ICME Summer Workshop
Fundamentals of Data Science

# Data Privacy and Ethics

Day 1 of 2

Prof. Johan Ugander
jugander@stanford.edu

Adapted from **MS&E 234** at Stanford,
see homepage for literature references and more:
http://msande234.stanford.edu/

ICME
Stanford | Institute for Computational & Mathematical Engineering

# What's in focus?

- **In focus:**
  - Machine learning-based privacy attacks and defenses
  - "Could/should dilemmas" in data products (search engines, recsys, personalization, adtech)
  - Evaluation of data products: Why experimentation?
  - Privacy regulation (EU GDPR) and effects on industry

- **Will come up, but <u>not</u> in focus:**
  - Cryptography & cybersecurity (CS)
  - Misinformation, fake news (Comm)
  - Privacy law (Law)

# Schedule

**Monday**

- **Part 1 - Introduction, Digital Exhaust**
  - 1:00p-2:05p Lecture
  - 2:05p-2:10p Break

- **Part 2 - Case study: “People You May Know”**
  - 2:10-3:00p In/out of breakout rooms
  - 3:00p-3:05p Break

- **Part 3 - Machine learning with relational data**
  - 3:05p-3:20p Jupyter notebook tour
  - 3:20p-4:00p Lecture

# Schedule

**Tuesday**

- **Part 4 - Differential Privacy**
  - 1:00p-2:05p Lecture
  - 2:05p-2:10p Break

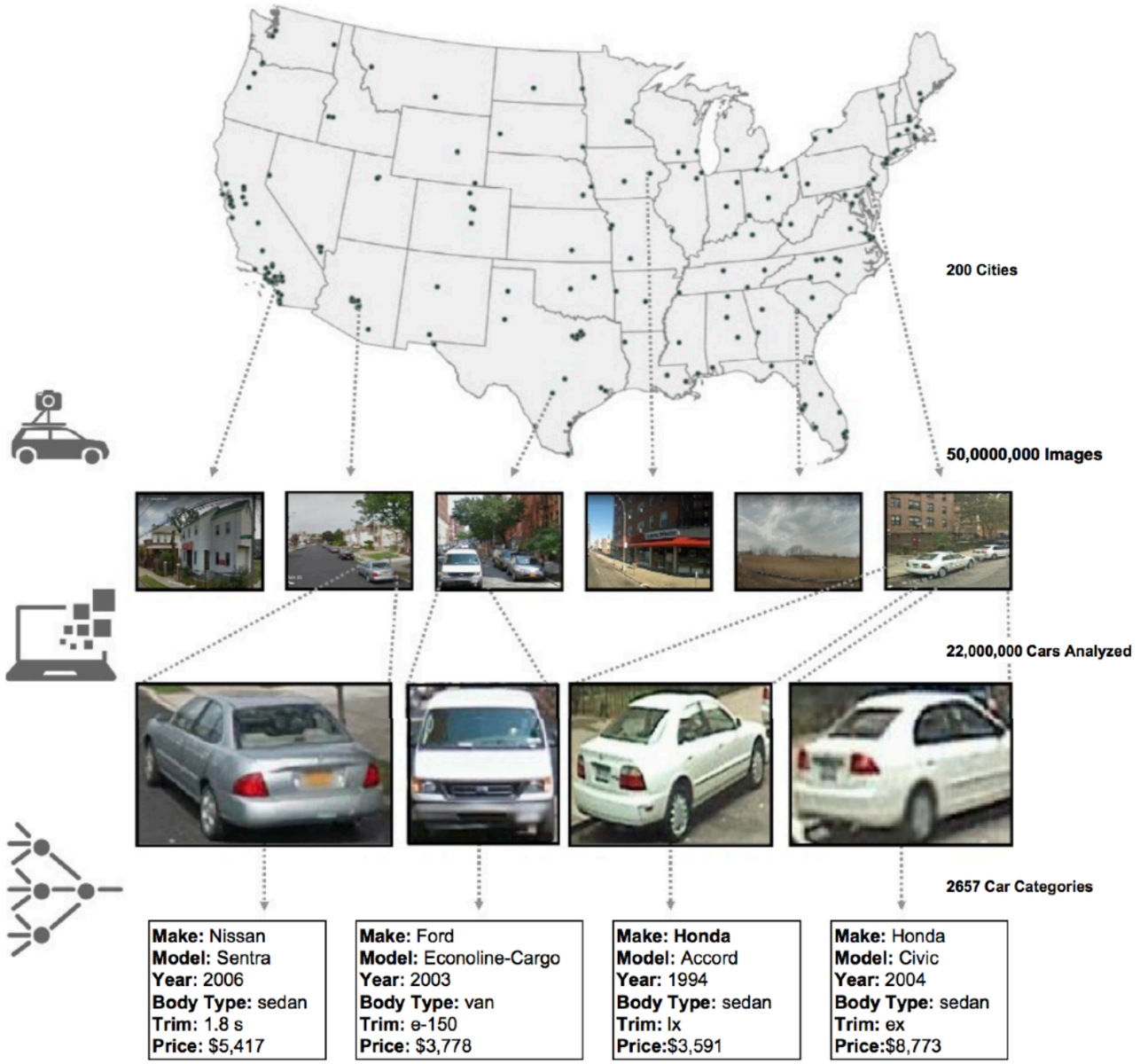
- **Part 5 - Discussion: transparency & public records**
  - 2:10-3:00p Voting on examples, discussion
  - 3:00p-3:05p Break

- **Part 6 - GDPR, Regulation**
  - 3:05p-4:00p Lecture

# Questions?

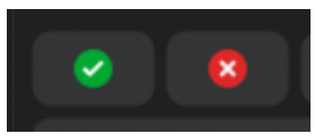
# Digital exhaust

NETFLIX

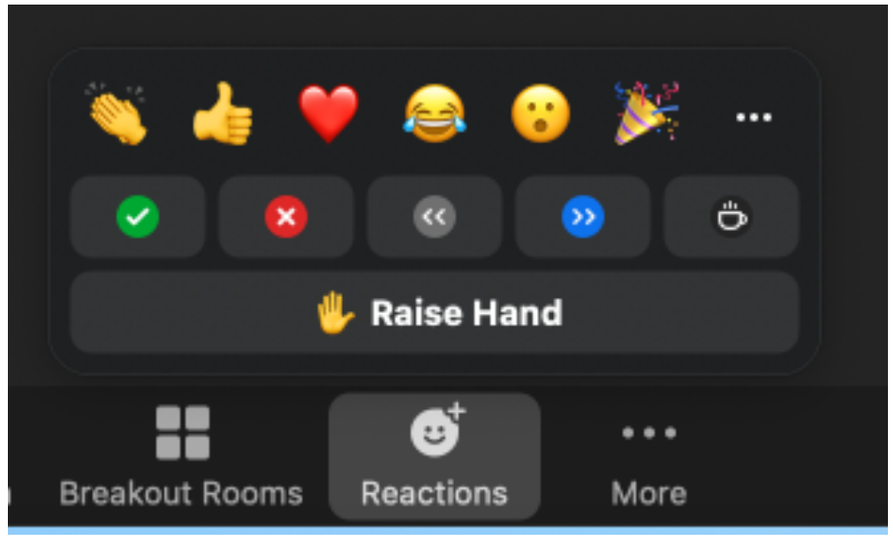



Gebru et al. 2017

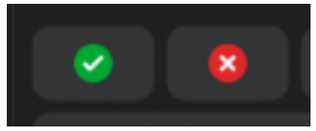
# Identity in high dimensions

- In 2022, how many of you have used your credit card …
  - At an airport? 
  - At a Starbucks?
  - At an IKEA?

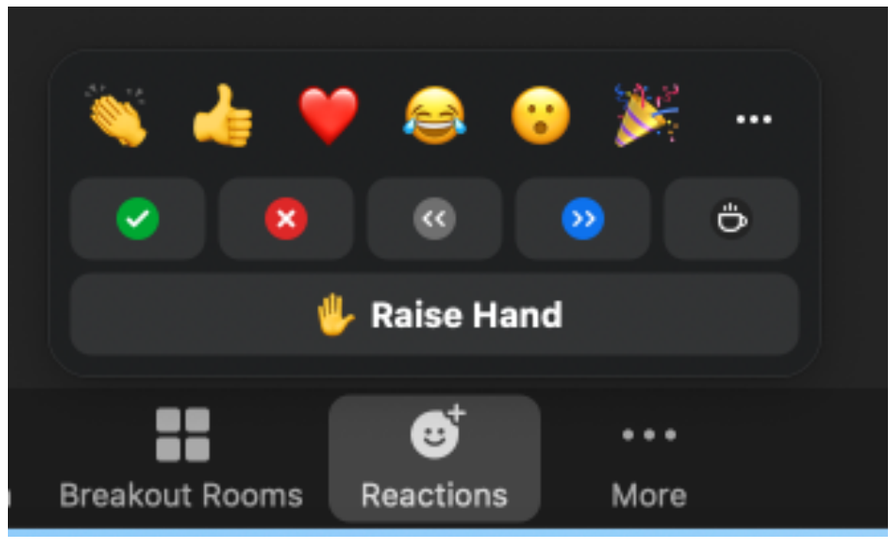
**Zoom reactions**



# Identity in high dimensions

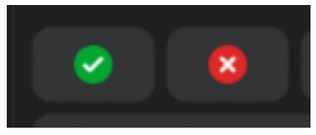
- In 2022, how many of you have used your credit card …
  - At an airport?
  - At a Starbucks? 
  - At an IKEA?

**Zoom reactions**

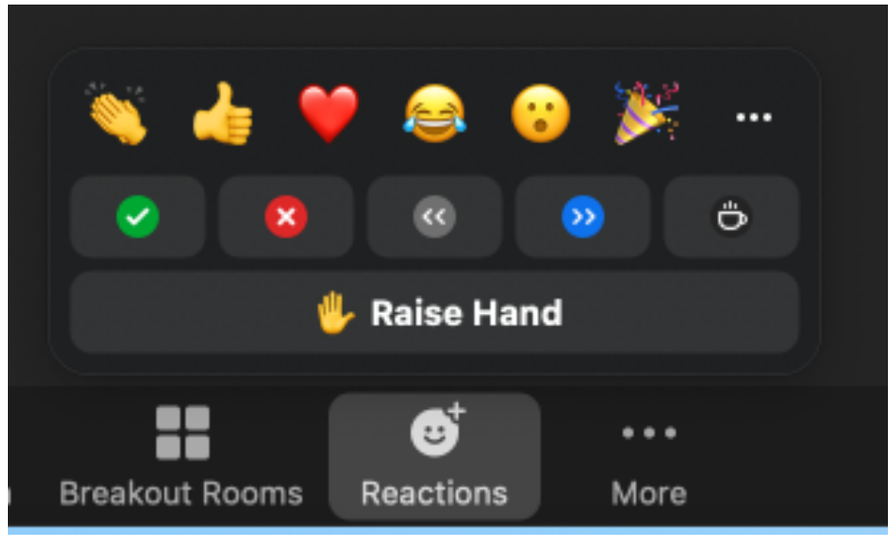


# Identity in high dimensions

- In 2022, how many of you have used your credit card …
  - At an airport?
  - At a Starbucks?
  - At an IKEA?



**Zoom reactions**



# Bits of information

- World population: ~7 billion.
- Log2(7 billion) = 33 bits
- Lower bound of 33 binary questions needed to identify.

# Bits of information

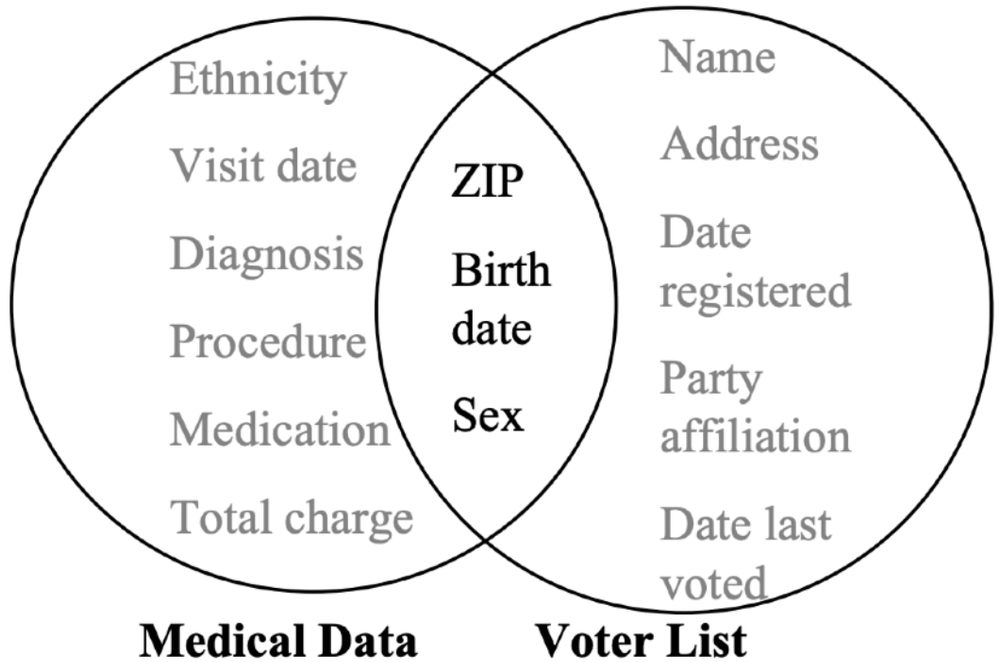
- World population: ~7 billion.
- Log2(7 billion) = 33 bits
- Lower bound of 33 binary questions needed to identify.

- Some data types, e.g. a location, contains many(!) bits.

# Bits of information

- World population: ~7 billion.
- Log2(7 billion) = 33 bits
- Lower bound of 33 binary questions needed to identify.

- Some data types, e.g. a location, contains many(!) bits.




- Sweeney 2000: 87% of US population: unique DOB/sex/postal code.

# Anonymity in data

- Golle & Partridge 2009 study of (home, work) location pairs

- Longitudinal Employer-Household Dynamics (LEHD)
  - U.S. Census Bureau program to compile information about where people work and live (towards reasonable goals)
    - County
    - Census tract (~ ZIP code)
  - 103 million workers from 42 states.

- "[with] approximate locations of an individual's **home and workplace** … then the median size of the individual's **anonymity set** in the U.S. working population is **1**, **21** and **34,980**, for locations known at the granularity of a **census block**, **census track,** and **county** respectively."

# Definition: k-anonymity

- **Definition (k-anonymity)**: A release of data is said to have the ***k*-anonymity property** if the information for each person in the data cannot be distinguished from at least **k-1** other individuals in the data.

# Home, work locations

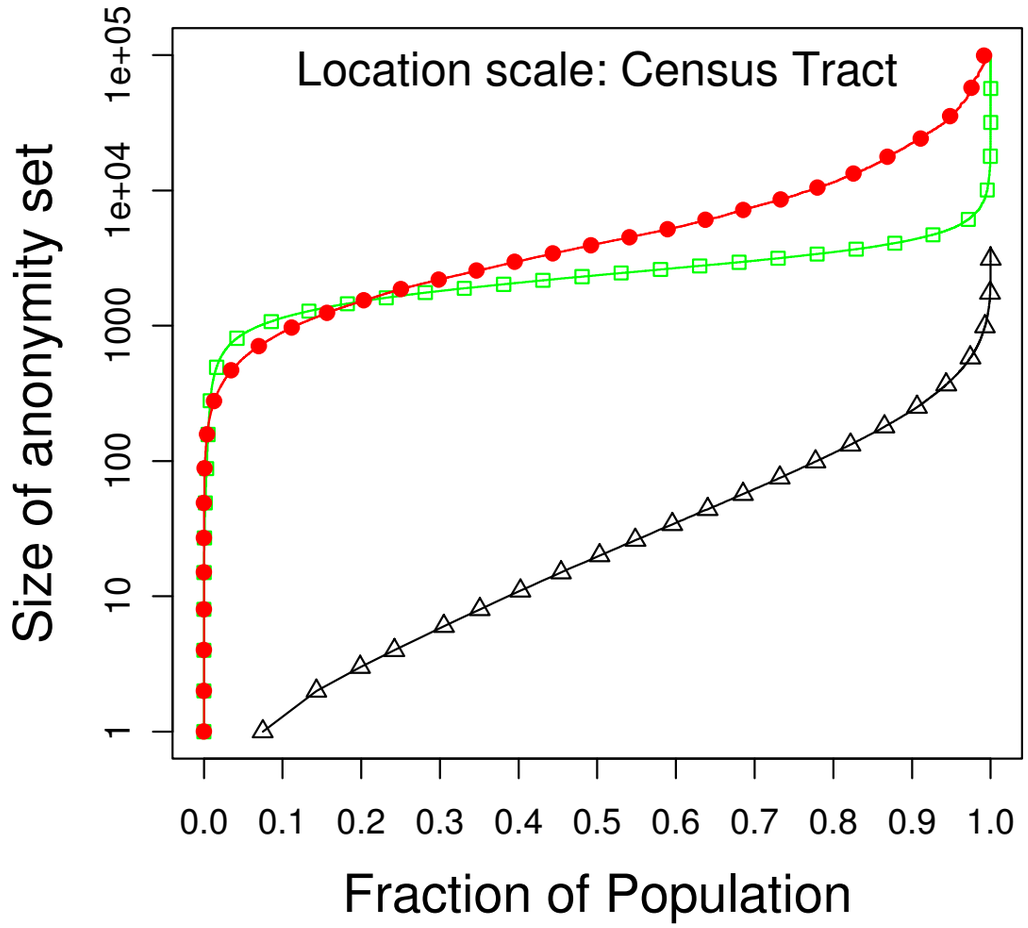



**Fig. 1.** Size of anonymity set under disclosure of work location (red circles), home location (green squares) or both (black triangles). Location granularity is either census tract (left graph) or county (right graph). Note the different scales on the Y-axes.

# Home, work locations

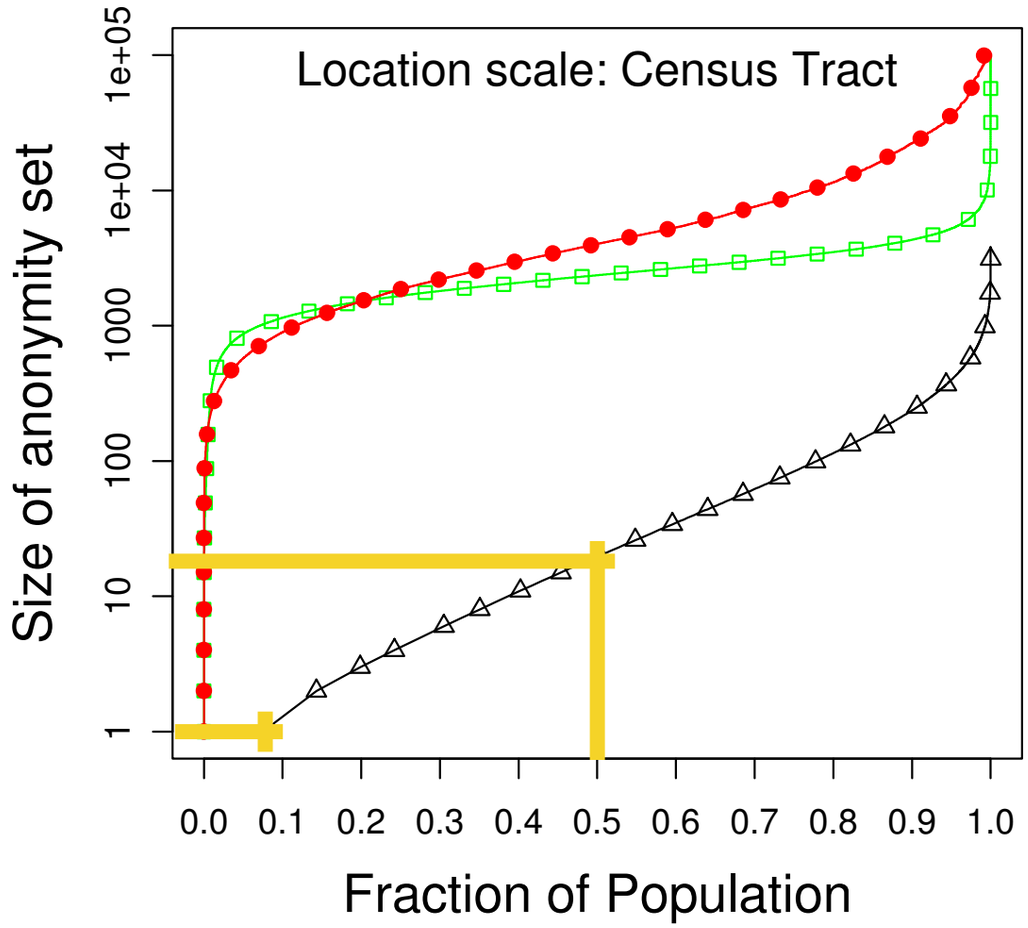



**Fig. 1.** Size of anonymity set under disclosure of work location (red circles), home location (green squares) or both (black triangles). Location granularity is either census tract (left graph) or county (right graph). Note the different scales on the Y-axes.

**Left plot: 7% have k-anonymity of 1, 50% of 21.**

# Home, work locations

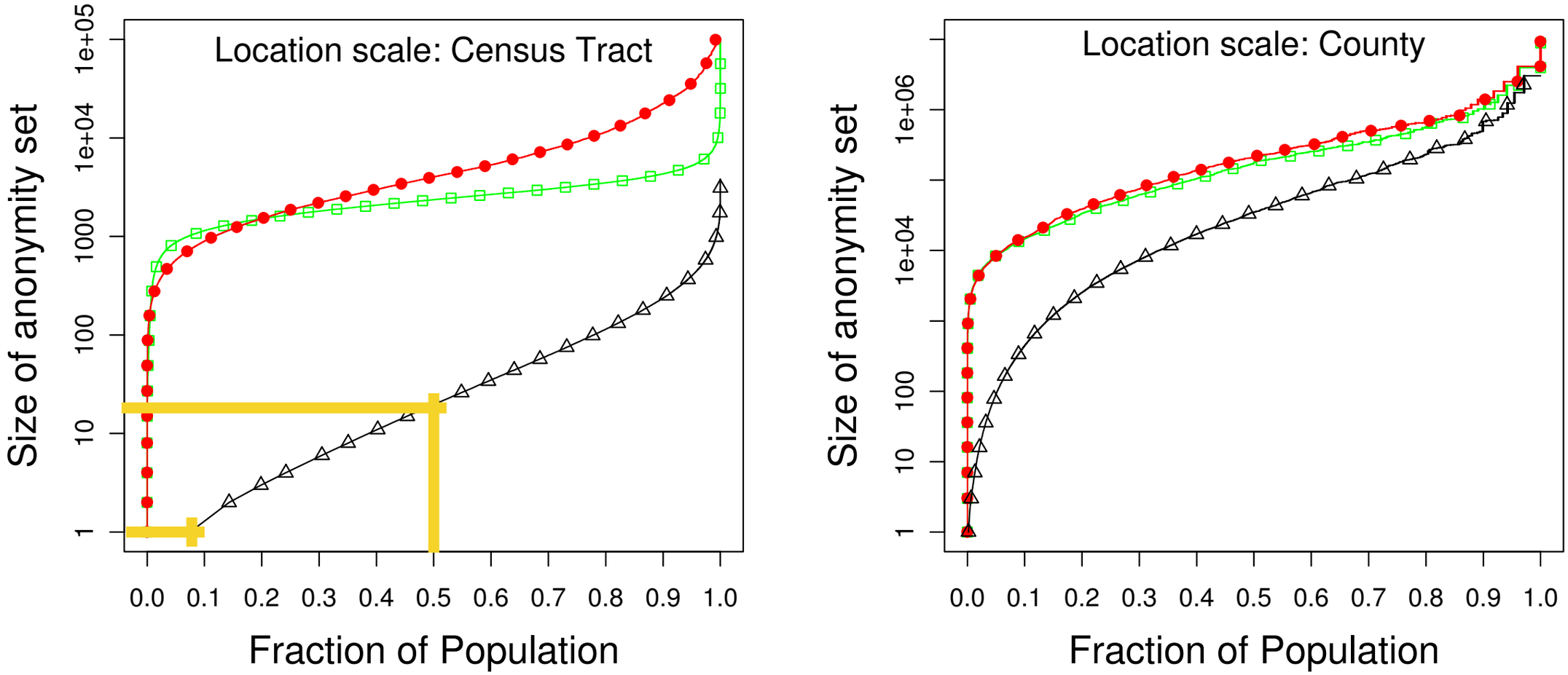



**Fig. 1.** Size of anonymity set under disclosure of work location (red circles), home location (green squares) or both (black triangles). Location granularity is either census tract (left graph) or county (right graph). Note the different scales on the Y-axes.

**Left plot: 7% have k-anonymity of 1, 50% of 21.**

# De-anonymization & linkage attacks

- Linkage attack: identify enough bits to find a person.
  - **Specific person**: Sweeney 2000
    - Medical data & voter data
  - **Some person:** Narayanan & Shmatikov 2008
    - Netflix data & IMDB data

- Statistical linkage attack: data contains traits X about a population, you want to know about Y for that population. Find other data that correlates X & Y (for that population!).

- Sometimes the risk is releasing correlation between X & Y.
- Sometimes the risk is releasing X.

# Linkage “attack”: One weird trick

- Examples:
  - U.S. Census (1990)
  - AOL search log (2006)
  - Netflix Prize (2007)
  - Facebook Beacon (2007-2009)
  - Location data, cell phones (2013)

- Others:
  - Race from names (Chang et al.)
  - Wealth/air quality from cars (Gebru et al.)
  - Revenue from satellite photos (Orbital Insights)

# Netflix

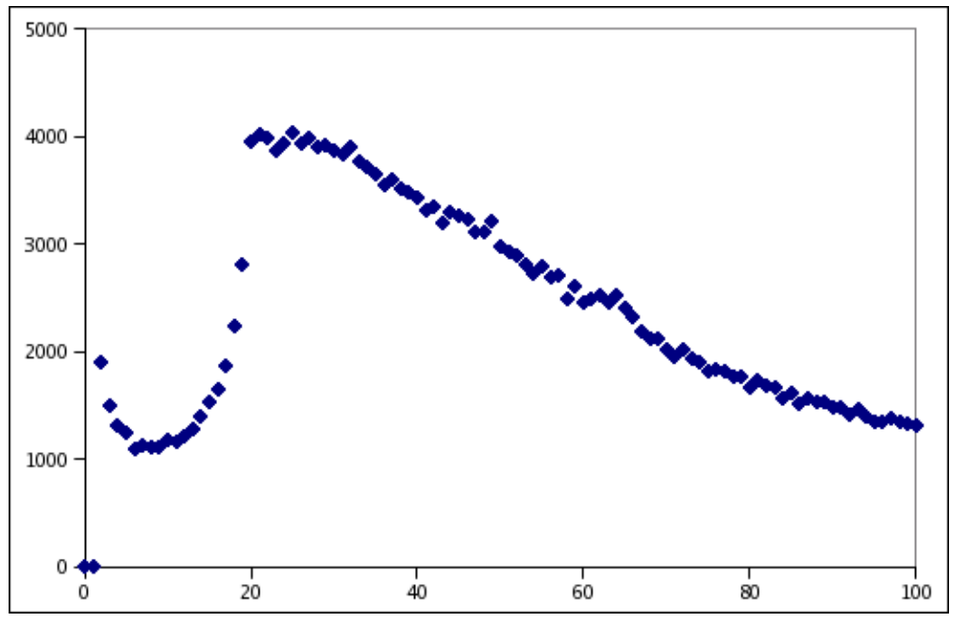
- In October 2006, released real movie ratings of 500,000 subscribers
- 10% of all Netflix users in 2005
- Names removed, data “perturbed”

- 500,000 users
- 17,000 movies – high dim!
- Avg 214 dated ratings/user

| | Movie 1 | Movie 2 | Movie 3 | … … |
|---|---|---|---|---|
| Alice | Rating/time | Rating/time | Rating/time | …… |
| Bob | | | | |
| Charles | | | | |
| David | | | | |
| Evelyn | | | | |
| … | | | | |

# Netflix

- In October 2006, released real movie ratings of 500,000 subscribers
- 10% of all Netflix users in 2005
- Names removed, data “perturbed”




**# of users with ≥ x ratings**

- 500,000 users
- 17,000 movies – high dim!
- Avg 214 dated ratings/user

| | Movie 1 | Movie 2 | Movie 3 | … … |
|---|---|---|---|---|
| Alice | Rating/time | Rating/time | Rating/time | …… |
| Bob | | | | |
| Charles | | | | |
| David | | | | |
| Evelyn | | | | |
| … | | | | |

# Netflix

- x-axis: similarity
- y-axis: fraction of users

- Curse of dimensionality:
  For 90% of records, there isn't a single other person more than 30% similar.

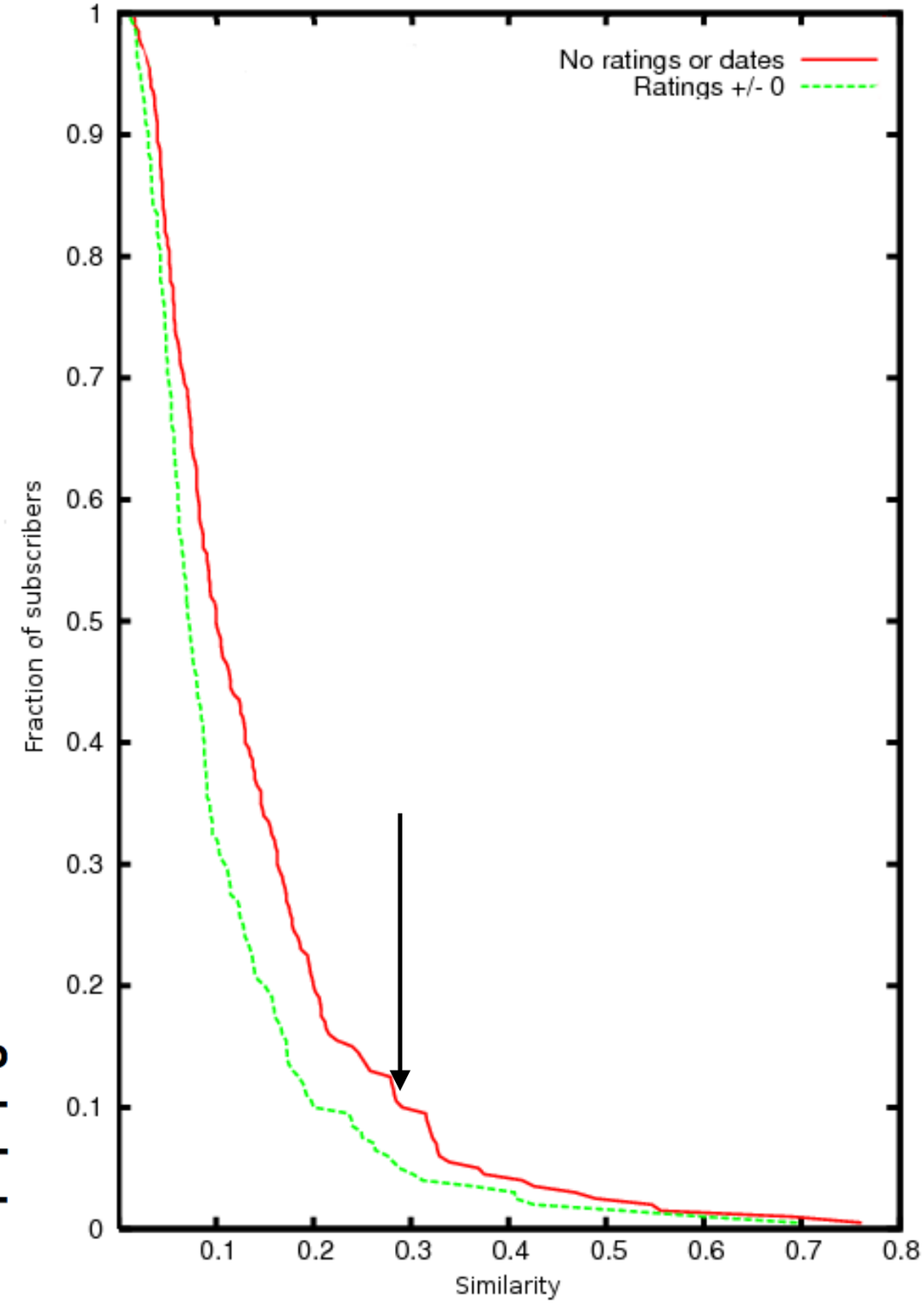



**Figure 1. X-axis ($x$) is the similarity to the "neighbor" with the highest similarity score; Y-axis is the fraction of subscribers whose nearest-neighbor similarity is at least $x$.**

# Netflix

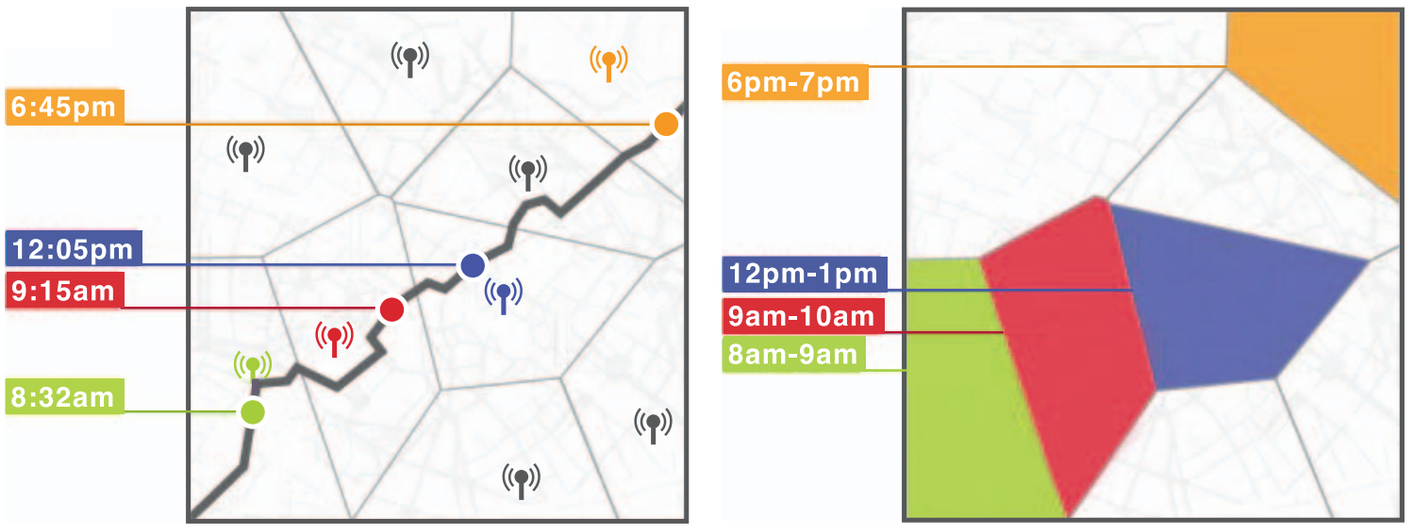
- How many does the attacker need to know to identify her target's record in the dataset?

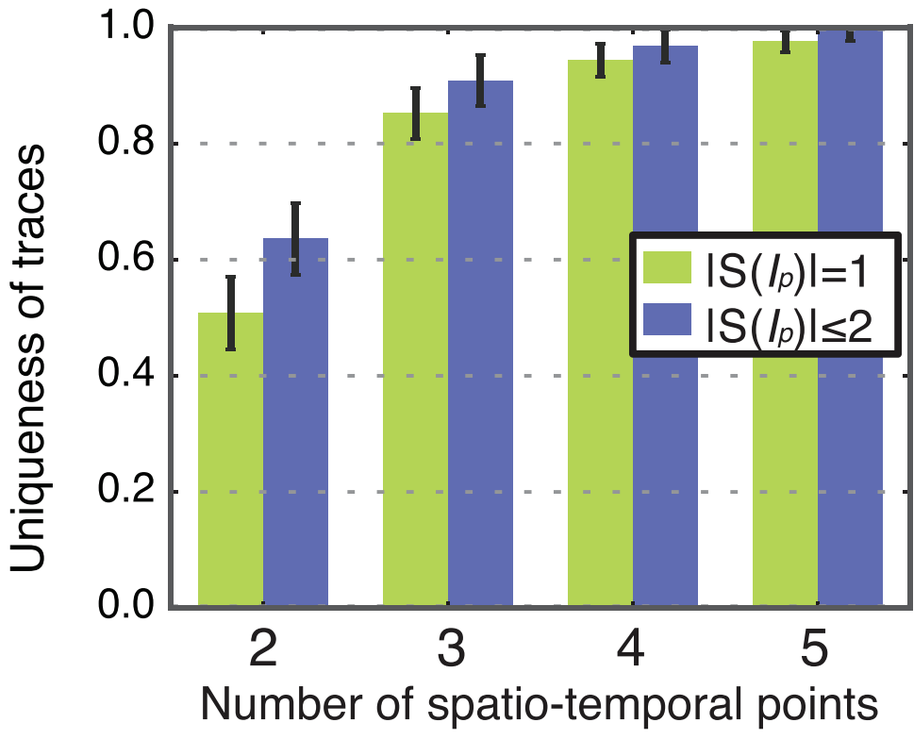
- On average:
  - **Two movies** is enough to reduce to 8 candidates
  - **Four movies** is enough for uniqueness

- Driven by the long tail, not "Star Wars"
- Temporal sequence also provides rich information.

# “Unique in the crowd”

- Timestamped cell phone tower data.
- 95% of users uniquely identified by four (tower, time) data points.
- Detailed study of effects of coarsening =>
  Even coarse data provides little anonymity.

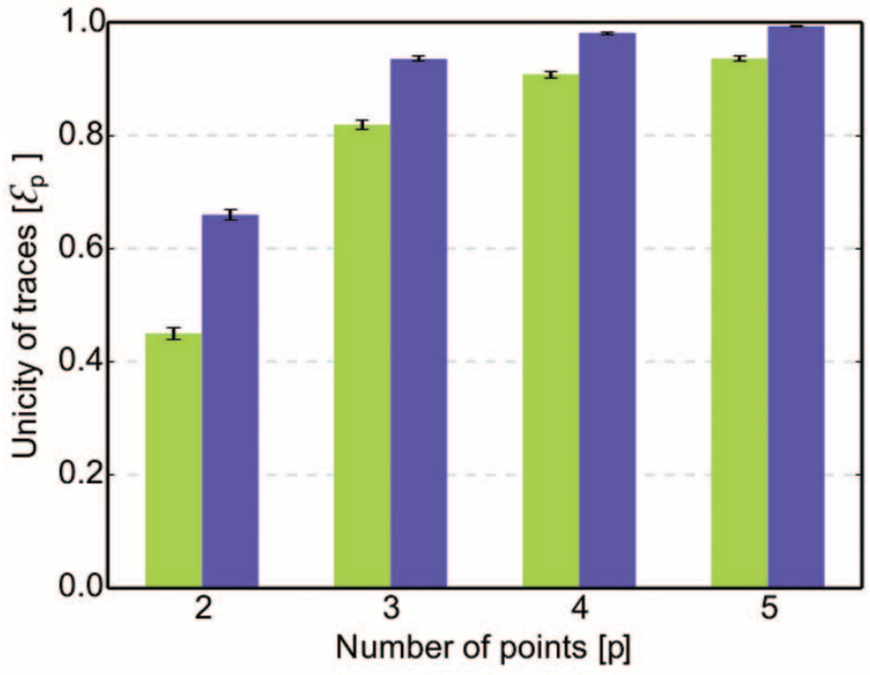


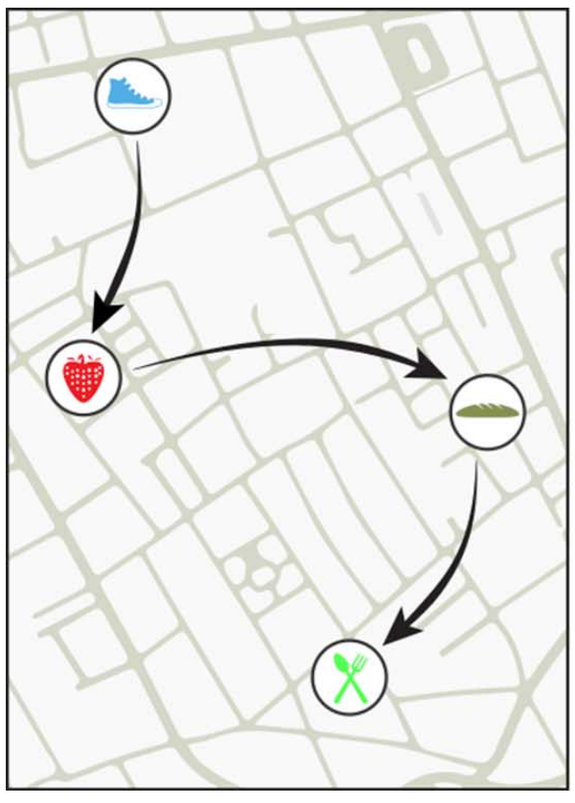



# “Unique in the shopping mall”

- Timestamped credit card purchase data.

- Green bars: space/time
- Blue bars: space/time/price



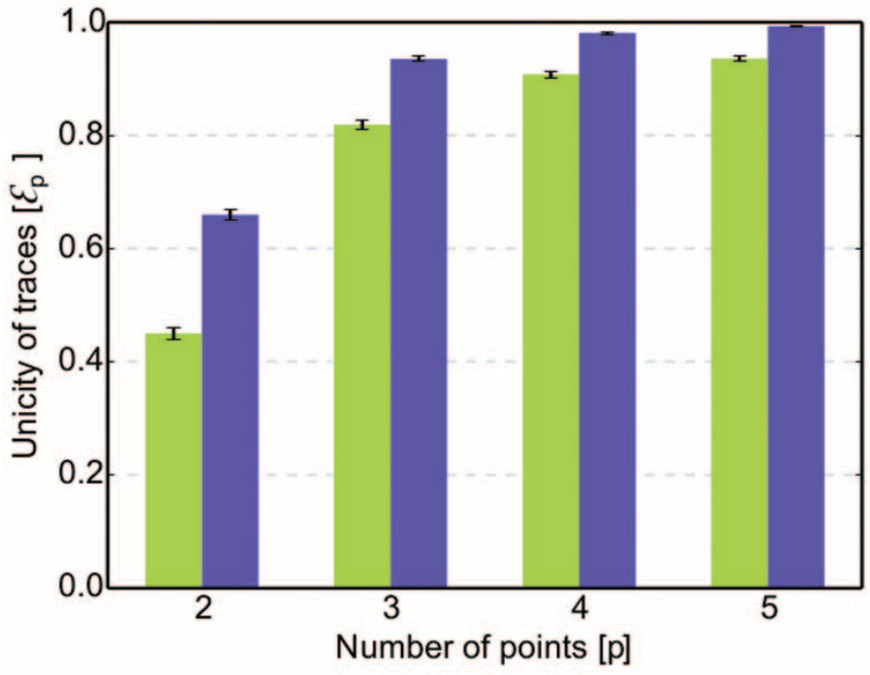



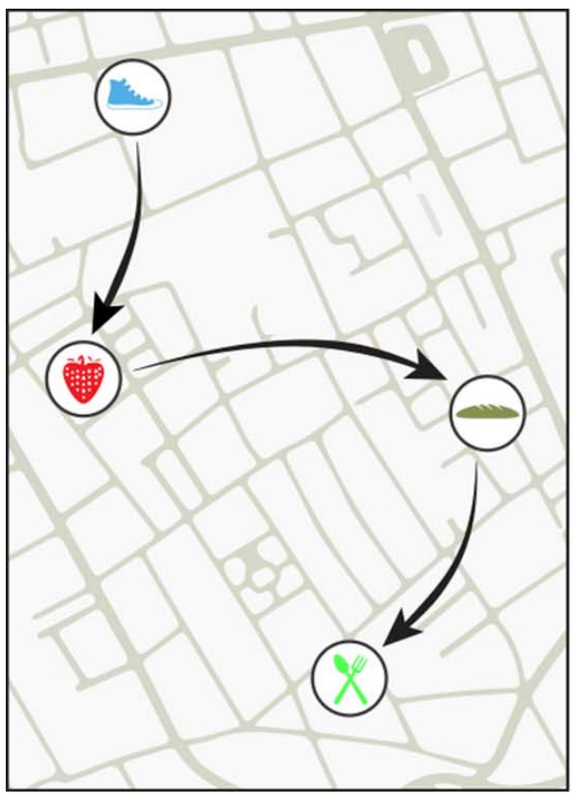
| shop | user_id | time | price | price_bin |
|---|---|---|---|---|
| | 7abc1a23 | 09/23 | $97.30 | $49 – $146 |
| | 7abc1a23 | 09/23 | $15.13 | $5 – $16 |
| | 3092fc10 | 09/23 | $43.78 | $16 – $49 |
| | 7abc1a23 | 09/23 | $4.33 | $2 – $5 |
| | 4c7af72a | 09/23 | $12.29 | $5 – $16 |
| | 89c0829c | 09/24 | $3.66 | $2 – $5 |
| | 7abc1a23 | 09/24 | $35.81 | $16 – $49 |

# “Unique in the shopping mall”

- Timestamped credit card purchase data.

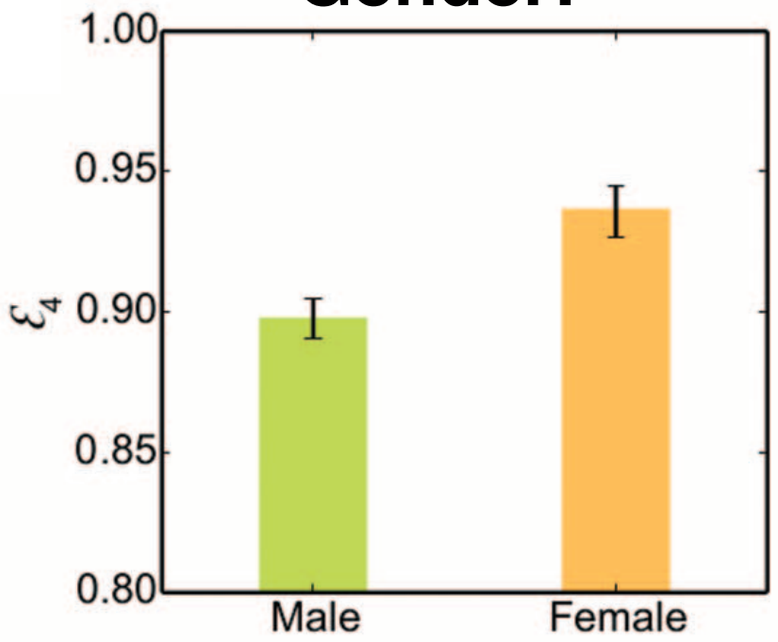
- Green bars: space/time
- Blue bars: space/time/price



| shop | user_id | time | price | price_bin |
|---|---|---|---|---|
| | 7abc1a23 | 09/23 | $97.30 | $49 – $146 |
| | 7abc1a23 | 09/23 | $15.13 | $5 – $16 |
| | 3092fc10 | 09/23 | $43.78 | $16 – $49 |
| | 7abc1a23 | 09/23 | $4.33 | $2 – $5 |
| | 4c7af72a | 09/23 | $12.29 | $5 – $16 |
| | 89c0829c | 09/24 | $3.66 | $2 – $5 |
| | 7abc1a23 | 09/24 | $35.81 | $16 – $49 |

## Gender:



Harvard Business Review

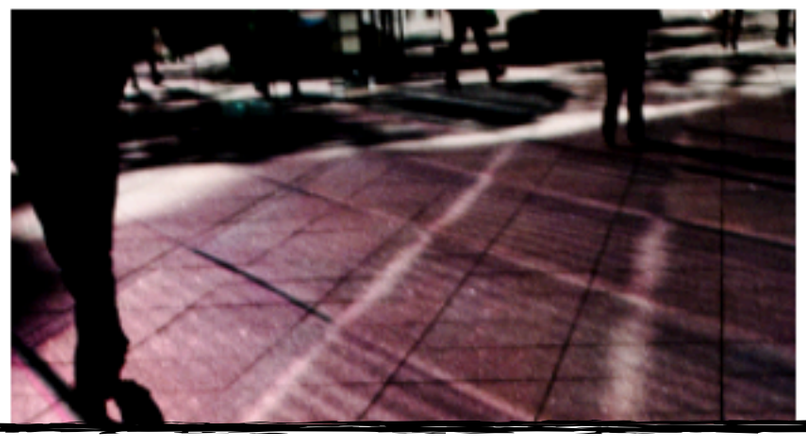
REGULATION

# There's No Such Thing as Anonymous Data

by Scott Berinato

FEBRUARY 09, 2015





About a decade ago, a hacker said to me, flatly, "Assume every card in your wallet is compromised, and proceed accordingly." He was right. Consumers have adapted to a

# What your ISP knows

- (Reed et al. 2018) Encrypted (!) data traffic:

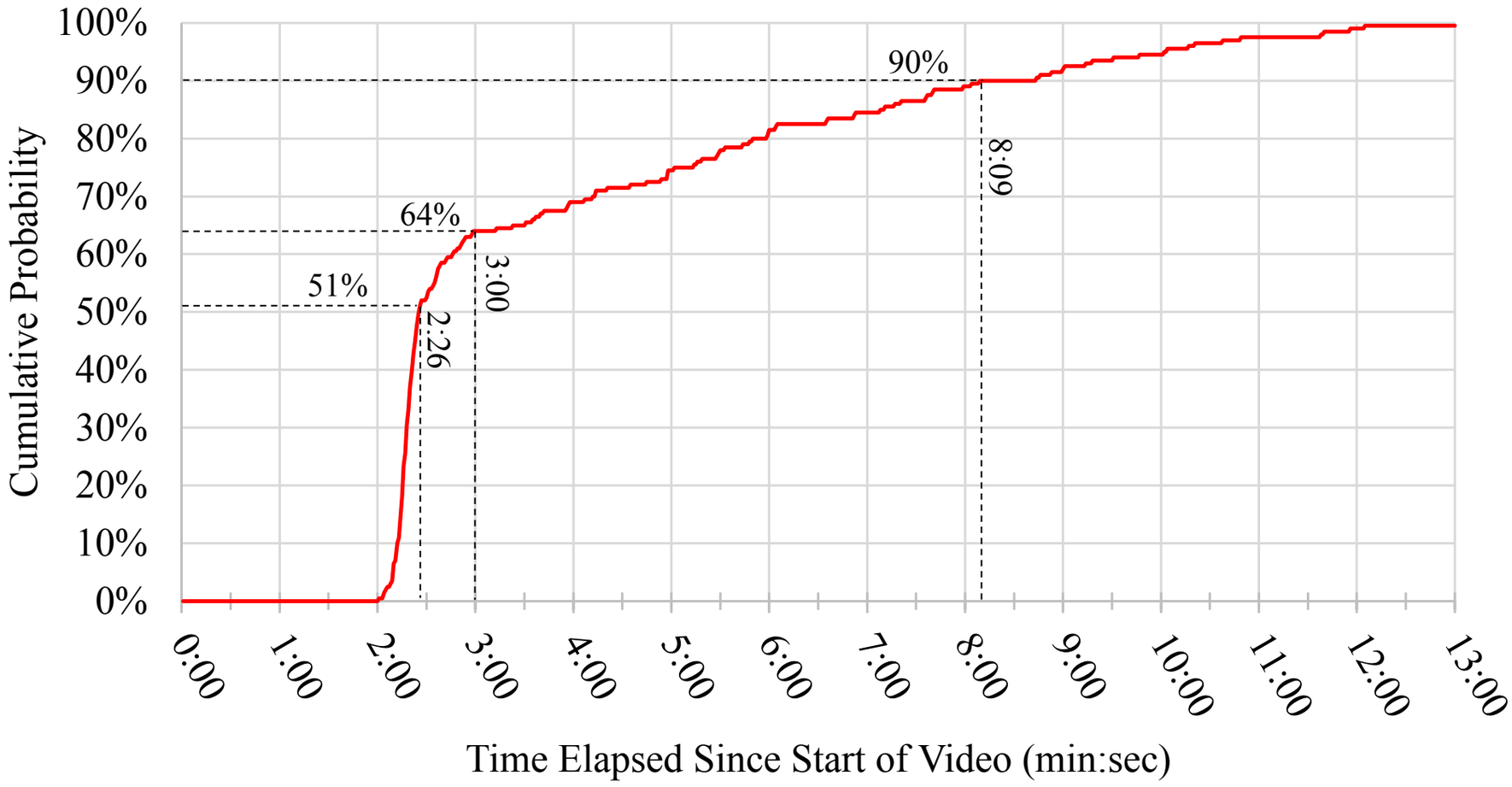



**Figure 4: Cumulative probability of identifying a video before a specified amount of time has elapsed.**

# Beyond uniqueness

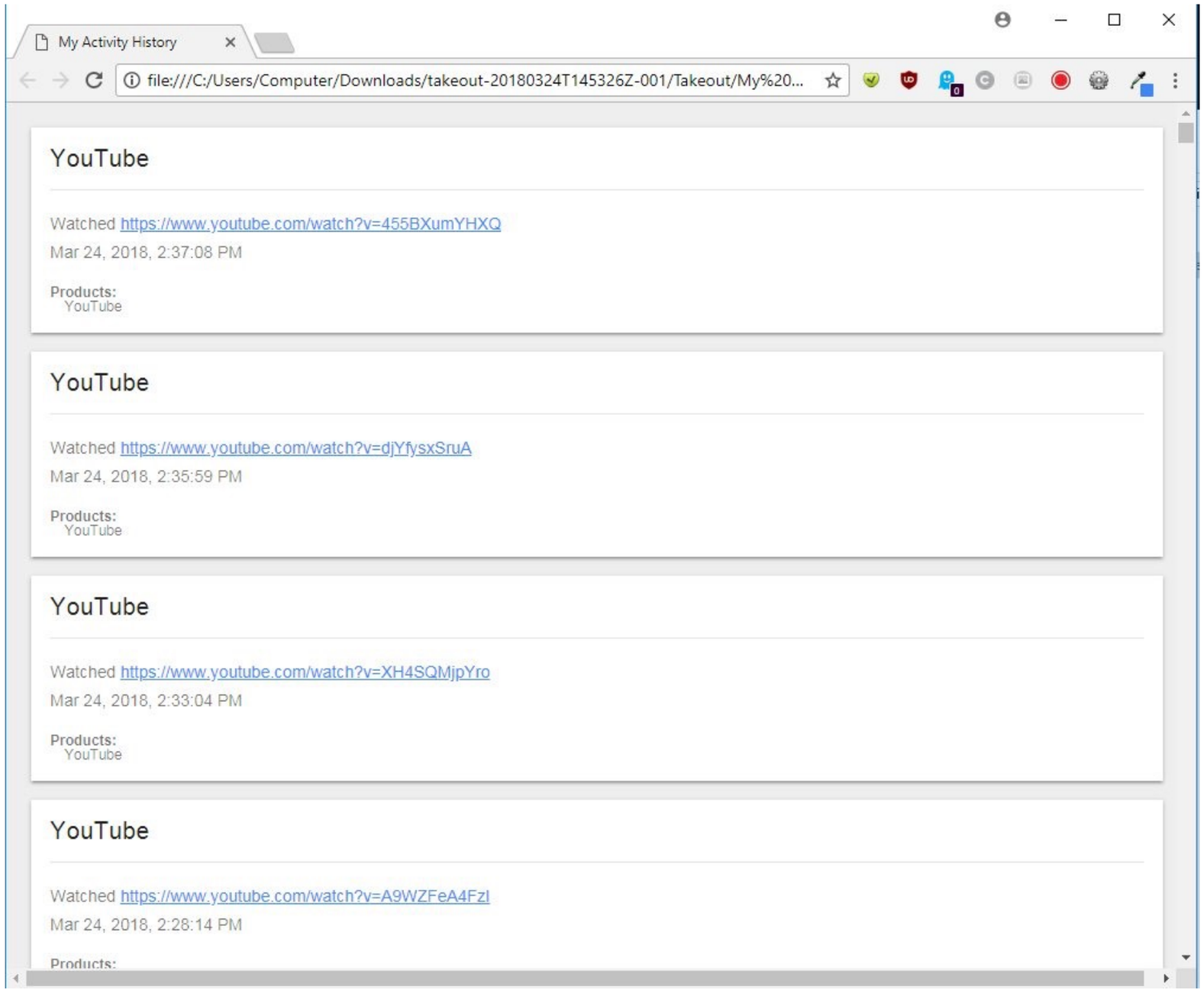
- ISP example is not about uniqueness.

- Statistical linkage attack: data contains traits X about a population, you want to know about Y for that population. Find other data that correlates X & Y (for that population!).

- ISP example: ISP doesn't know much about you on paper. But they know what movies you watch. Scrape IMDB, or social network survey data for demographics of people who watch X and they obtain a very good estimate of your demographics.

# Behavioral data is high dim!

- Youtube videos watched (from twitter thread by @iamdylancurran):



# A useful tool for missing data

- Chang et al. 2010
- Facebook didn't/doesn't collect information about race.
- But they have names.
- Census has names.

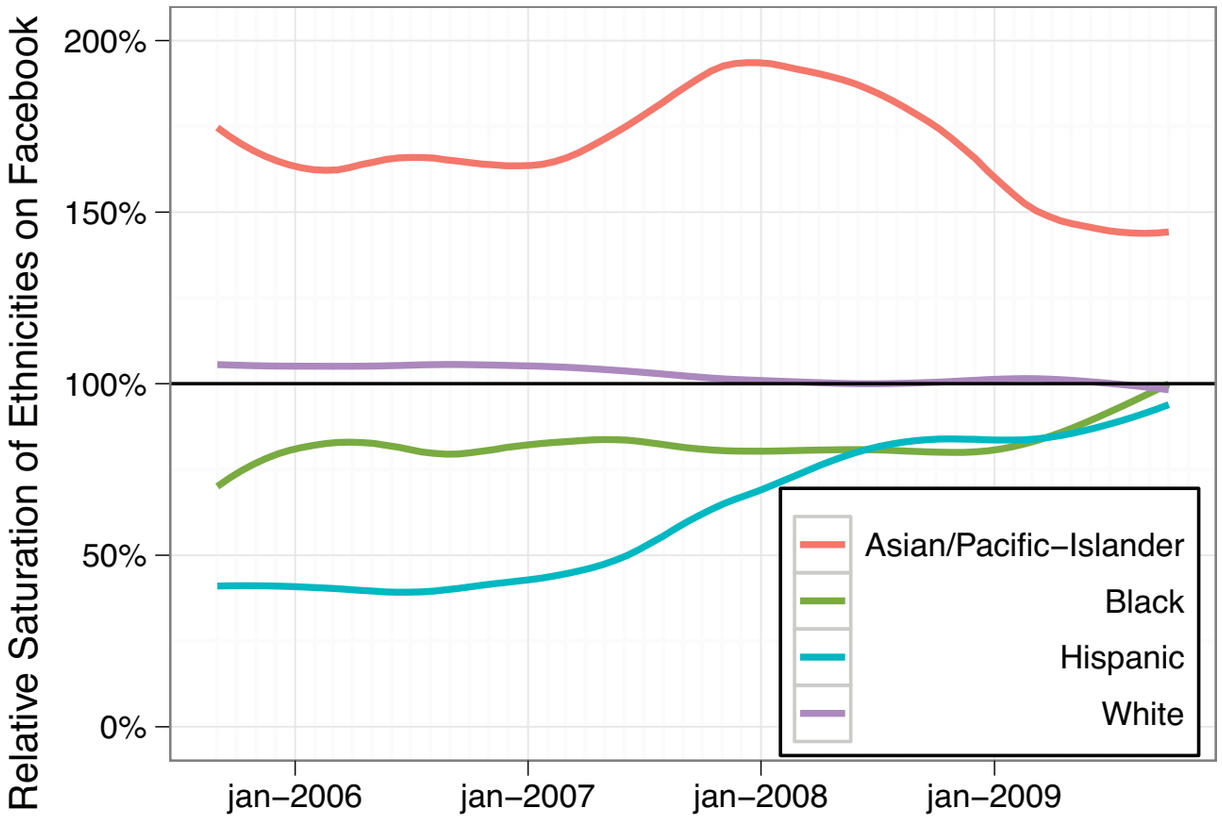
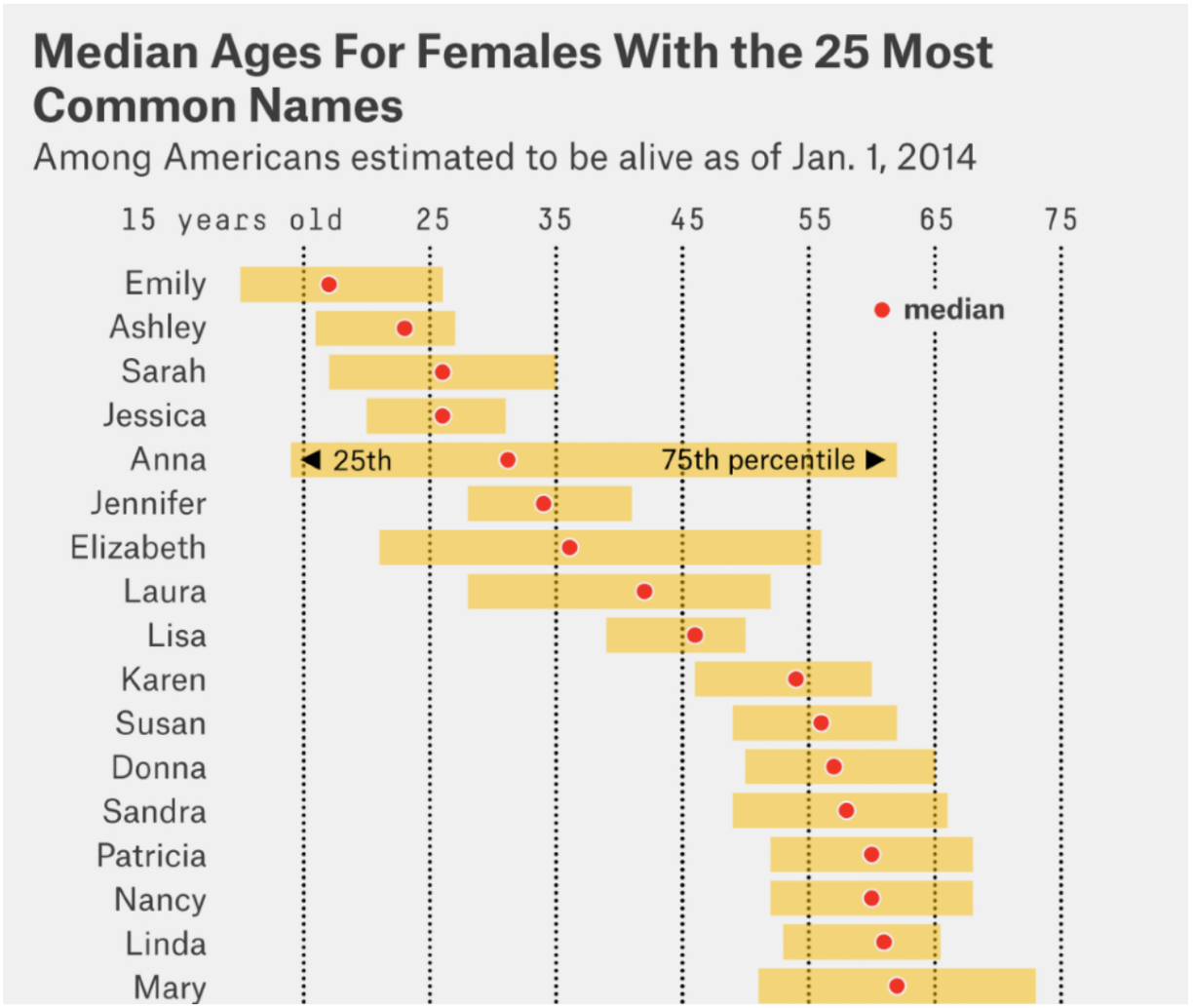



Figure 6: Relative saturation of ethnicities on Facebook. As the lines converge towards 100% (center), the makeup of U.S. Facebook converges towards that of the addressable Internet population.

# FiveThirtyEight example

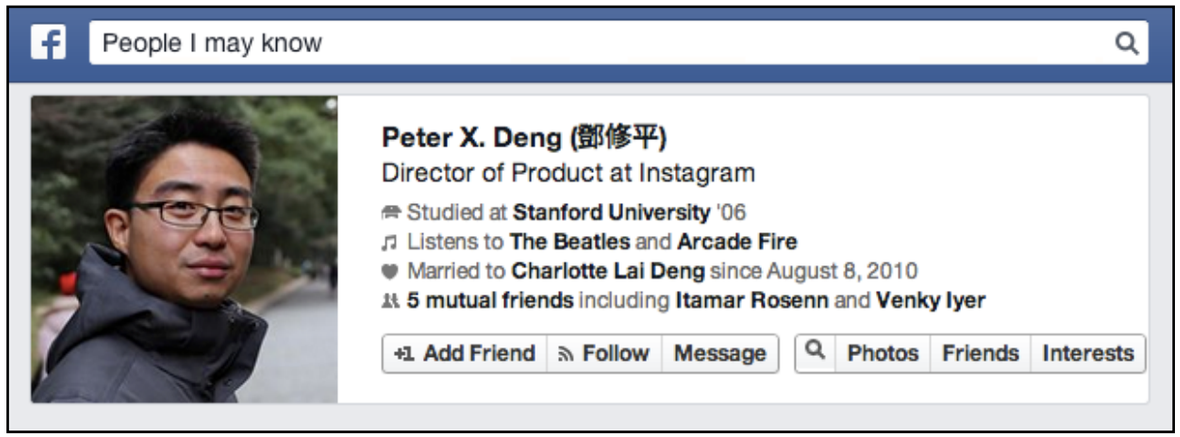
- After discussion: https://tinyurl.com/icme538




- User/customer records often record names, but not ages
- Uses US Social Security Administration data to predict ages from names.
- Inaccurate for individuals, accurate in aggregate.

# After break: discussion!

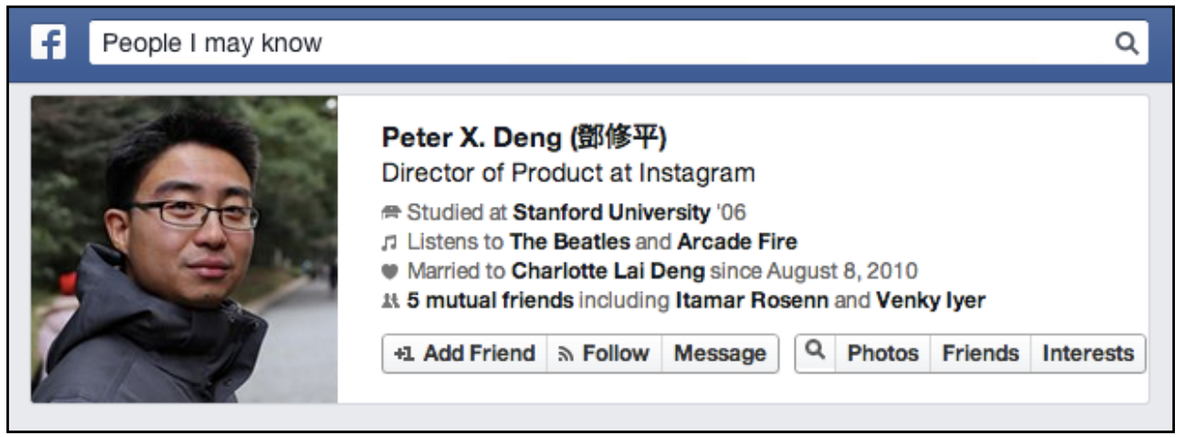
- Friend recommendations on social platforms
  - “People You May Know” (LinkedIn, Facebook), “Who to Follow” (Twitter), etc.




- Imagine that we’re an engineering team being tasked with overhauling Facebook’s PYMK recommendations.
  - Goals?
  - Good data sources?
  - What could we do, what should we do?

# After break: discussion!

- Friend recommendations on social platforms
  - “People You May Know” (LinkedIn, Facebook), “Who to Follow” (Twitter), etc.



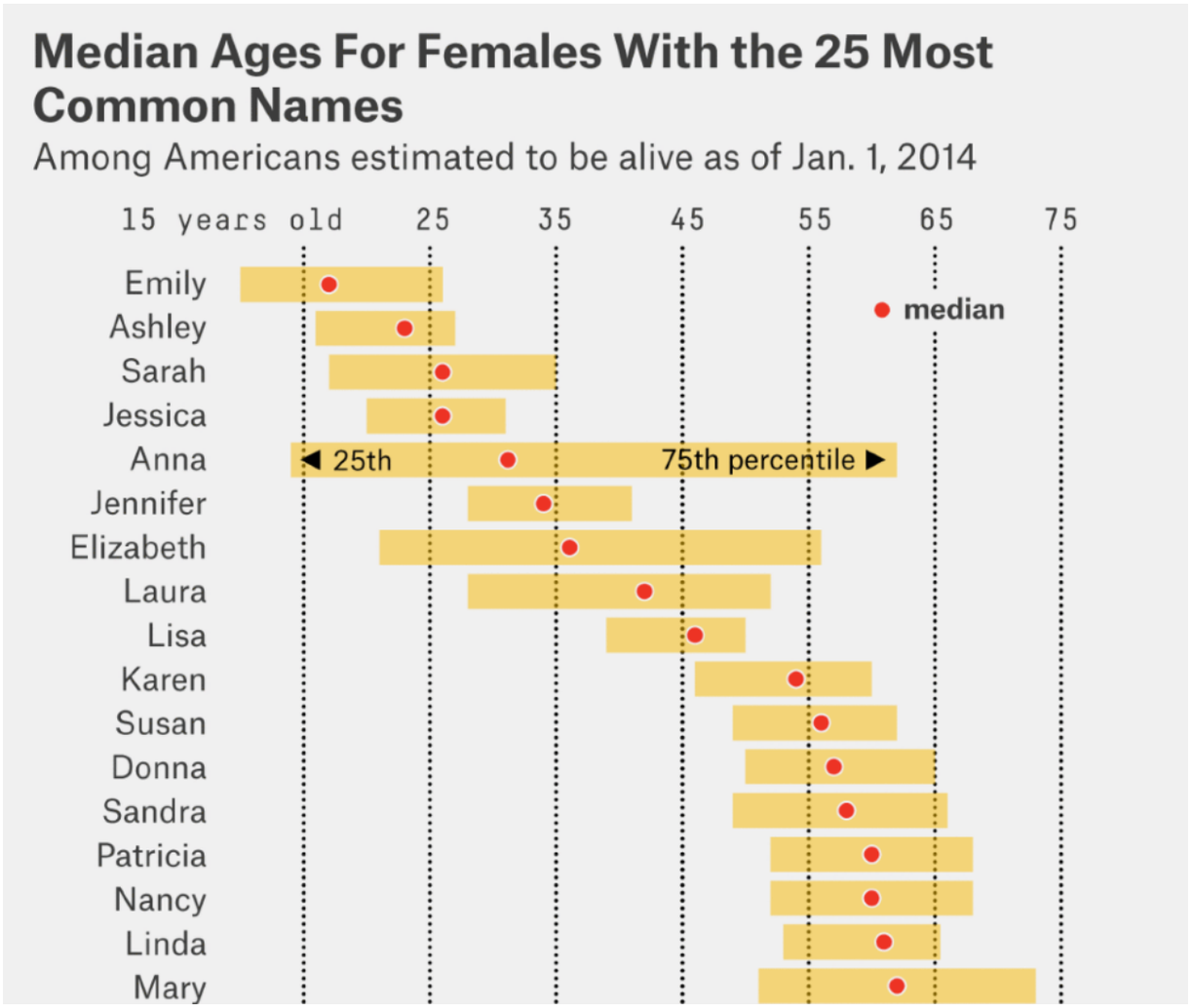

- Imagine that we’re an engineering team being tasked with overhauling Facebook’s PYMK recommendations.
  - Goals?
  - Good data sources?
  - What could we do, what should we do?

# Break!

# Discussion

# FiveThirtyEight example

- After discussion: https://tinyurl.com/icme538




- User/customer records often record names, but not ages
- Uses US Social Security Administration data to predict ages from names.
- Inaccurate for individuals, accurate in aggregate.

# Part 3 - Machine learning with relational data

# Identity in high dimensions

- Recall…
- World population: ~7 billion.
- Log2(7 billion) = 33 bits

- Relational data is very high-dimensional.

# Relational data

- Affiliation networks:
    - Movies watched (Netflix)
    - Facebook likes
    - Contact list phone numbers
    - Credit card retailers
    - Restaurant check-ins
    - URLs clicked
- Or straight-up social networks:
    - Facebook friends
    - Twitter followers and followees

- 1-2p: focus on identification. Now: focus on prediction.

# Relational data & prediction

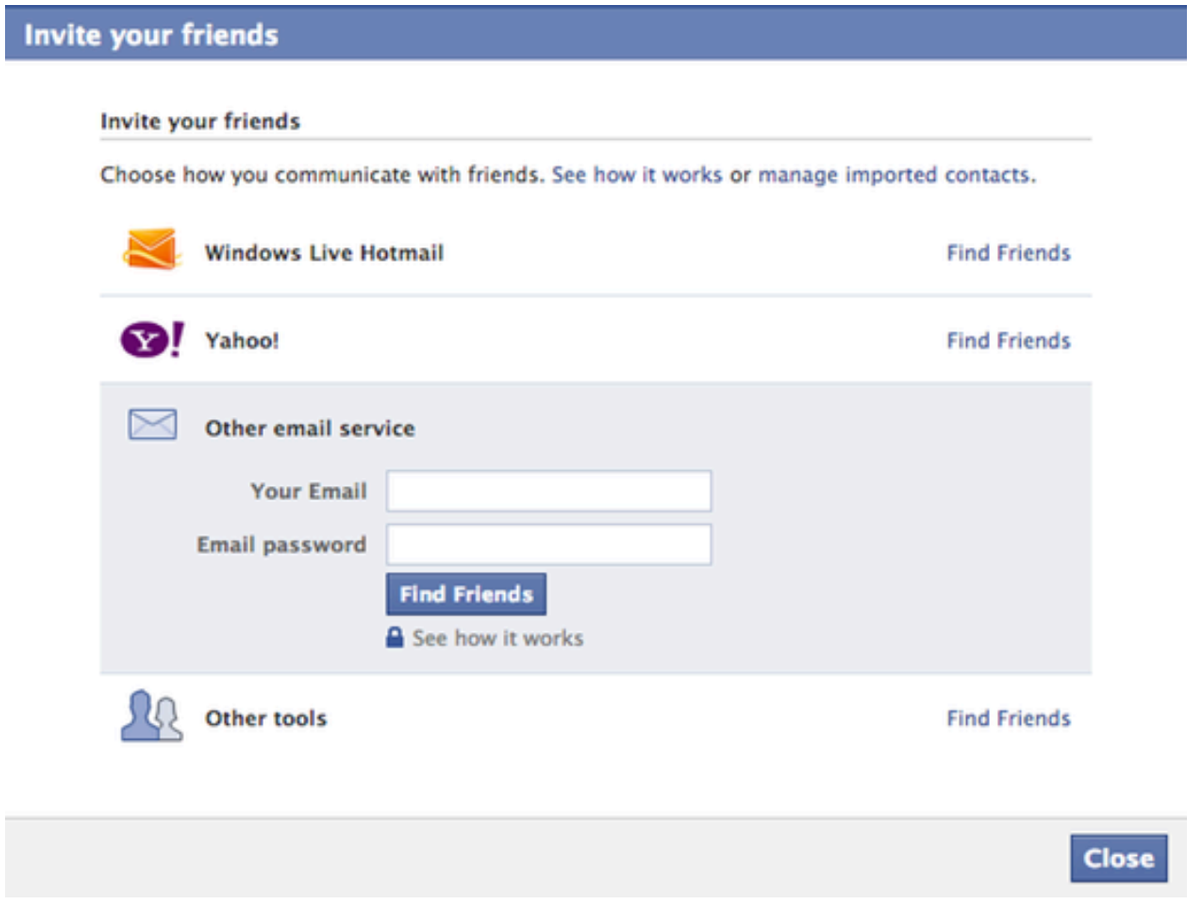
- Predict a trait based on affiliations:
  - Demographics (Zheleva & Getoor, 2009)
  - Psychological traits (Kosinski et al., 2013)
  - Ad click through rate (CTR)

- Focus today on just predicting the trait, not on whether that trait is useful toward next steps.
  - For sample of drama there, see PNAS exchange between Matz et al. (2017) & Eckles et al. (2018)

# So much relational data

- Take “friends” as relations

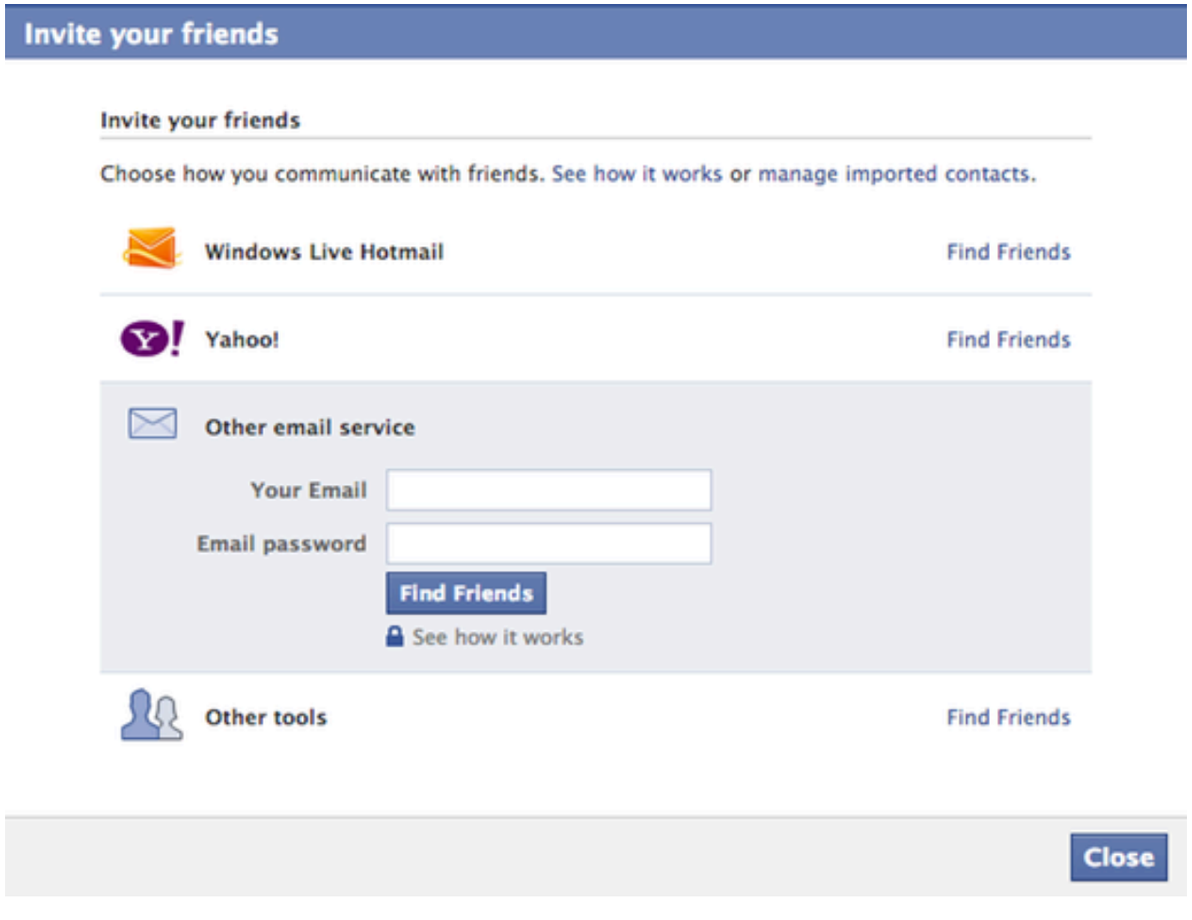
- …the people you’re friends with, the people you friended, the people in your phonebook, the people who’s phonebook you’re in, the people in your gmail, the people who’s gmail you’re in…



# So much relational data

- Take “friends” as relations

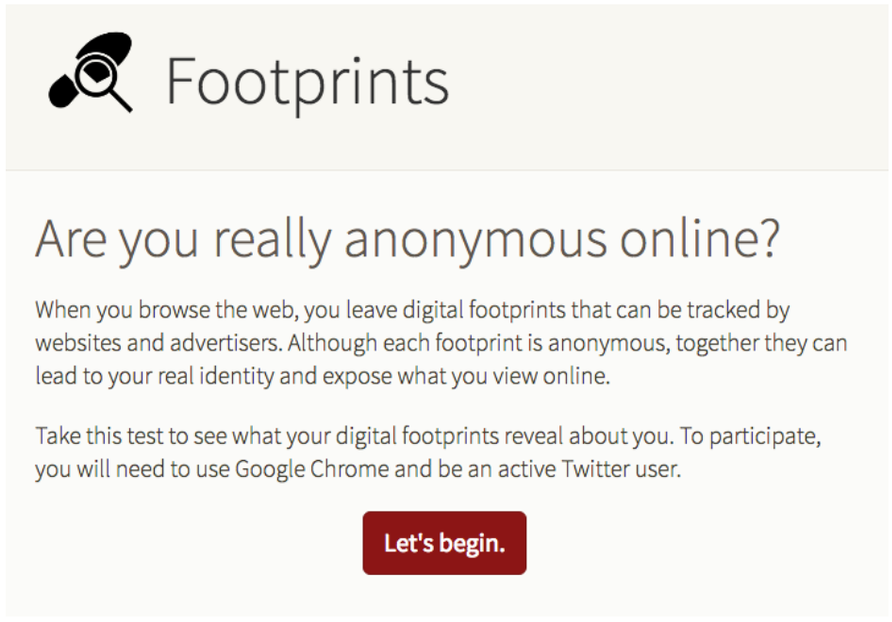
- …the people you’re friends with, the people you friended, the people in your phonebook, **the people who’s phonebook you’re in**, the people in your gmail, **the people who’s gmail you’re in**…

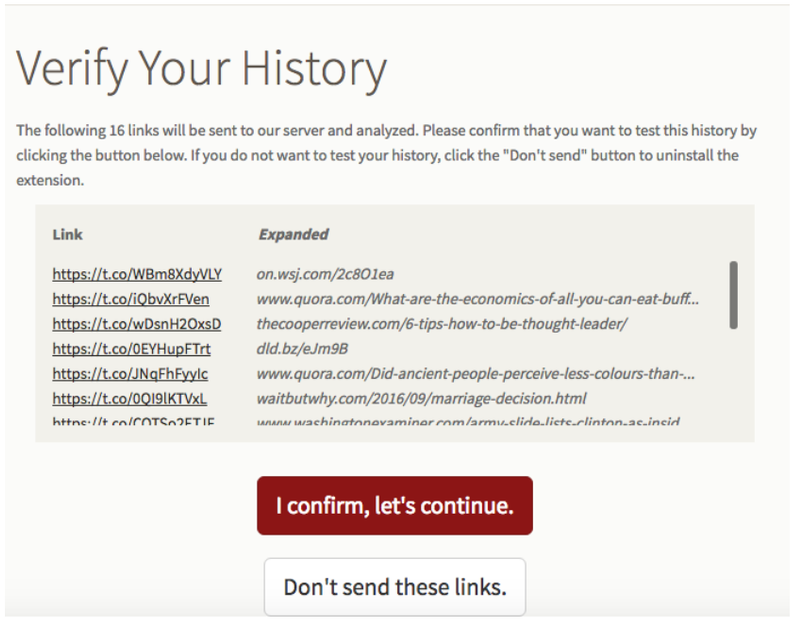


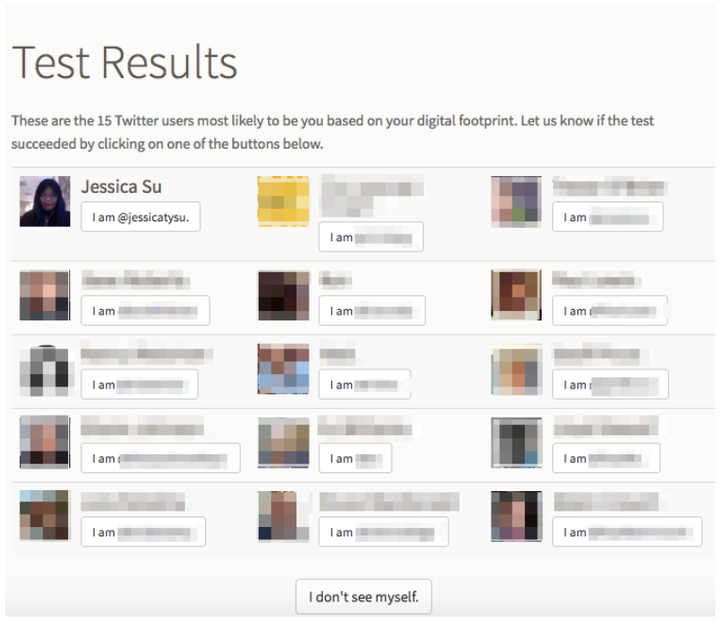
# Social networks as a backboard

- If you know somebody's twitter ID, you can learn from their twitter affiliations. But you can learn twitter ID from…

- Su et al. 2017: looking at URLs in your browser history (Chrome extension), predict your twitter ID.

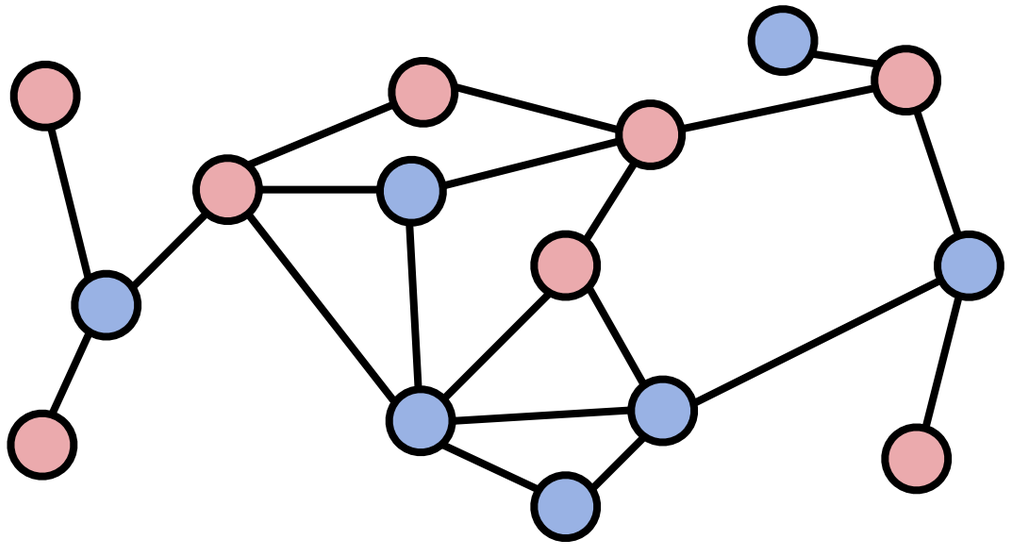


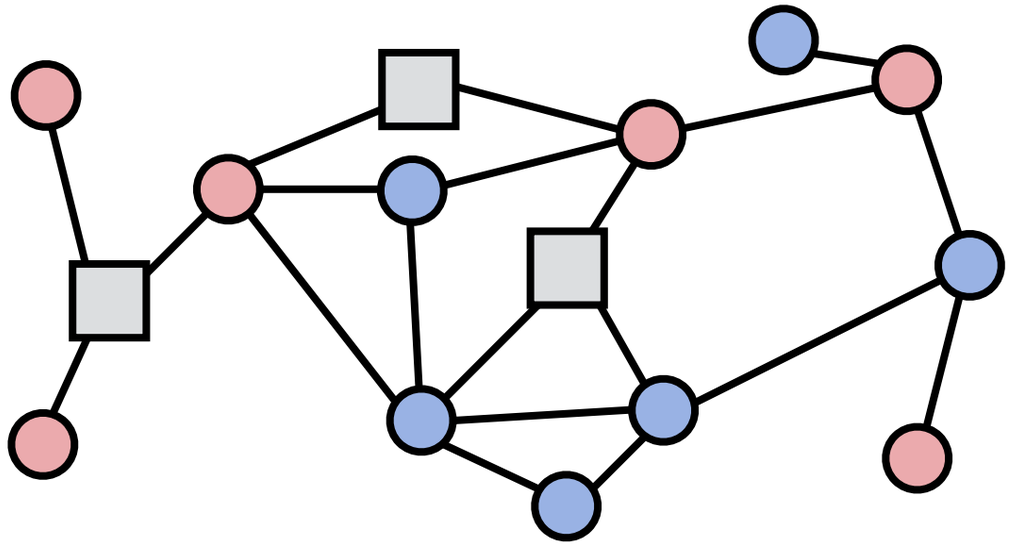



# Some basic techniques

- Network of red/blue nodes:



- Use diffusion, based on homophily, to infer values:

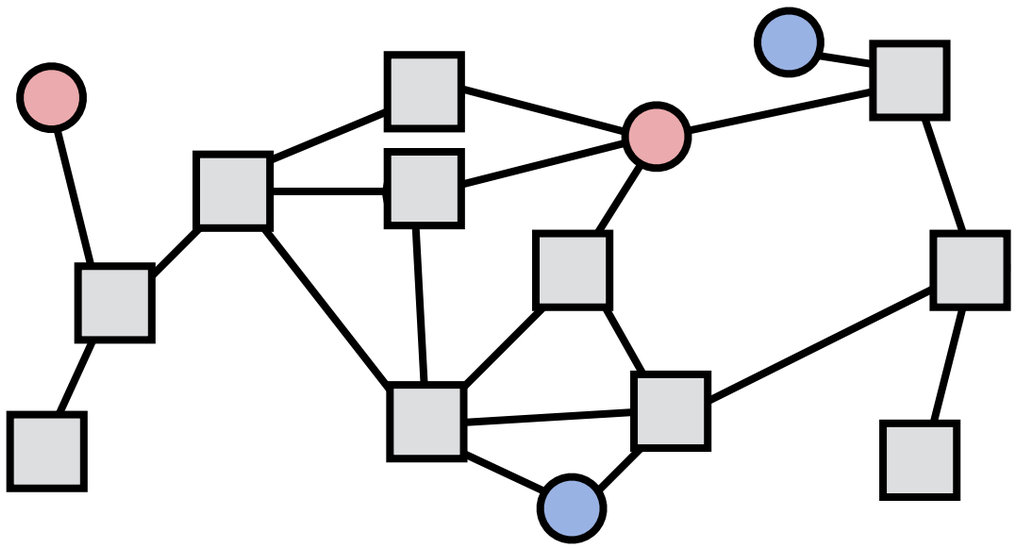


# Some basic techniques

- **Proportion friends same?** “Birds of a feather flock together” (Lazarsfeld & Merton 1954; Ibarra 1992; McPherson et al, 2001)
- **Majority Vote/Label Propagation:** Each labelled neighbor given one vote; majority rules. (Macskassy-Provost 2003)
- **Semi-supervised learning:** (Zhu-Ghahramani-Lafferty 2003, Zhou et al. 2003, Xu-Dyer-Owen 2010)
- **Personalized PageRank:** (Jeh-Widom 2003)
- **Graph neural networks:** active research area…

# Diffusion on graphs

- **Diffusion intuition:**

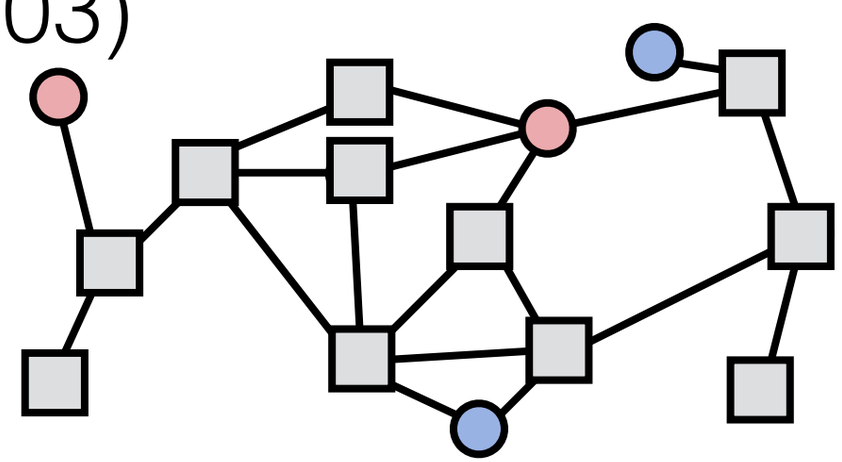


# ZGL and semi-supervised learning

- Consider a graph G=(V, E) with labelled nodes.
- Looking for a “function” f(v), defined on the node set, where f(v)=1 for red nodes, f(v)=0 for blue nodes, and minimizes:

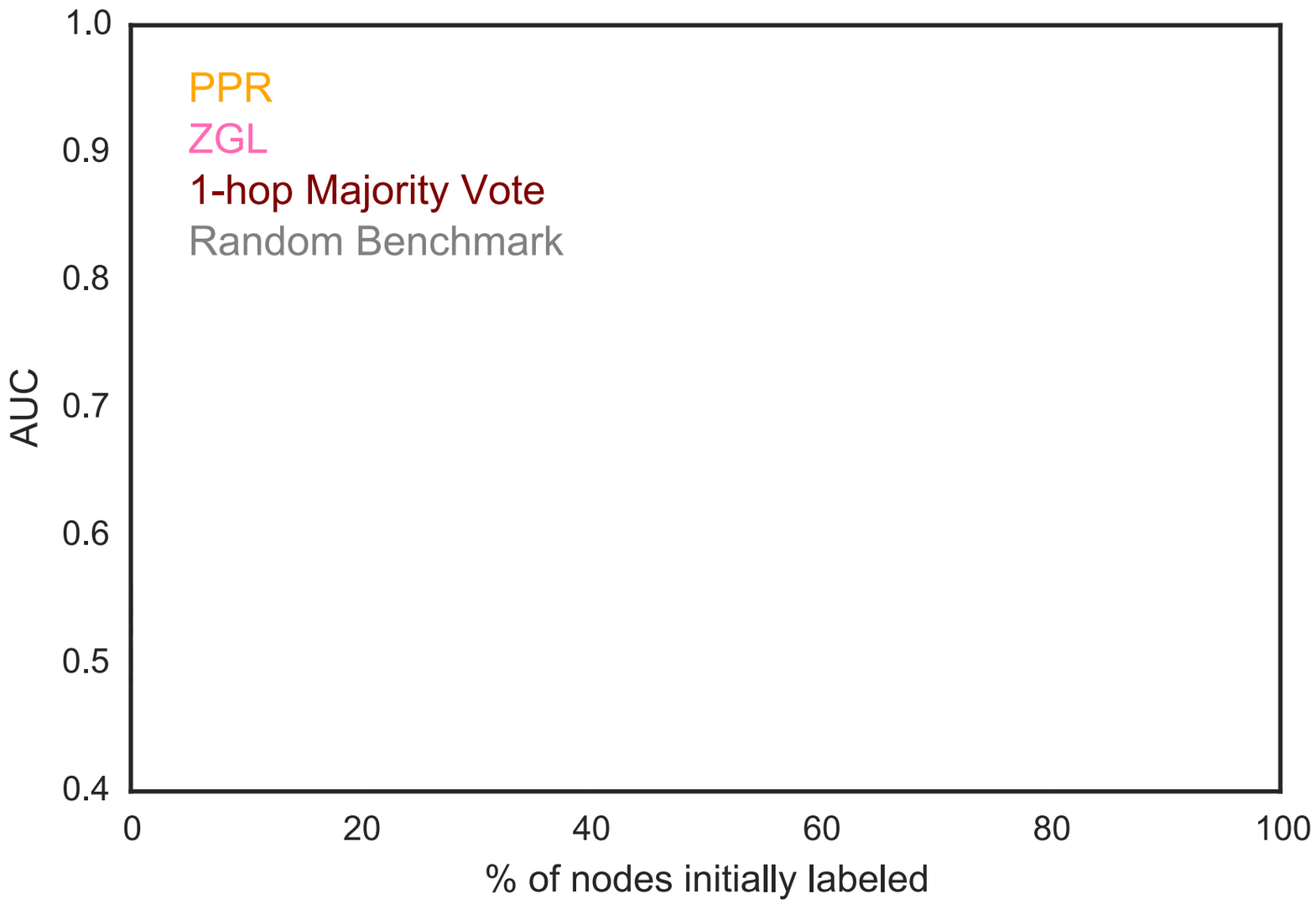
$$\sum_{(u,v)\in E} [f(u) - f(v)]^2$$

- “Turns out”: solution from solving a linear system.
- See (Zhu, Ghahramani, & Lafferty 2003) related to: spectral graph theory, Kriging, transductive learning, harmonic functions, …



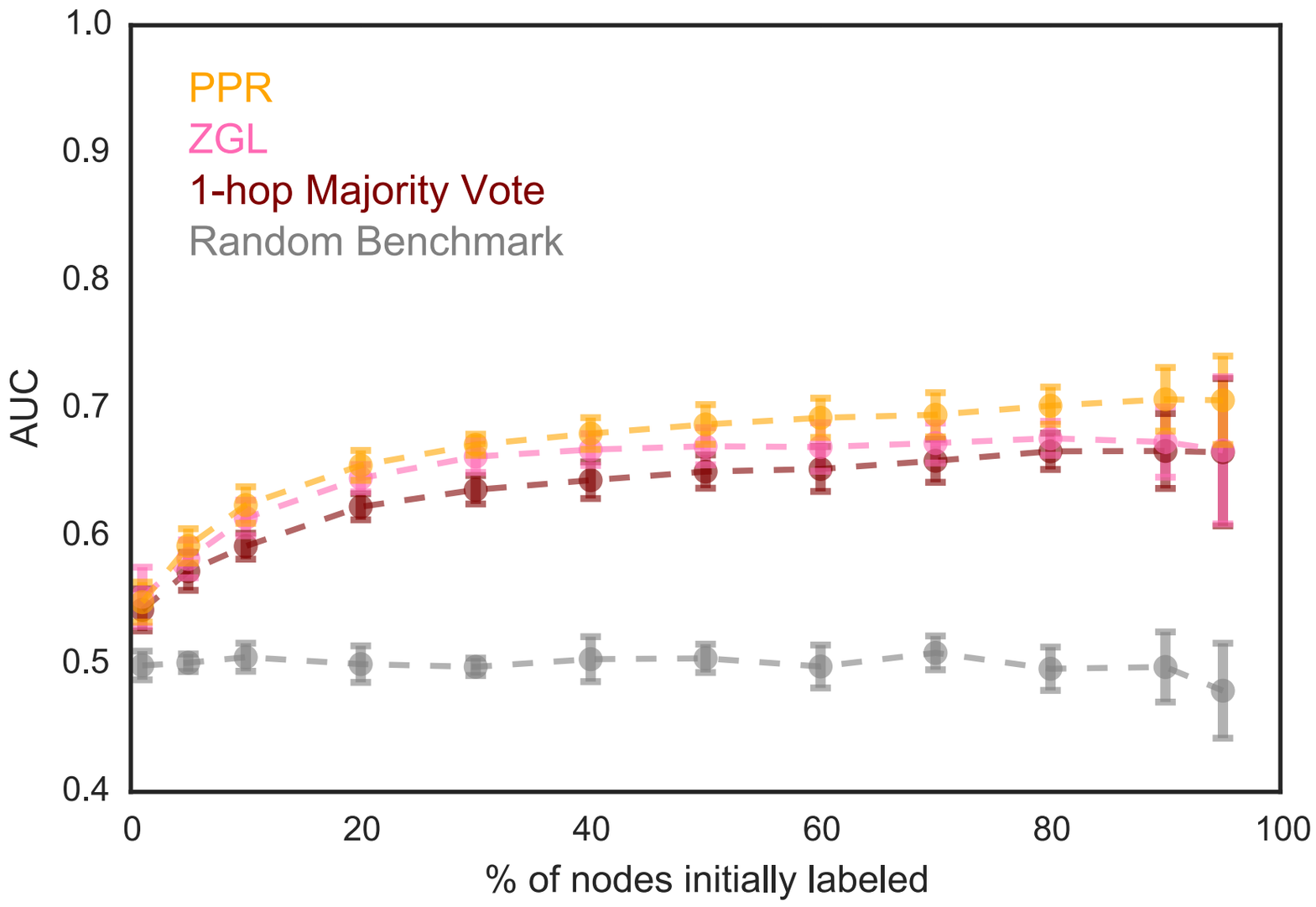
# Some less good techniques

- **Predict gender, Amherst FB, 2005:**



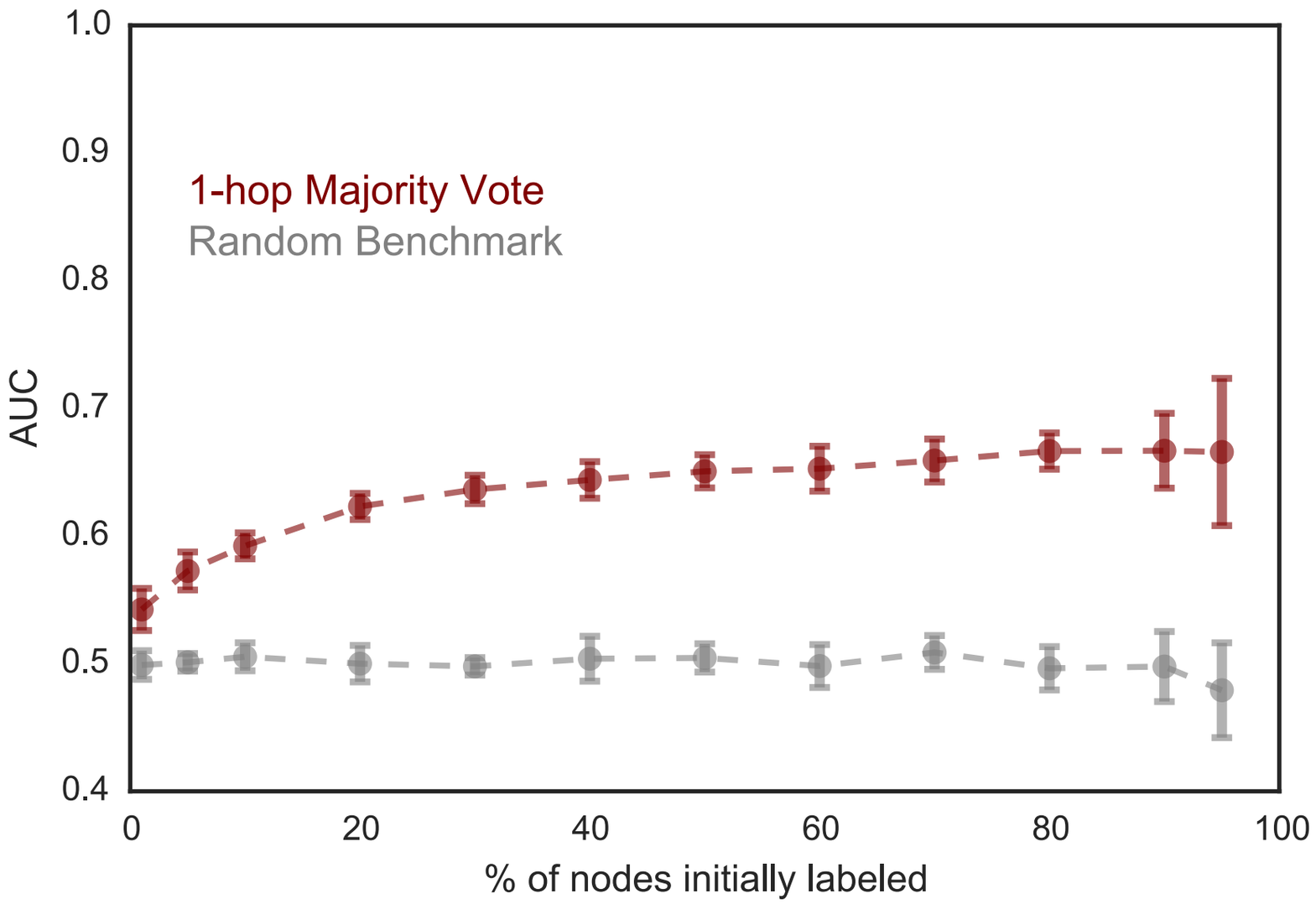
# Some less good techniques

- **Predict gender, Amherst FB, 2005:**



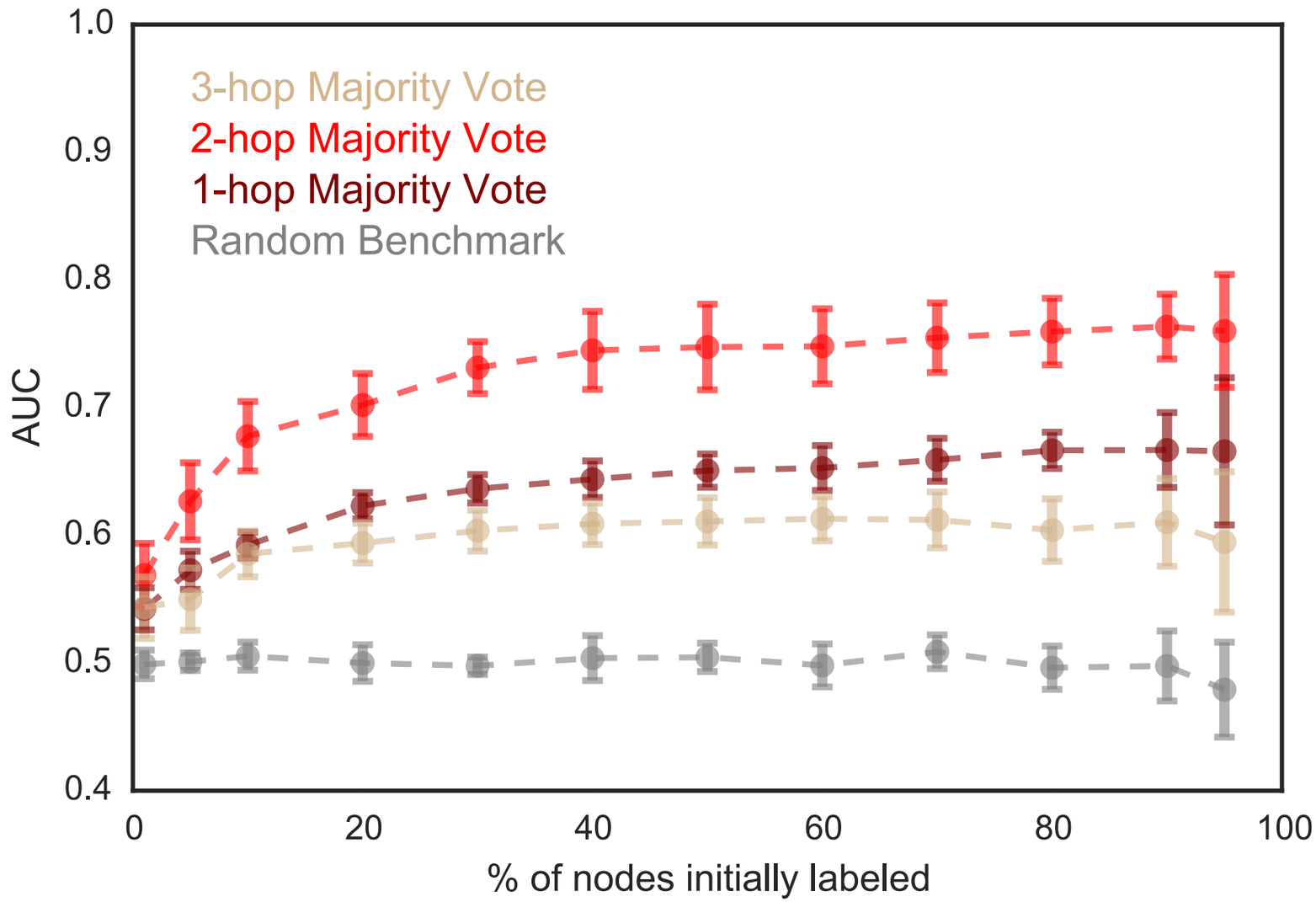
# Some less good techniques

- **Predict gender, Amherst FB, 2005:** 1-hop MV



# Some less good techniques

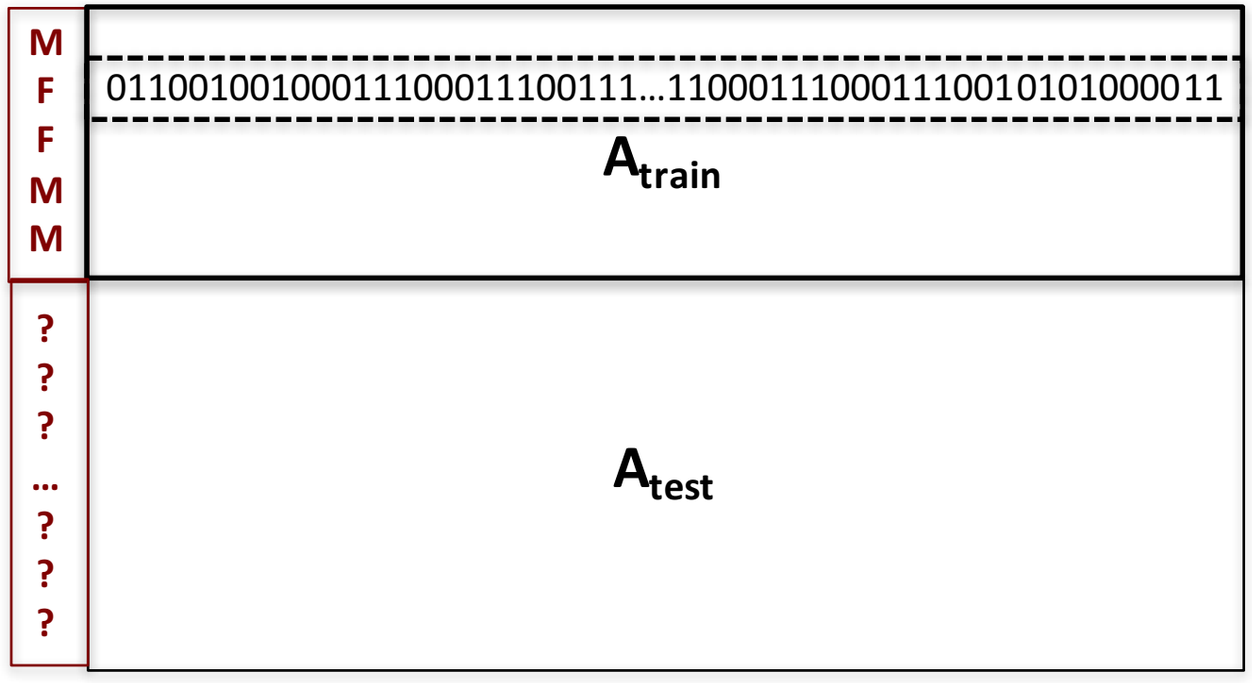
- **Predict gender, Amherst FB, 2005:** k-hop?




- There's something special about 2 hops….

# Classification from likes/friends

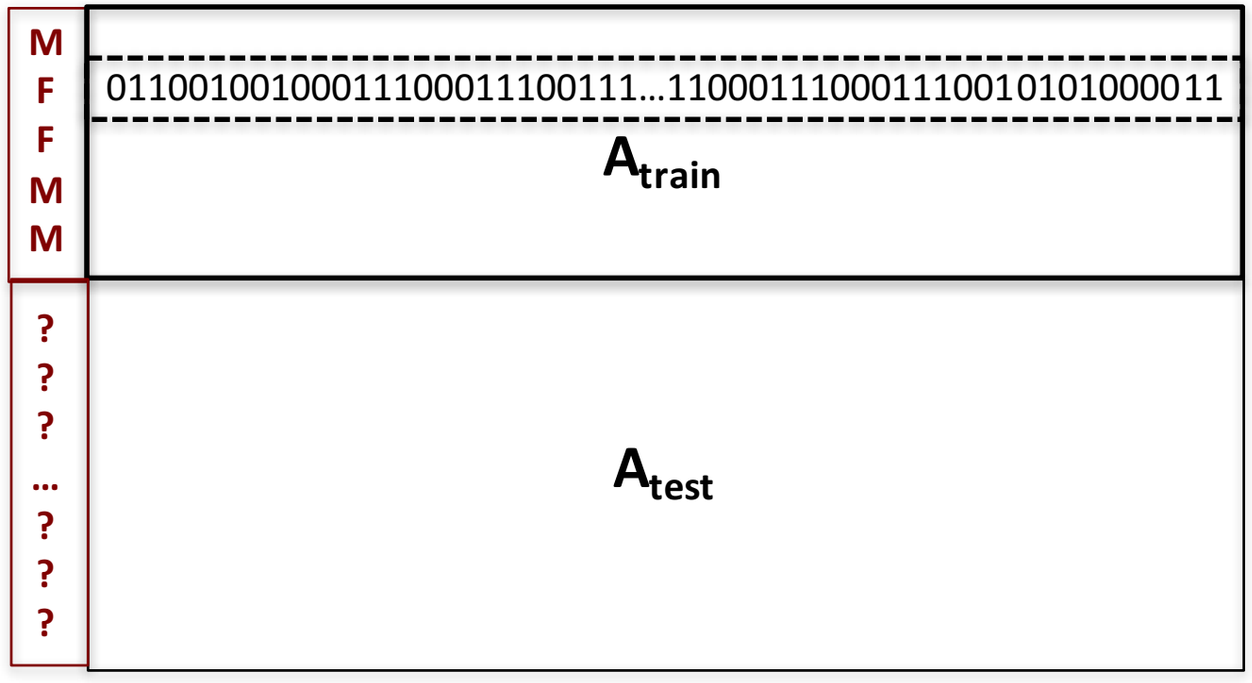
- LINK algorithm (Zheleva and Getoor, 2009) proposed using adjacency matrix as a large sparse feature vector.
- For example, predicting gender:




- Logistic regression, Random Forests, Naive Bayes, …

# Classification from likes/friends

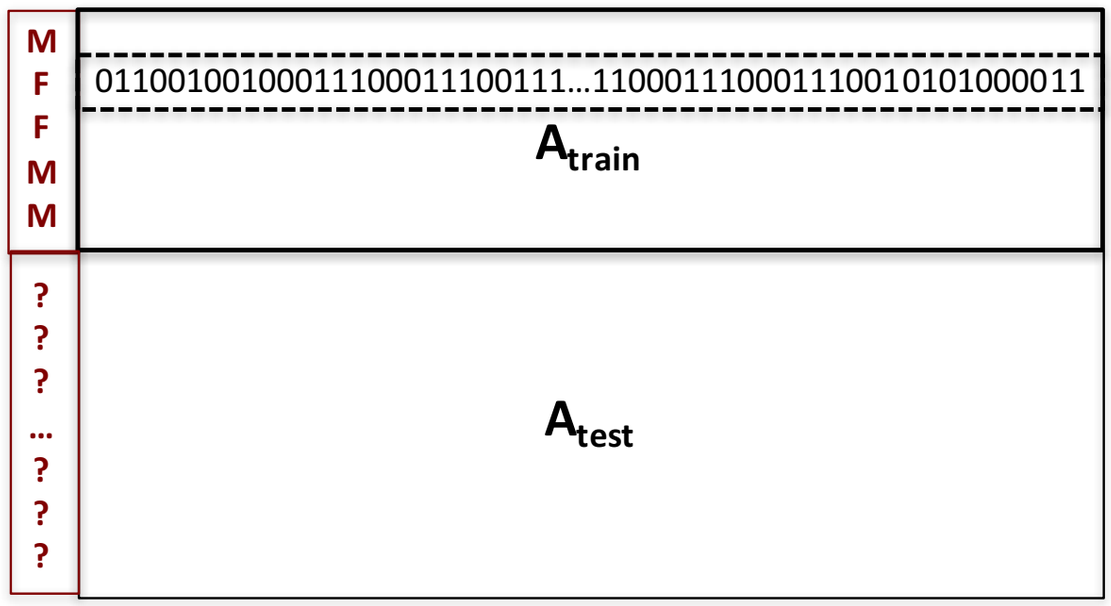
- LINK algorithm (Zheleva and Getoor, 2009) proposed using adjacency matrix as a large sparse feature vector.
- For example, predicting gender:




- Logistic regression, Random Forests, Naive Bayes, …
- **Log-Reg** plug-and-play; **Naive Bayes** now for intuition.

# Logistic regression

- LINK Logistic regression: $\sum_{j=1}^{n} \beta_j A_{ij} > 0$.
- $\beta_j$ is interpretable as the contribution that "being friends with j" makes to the log-odds of being F.
- Kosinski 2013 uses logistic regression on features that are PCA'ed down from the user-like matrix. Same idea as LINK.



# Naïve Bayes in a nutshell

- Features $X_1,\ldots,X_N$, class Y.
- Pick Y to maximize $\Pr(Y|X_1,\ldots,X_N)$.

$$\hat{y} = h(X) = \arg\max_{y \in Y} Pr(Y|X_1, ..., X_N)$$

- Bayes theorem + assume $X_1,\ldots,X_N$ are conditionally independent:

$$h(X) = \arg\max_{y \in Y} \frac{Pr(X|Y)Pr(Y)}{Pr(X)} = \arg\max_{y \in Y} Pr(Y) \prod_{i=1}^{N} Pr(X_i|Y)$$

# NB likelihood ratio

- For binary problem (two classes):

$$LR(X) = \frac{\Pr(Y=1)\prod_{i=1}^{n}\Pr(X_i|Y=1)}{\Pr(Y=0)\prod_{i=1}^{n}\Pr(X_i|Y=0)} = \frac{\Pr(Y=1)}{\Pr(Y=0)}\prod_{i=1}^{n}\frac{\Pr(X_i|Y=1)}{\Pr(X_i|Y=0)}$$

# NB likelihood ratio

- For binary problem (two classes):

$$LR(X) = \frac{\Pr(Y=1)\prod_{i=1}^{n}\Pr(X_i|Y=1)}{\Pr(Y=0)\prod_{i=1}^{n}\Pr(X_i|Y=0)} = \frac{\Pr(Y=1)}{\Pr(Y=0)}\prod_{i=1}^{n}\frac{\Pr(X_i|Y=1)}{\Pr(X_i|Y=0)}$$

- Then classification problem is:

$$\log LR(X) \geq 0 \Leftrightarrow \hat{y} = 1,$$
$$\log LR(X) < 0 \Leftrightarrow \hat{y} = 0.$$

# NB with likes/friends

- Binary features too; separate into $X_i=1$ and $X_i=0$:

$$LR(X) = \frac{\Pr(Y=1)}{\Pr(Y=0)} \prod_{i:X_i=1} \frac{\Pr(X_i=1|Y=1)}{\Pr(X_i=1|Y=0)} \prod_{i:X_i=0} \frac{\Pr(X_i=0|Y=1)}{\Pr(X_i=0|Y=0)}$$

# NB with likes/friends

- Binary features too; separate into $X_i=1$ and $X_i=0$:

$$LR(X) = \frac{\Pr(Y=1)}{\Pr(Y=0)} \prod_{i:X_i=1} \frac{\Pr(X_i=1|Y=1)}{\Pr(X_i=1|Y=0)} \prod_{i:X_i=0} \frac{\Pr(X_i=0|Y=1)}{\Pr(X_i=0|Y=0)}$$

- In sparse data, $X_i=0$ true for the vast majority of data points; as a result we can focus on $X_i=1$. Take logs:

$$\log LR(X) = C + \sum_{i:X_i=1} \log \frac{\Pr(X_i=1|Y=1)}{\Pr(X_i=1|Y=0)}$$

# NB with likes/friends

- In NB, “feature probabilities” are “parameters”:

$$\log LR(X) = C + \sum_{i:X_i=1} \log \frac{\Pr(X_i = 1|Y = 1)}{\Pr(X_i = 1|Y = 0)}$$

- How do we estimate parameters?

# NB with likes/friends

- In NB, “feature probabilities” are “parameters”:

$$\log LR(X) = C + \sum_{i:X_i=1} \log \frac{\Pr(X_i = 1|Y = 1)}{\Pr(X_i = 1|Y = 0)}$$

- How do we estimate parameters?
  - Deriving NB parameter MLEs is an exercise in constrained optimization (Lagrangian multipliers).

# NB with likes/friends

- In NB, “feature probabilities” are “parameters”:

$$\log LR(X) = C + \sum_{i:X_i=1} \log \frac{\Pr(X_i = 1|Y = 1)}{\Pr(X_i = 1|Y = 0)}$$

- How do we estimate parameters?
  - Deriving NB parameter MLEs is an exercise in constrained optimization (Lagrangian multipliers):

$$\widehat{\Pr}(X_i = 1|Y = 1) = \frac{d_{i,1} + 1}{n_1 + 2}$$
$$\widehat{\Pr}(X_i = 1|Y = 0) = \frac{d_{i,0} + 1}{n_0 + 2}$$

  where $d_{i,0}$ are the number of Y=0 units with feature i.

# NB with likes/friends

- In NB, “feature probabilities” are “parameters”:

$$\log LR(X) = C + \sum_{i:X_i=1} \log \frac{\Pr(X_i = 1|Y = 1)}{\Pr(X_i = 1|Y = 0)}$$

- How do we estimate parameters?
  - Deriving NB parameter MLEs is an exercise in constrained optimization (Lagrangian multipliers):

$$\widehat{\Pr}(X_i = 1|Y = 1) = \frac{d_{i,1} + 1}{n_1 + 2}$$

$$\widehat{\Pr}(X_i = 1|Y = 0) = \frac{d_{i,0} + 1}{n_0 + 2}$$

Laplace smoothing

where $d_{i,0}$ are the number of Y=0 units with feature i.

# NB with likes/friends

- Put it together:

$$\log LR(X) = C' + \sum_{i:X_i=1} \log \frac{d_{i,1}+1}{d_{i,0}+1}$$

# NB with likes/friends

- Put it together:

$$\log LR(X) = C' + \sum_{i:X_i=1} \log \frac{d_{i,1}+1}{d_{i,0}+1}$$

- In graph language, where *N(j)* is the neighborhood of j:

$$\sum_{i \in N(j)} \log \frac{d_{i,1}+1}{d_{i,0}+1} > C'$$

- So predicted class of j depends on how many 1s vs. 0s each of j's neighbors ("each i") have: **"friends of friends"**

# NB with likes/friends

- Put it together:

$$\log LR(X) = C' + \sum_{i:X_i=1} \log \frac{d_{i,1}+1}{d_{i,0}+1}$$

- In graph language, where *N(j)* is the neighborhood of j:

$$\sum_{i \in N(j)} \log \frac{d_{i,1}+1}{d_{i,0}+1} > C'$$

- So predicted class of j depends on how many 1s vs. 0s each of j's neighbors ("each i") have: **"friends of friends"**
- Here C' depends on the class balance; important for classification, less so for ranking.

# NB with likes/friends

- One more time:

$$\sum_{i \in N(j)} \log \frac{d_{i,1}+1}{d_{i,0}+1} > C'$$

- **Intuition:** look at each of j's friends. They're likely to be male if their friends mostly friend male nodes.

# NB with likes/friends

- One more time:

$$\sum_{i \in N(j)} \log \frac{d_{i,1}+1}{d_{i,0}+1} > C'$$

- **Intuition:** look at each of j's friends. They're likely to be male if their friends mostly friend male nodes.

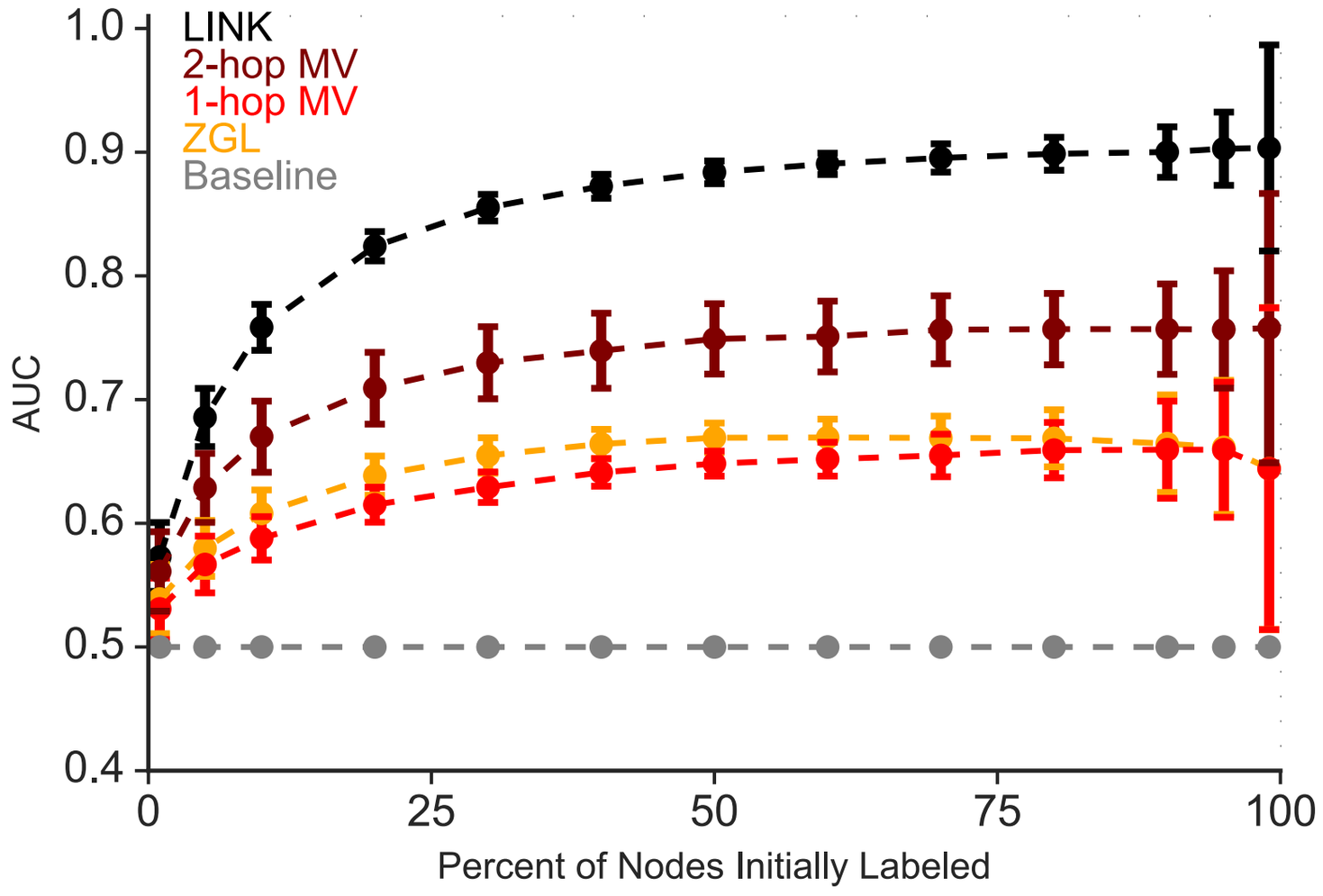
- Call

$$\beta_i = \log \frac{d_{i,1}+1}{d_{i,0}+1}$$

the LINK weight of node i. Notice: $\sum_{i \in N(j)} \beta_i = \beta^T X$

# How well does LINK do?

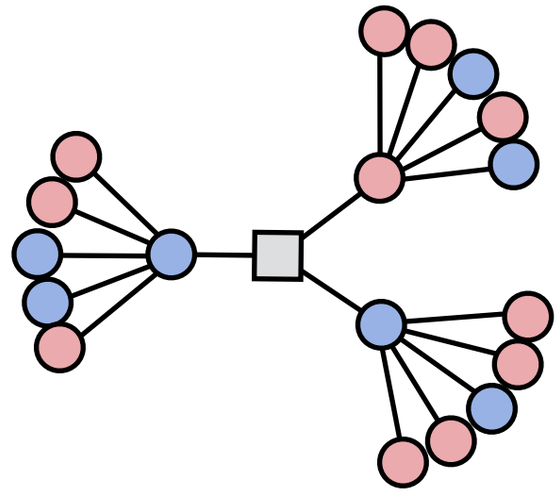
- Facebook Amherst, gender:

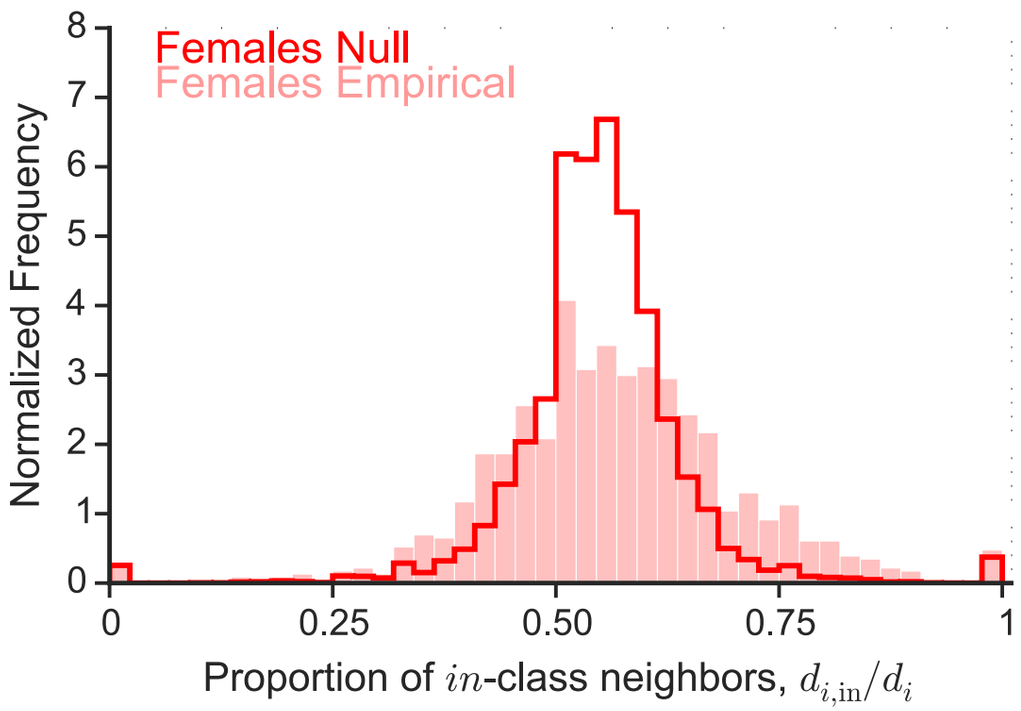


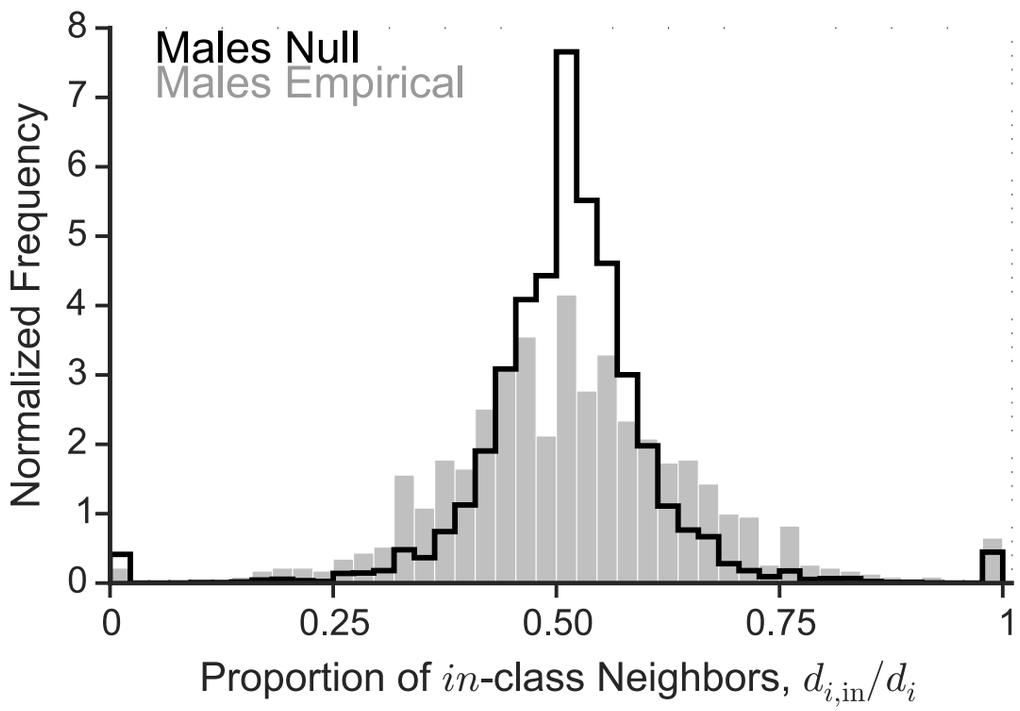

(From Altenburger & Ugander, 2018)

# Why? Not homophily!

- Extreme gender-frienders, with high-magnitude LINK weights, drive predictions.





# Technical summary

- Homophily-based inference (semi-supervised learning and related methods) work well when homophily is present.

- Homophily is a sufficient but not necessary condition for relational inference.

- LINK doesn't require homophily to work, instead can work under a separate sufficient condition: an overdispersion of preferences or "monophily".

# Implications

- Even if you don't have biased friending habits, your friends might, empowering inferences about you.

- Used in missing data models and ad targeting, but also fraud and abuse prevention models.

- **Ongoing debates** in the graph neural network literature about how different architectures can/can't harness varied types of relational information to make predictions.

**Beyond Homophily in Graph Neural Networks: Current Limitations and Effective Designs**


**Jiong Zhu**
University of Michigan
jiongzhu@umich.edu

**Yujun Yan**
University of Michigan
yujunyan@umich.edu

**Lingxiao Zhao**
Carnegie Mellon University
lingxia1@andrew.cmu.edu

**Mark Heimann**
University of Michigan
mheimann@umich.edu

**Leman Akoglu**
Carnegie Mellon University
lakoglu@andrew.cmu.edu

**Danai Koutra**
University of Michigan
dkoutra@umich.edu


**Is Homophily a Necessity for Graph Neural Networks?**


**Yao Ma**
New Jersey Institute of Technology
majunyao@gmail.com

**Xiaorui Liu**
Michigan State University
xiaorui@msu.edu

**Neil Shah**
Snap Inc.
nshah@snap.com

**Jiliang Tang**
Michigan State University
tangjili@msu.edu

# Summary, Day 1

- Data privacy is hard. Most common threats are variations on linkage/identification attacks.

- **Linkage** as a double edge sword:
    - Netflix privacy attack example
    - Monitoring aggregate unseen demographics

- **Attribute prediction** using relational machine learning: your high-dimensional data, and what it says about you

- **Day 2:** Differential privacy, data transparency, regulation.