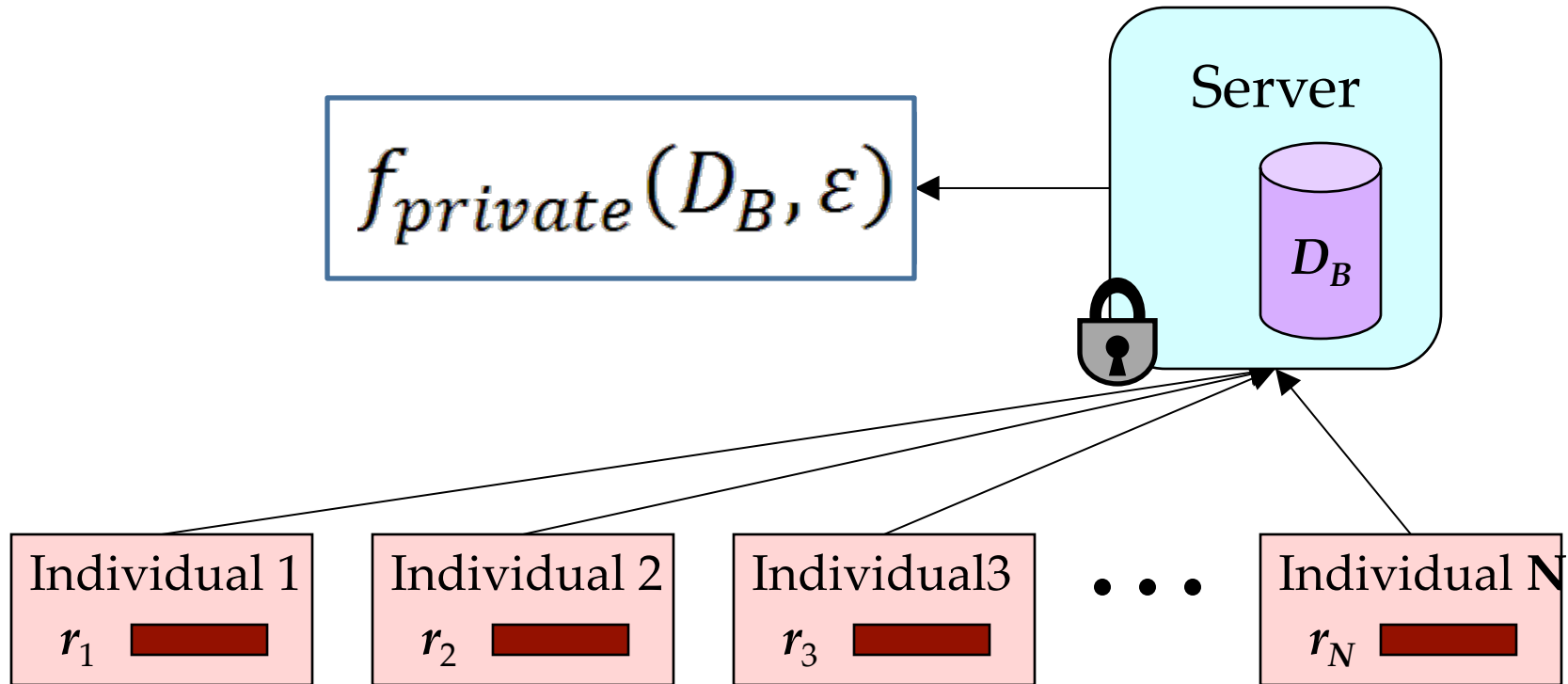
Can data analysis be done without breaching the privacy of individuals?



# Private data analysis problem

**Utility:**  $f_{private}$  approximates  $f$

**Privacy:** No breach about any individual



# Private data analysis examples

Application	Data Collector	Third Party (adversary)	Private Information	Function (utility)
Medical	Hospital	Epidemiologist	Disease	Correlation between disease and geography
Genome analysis	Hospital	Statistician/Researcher	Genome	Correlation between genome and disease
Advertising	Google/FB/Y!	Advertiser	Clicks/Browsing	Number of clicks on an ad by age/region/gender ...
Social Recommendations	Facebook	Another user	Friend links / profile	Recommend other users or ads to users based on social network
Location Services	Verizon/AT&T	Verizon/AT&T	Location	Local Search

# Private data analysis methods

- Bare Minimum protection: K-anonymity
- Ideal (state-of-the-art): Differential Privacy



# K-Anonymity

- If every row corresponds to one individual, then ...  
  
... every row should look like  $k-1$  other rows based on the *quasi-identifier* attributes

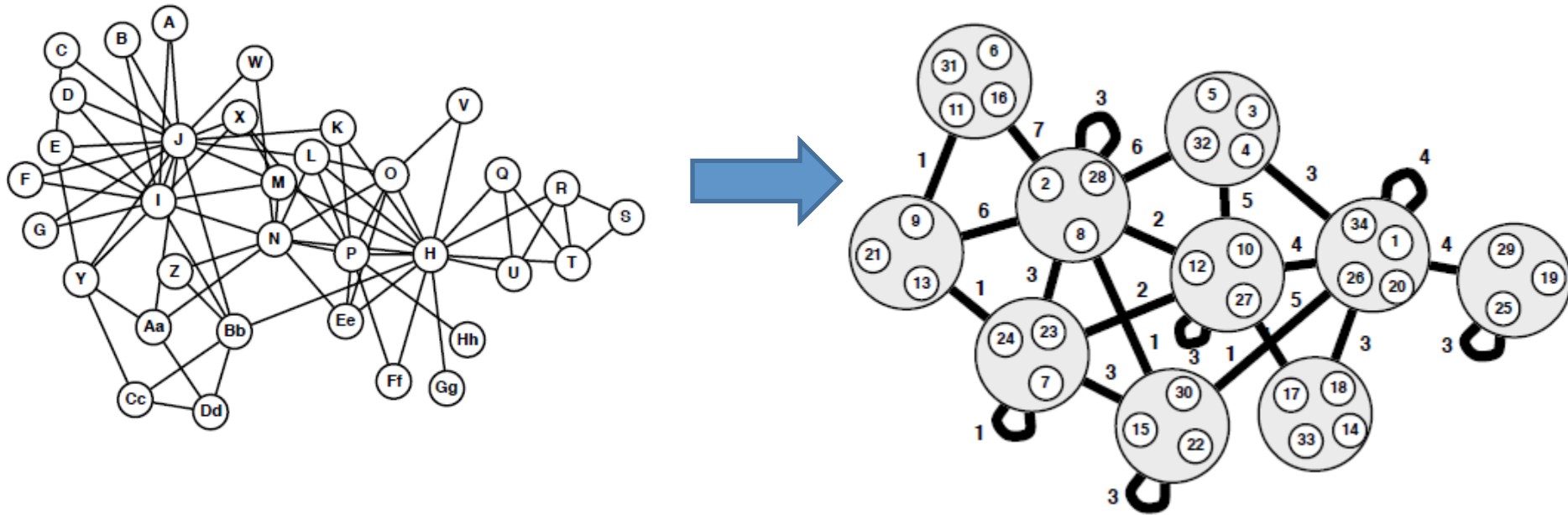
# K-Anonymity

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer



Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

# K-anonymity in graphs



# Problem: Homogeneity

Bob has Cancer

Name	Zip	Age	Nat.
Bob	13053	35	??

Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

# Problem: Composition

Zip Code	Age	Income	Disease
130**	[25-30]	$\geq 50k$	None
130**	[25-30]	$\geq 50k$	Stroke
130**	[25-30]	$\geq 50k$	Flu
130**	[23-30]	$\geq 50k$	Cancer
902**	[60-70]	$< 50k$	Flu
902**	[60-70]	$< 50k$	Stroke
902**	[60-70]	$< 50k$	Flu
902**	[60-70]	$< 50k$	Cancer

If Bob is in both datasets,  
then Bob has Stroke!

Zip Code	Age	Nationality	Disease
130**	$< 40$	*	Cold
130**	$< 40$	*	Stroke
130**	$< 40$	*	Rash
1485*	$\geq 40$	*	Cancer
1485*	$\geq 40$	*	Flu
1485*	$\geq 40$	*	Cancer

# Differential Privacy

- Consider two datasets
  - With Bob as one of the participants
  - Without Bob
- Answers are roughly the same whether or not Bob is in the data

# Differential Privacy

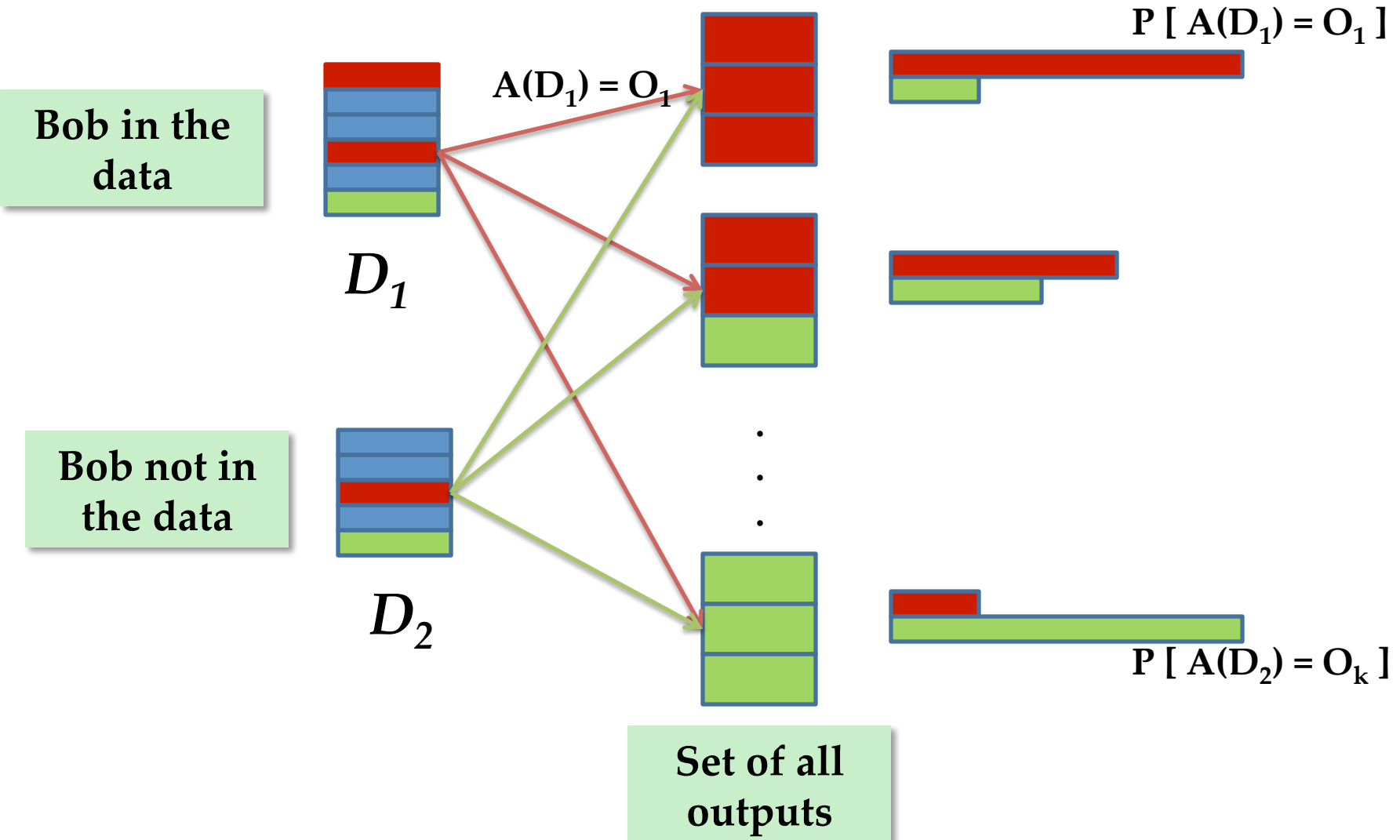
Algorithm A satisfies  $\epsilon$ -differential privacy if:

For **every pair** of *neighboring tables*  $D_1, D_2$

For **every output** O

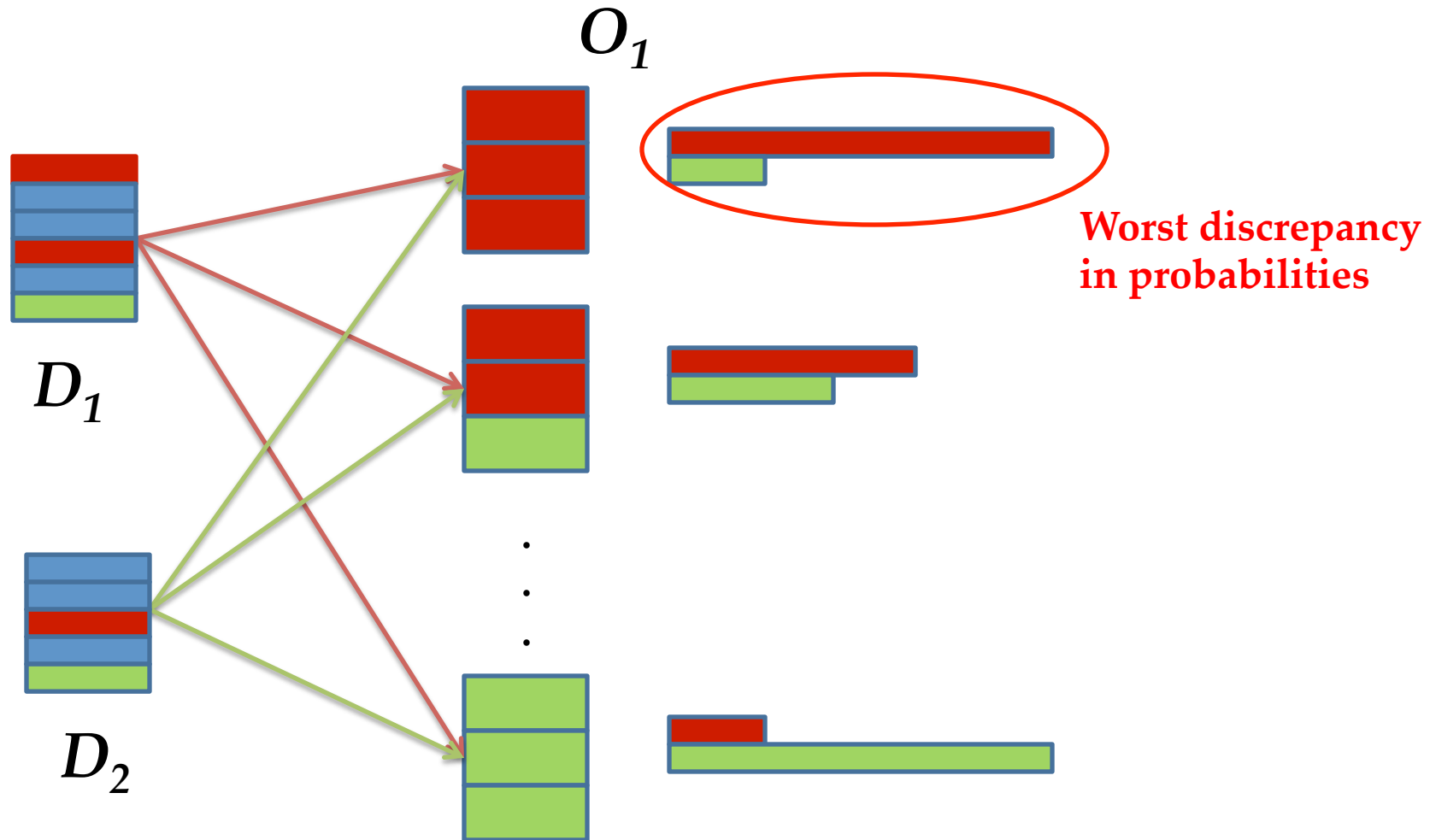
$$\Pr[A(D_1) = O] \leq e^\epsilon \Pr[A(D_2) = O]$$

# Meaning ...





# Meaning ...



# Privacy loss parameter $\epsilon$

Algorithm  $A$  satisfies  $\epsilon$ -differential privacy if:

For **every pair** of *neighboring tables*  $D_1, D_2$

For **every output**  $O$

$$\Pr[A(D_1) = O] \leq e^\epsilon \Pr[A(D_2) = O]$$

- Smaller the  $\epsilon$  more the privacy (and better the utility)

# Differential Privacy

Algorithm  $A$  satisfies  $\epsilon$ -differential privacy if:

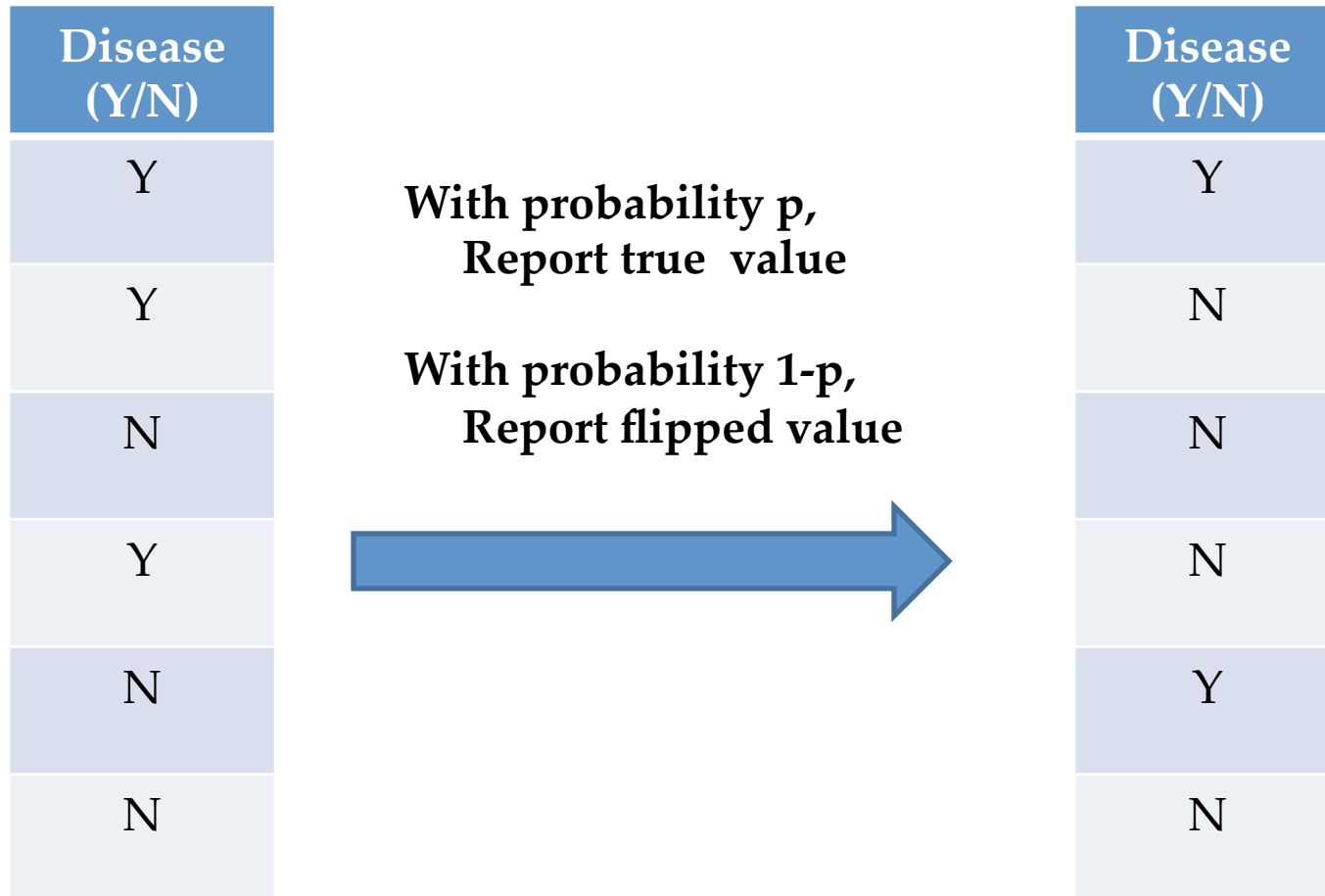
For **every pair** of *neighboring tables*  $D_1, D_2$

For **every output**  $O$

$$\Pr[A(D_1) = O] \leq e^\epsilon \Pr[A(D_2) = O]$$

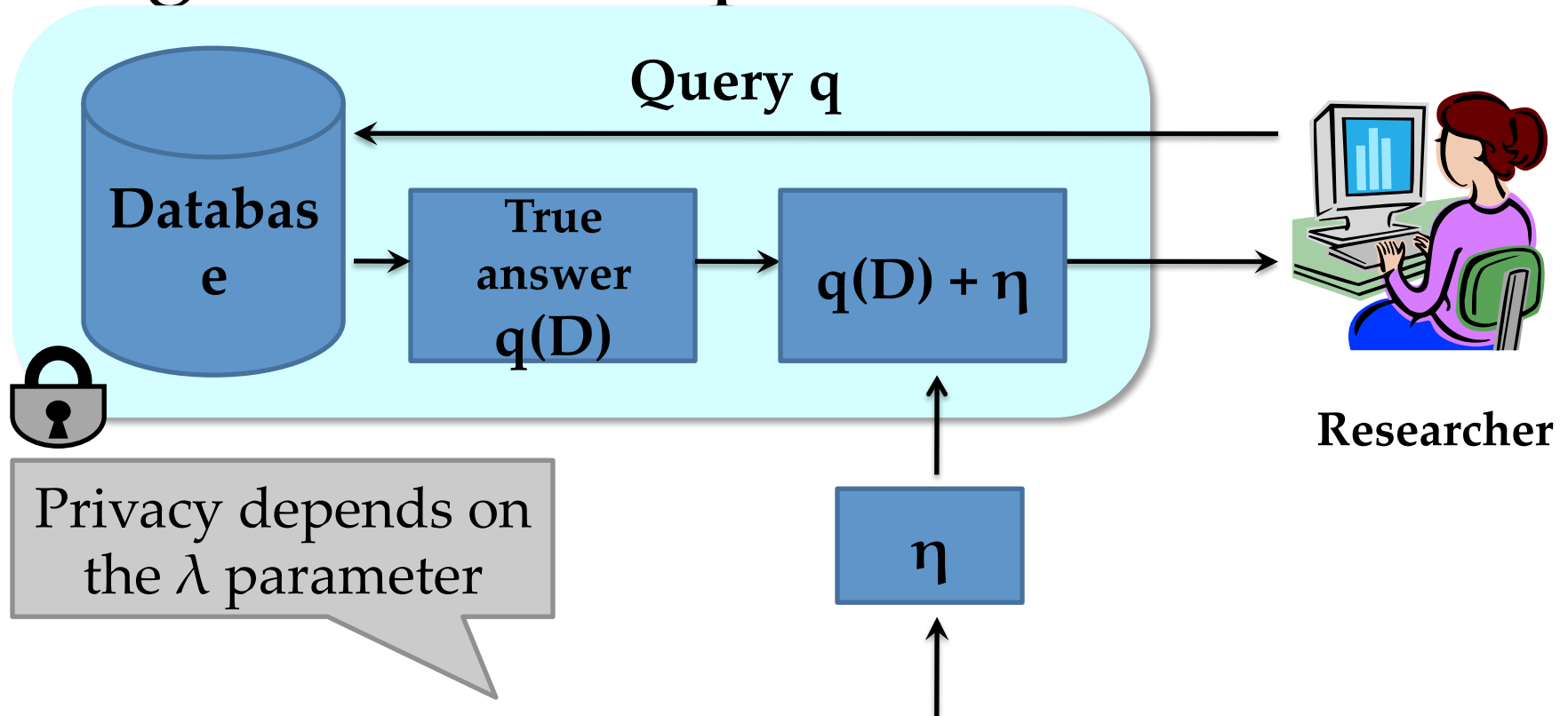
*what the adversary learns about an individual is the same even if the individual is not in the data (or lied about his/her value)*

# Algorithm 1: Randomized Response



*Can estimate the true proportion of Y in the data based on the perturbed values (since we know  $p$ )*

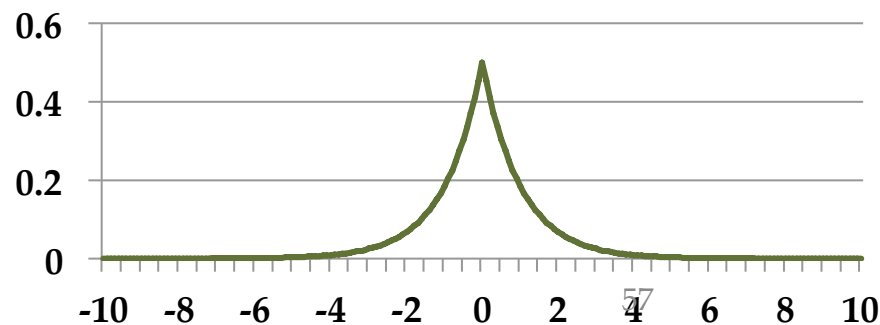
# Algorithm 2: Laplace Mechanism



$$h(\eta) \propto \exp(-\eta / \lambda)$$

Mean: 0,  
Variance:  $2 \lambda^2$

Laplace Distribution – Lap( $\lambda$ )



# Laplace Mechanism example

Qn: Release the histogram of admissions by diagnosis

Ans:

- Compute the true histogram
- Add noise to each count in the histogram using noise from  $\text{Lap}(1/\epsilon)$

*Noisy count is within  $\pm 1.38$  of true count  
for  $\epsilon = 1$*

# Composition

Qn: Release 2 histograms of admissions  
(a) by diagnosis, and (b) age

Ans:

- Compute the true histograms
- Add noise to each count in the histograms using noise from  $\text{Lap}(1/\epsilon)$

*Noisy counts are within  $\pm 1.38$  of true counts in both histograms ... but total privacy loss = 2*

# Differential Privacy summary

- Guarantees that the output of an analysis does not change much whether or not an individual is in the data
- Very active area of research
- Many sophisticated algorithms for a variety of analyses (*see my [other course](#)*)
  - Used by the US Census to release data



# Summary

- “Data-driven” revolution has transformed many fields ...
- ... but need to address the privacy problem
- Tools like differential privacy can foster ‘safe’ data collection, analysis and data sharing.