# MEASURING PRIVACY RISK IN ONLINE SOCIAL NETWORKS

Justin Becker, Hao Chen

UC Davis

May 2009

# Motivating example

## College admission

- Kaplan surveyed 320 admissions offices in 2008
- 1 in 10 admissions officers viewed applicants’ online profiles
- 38% said they had “negative impact” on applicants

If only we could measure privacy risk

# Scale of Facebook

- 200 million active users
- 100 million users log on once a day
- 1 billion pieces of content shared each week
- More than 20 million users update their status daily

http://www.facebook.com/press/info.php?statistics

# Will users take action?

Online survey using a simple tool

- Calculated privacy risk
  - Information revealed to third party applications
- Reported score to participant

- Results
  - 105 participants
  - 65% said they would change privacy settings

# Demographics

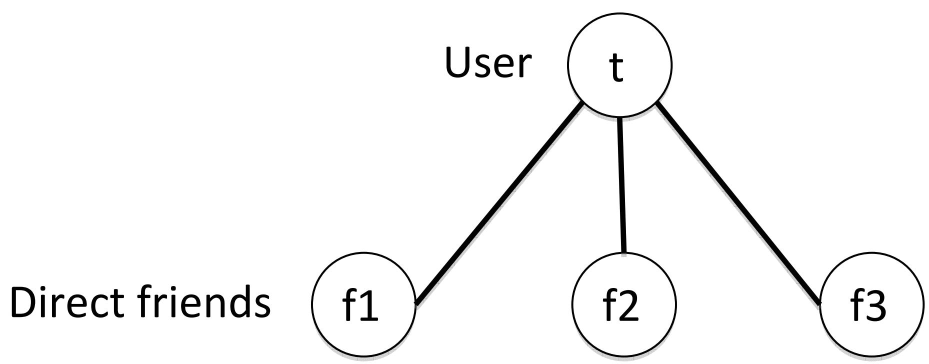
- 47 men and 24 women
- The average age was 23.89 with
  - standard deviation of 6.1 and a range of 14-44.
- 12 different countries
  - Canada, China, Ecuador, Egypt, Iran, Malaysia, New Zealand,Pakistan, Singapore, South Africa, United Kingdom, United States

# PrivAware

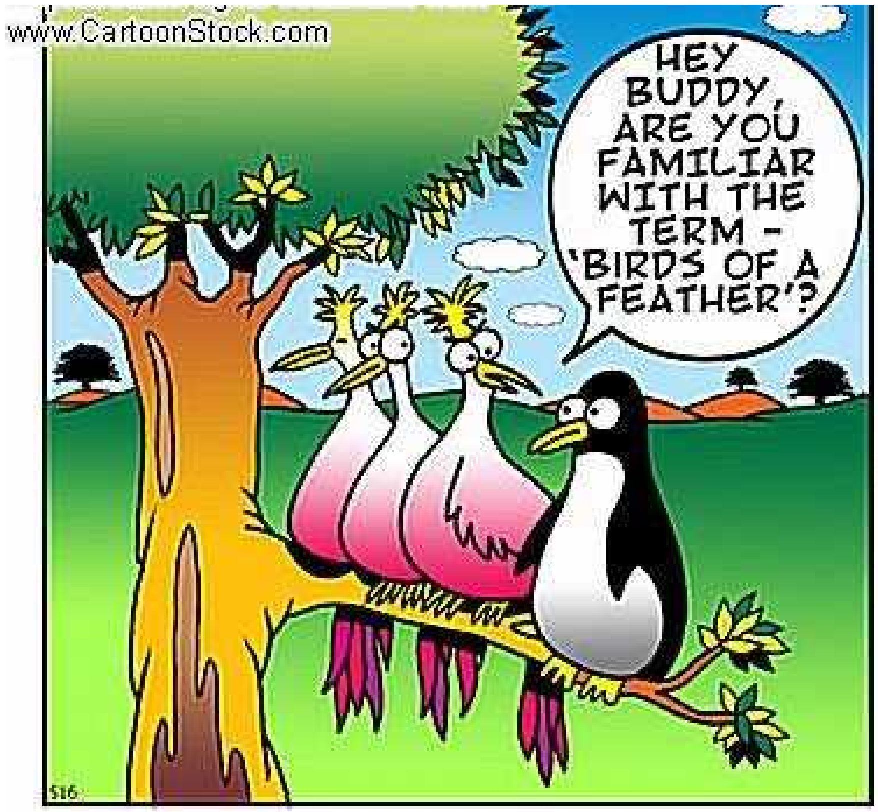
- A tool to
  - **measure** privacy risks
  - suggest user actions to **alleviate** privacy risks
- Developed using Facebook API
  - Can query user and direct friends profile information
  - Measures privacy risk attributed to social contacts

# Threat model

- Let **user** *t be the* inference *target.*
- *Let F be the* set of **direct friends**.
- **Infer** the **attributes** of *t* from *F.*

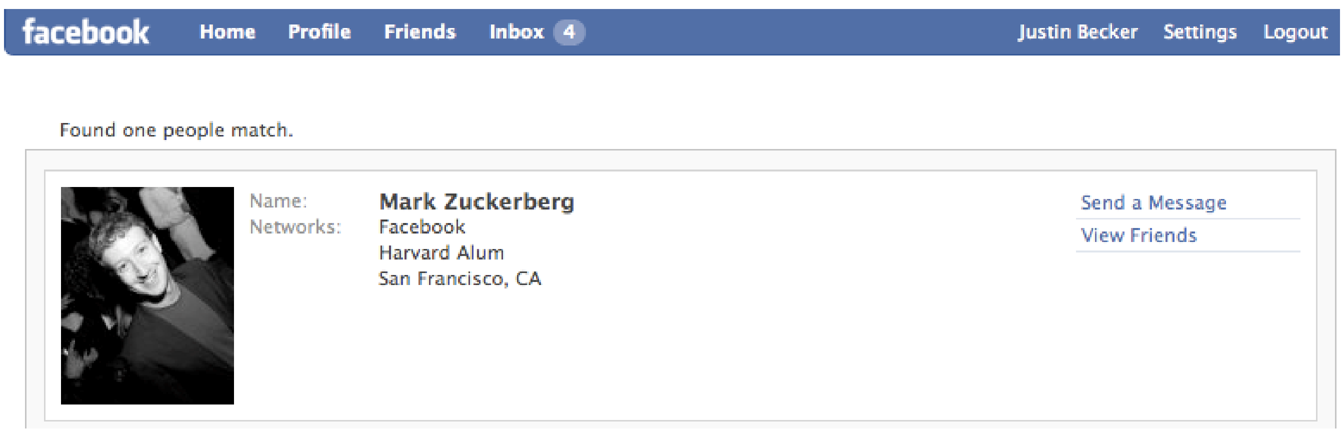


# Threat model

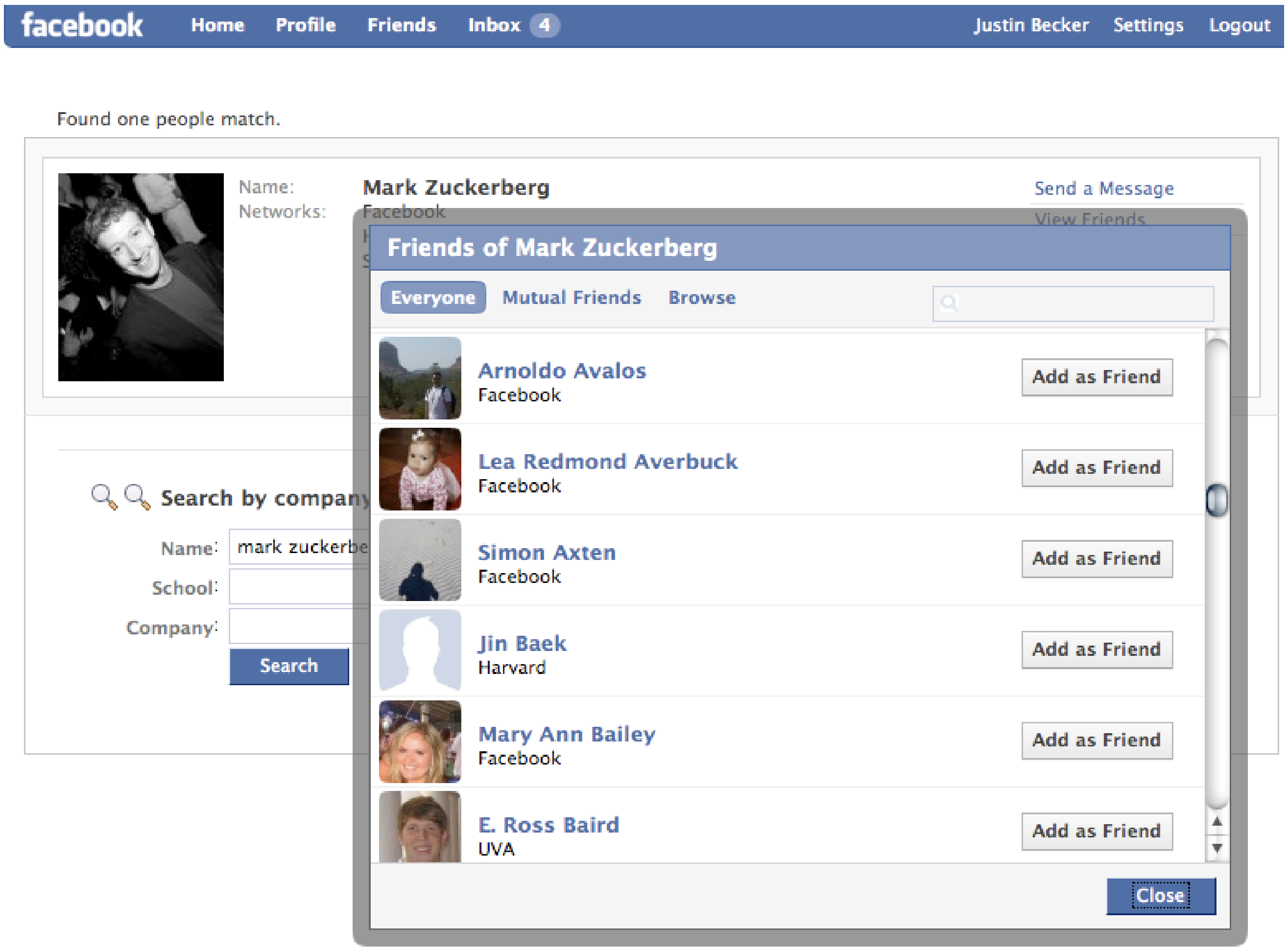


# Example

Can we derive a user affiliation from their friends?



# Example


facebook
Home
Profile
Friends
Inbox 4
Justin Becker
Settings
Logout
Found one people match.
Name:
Mark Zuckerberg
Networks:
Facebook
Send a Message
View Friends
Friends of Mark Zuckerberg
Everyone
Mutual Friends
Browse
Arnoldo Avalos
Facebook
Add as Friend
Lea Redmond Averbuck
Facebook
Add as Friend
Simon Axten
Facebook
Add as Friend
Jin Baek
Harvard
Add as Friend
Mary Ann Bailey
Facebook
Add as Friend
E. Ross Baird
UVA
Add as Friend
Close
Search by company
Name:
mark zuckerbe
School:
Company:
Search

# Example

| Affiliation | Frequency |
|---|---|
| Facebook | 32 |
| Harvard | 17 |
| San Francisco | 8 |
| Silicon Valley | 4 |
| Berkeley | 2 |
| Google | 2 |
| Stanford | 2 |

# PrivAware implementation

- A user must agree to install PrivAware
- Due to Facebook's liberal privacy policy PrivAware can
  - Access the user'*s* profile
  - Access the profiles of all the user's direct friends

# Threats

1) Friend threat
    - Derive private attributes via mutual friends
2) Non-friend threat
    - Derive private attributes via friends public attributes
    - Derive private attributes via mutual friends
3) Malicious applications
    - Derive private attributes via friends public attributes

# Inferring attributes

Algorithm: select the most frequent attribute value among the user's friends

**Friend attributes**

| | |
|---|---|
| Education | [**UC Davis**:7, Stanford:2, UCLA:4] |
| Employer | [**Google**:10, LLNL:8, Microsoft:2 ] |
| Relationship | [**Married**:9, Single:5, In a relationship:7] |

**Inferred values**

| | |
|---|---|
| Education | **UC Davis** |
| Employer | **Google** |
| Relationship | **Married** |

# Evaluation metrics

1) Inferable attributes
   - Attribute can be inferred

2) Verifiable inferences
   - Inferred attributes can be validated against profile

3) Correct inferences
   - Verifiable inferences equals profile attribute

# Validation example

| Classification | Score |
|---|---|
| Inferred attributes | 3 |
| Verifiable inferences | 2 |
| Correct inferences | 1 |

**Inferred values**

| | |
|---|---|
| Education | UC Davis |
| Employer | Google |
| Relationship status | Married |

**Actual values**

| | |
|---|---|
| Education | UC Davis |
| Employer | LLNL |

# Data disambiguation

Decide if different attribute values are **semantically equal**

Variants for University of California, Berkeley

- UC Berkeley
- Berkeley
- Cal

# Approaches for Disambiguation

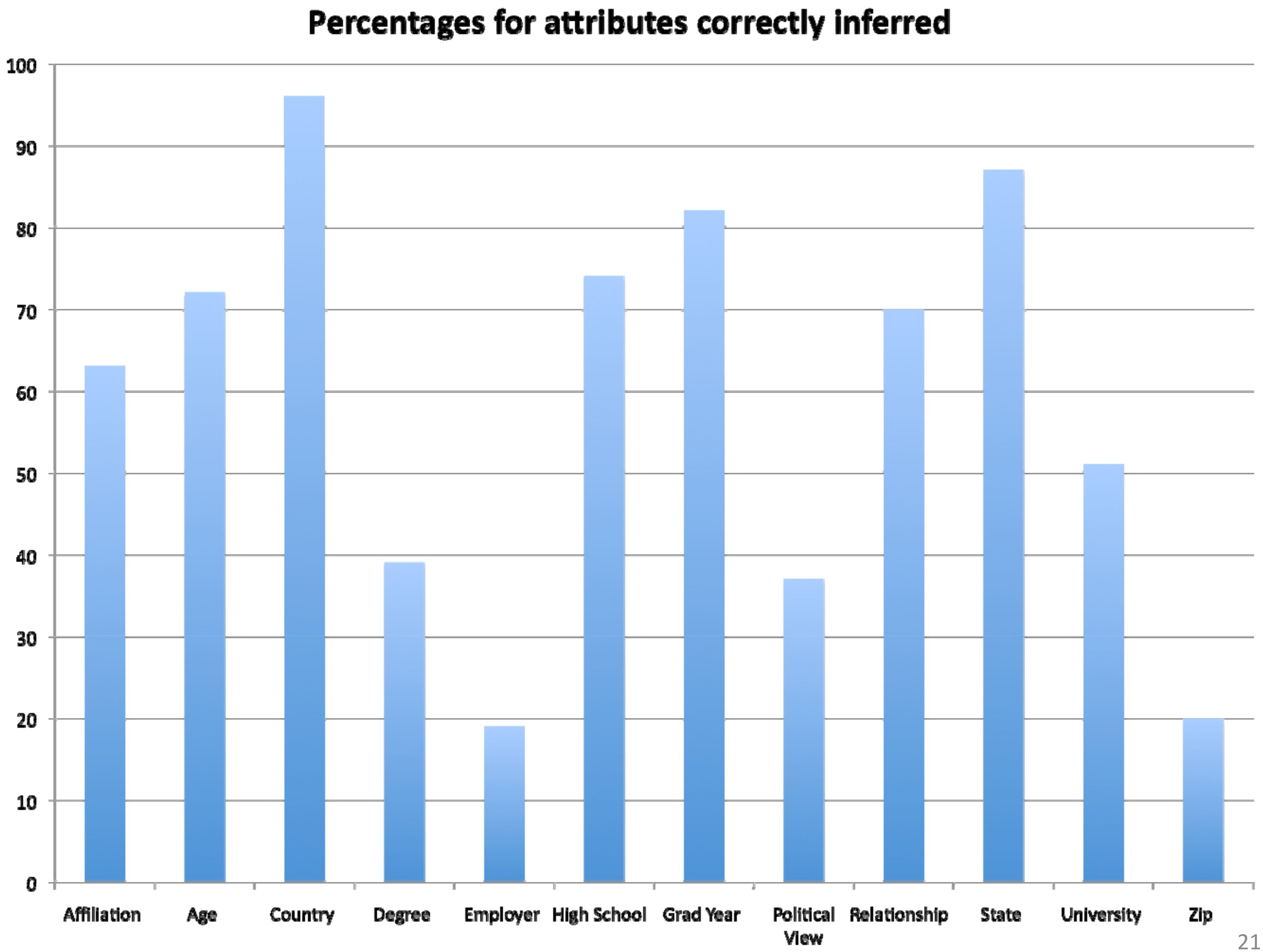
- Dictionary lookup
  - Keywords and synonyms
- Edit distance
  - Levenstein algorithm
- Named entity recognition

# Social contacts

| | |
|---|---|
| Total people | 93 |
| Total social contacts | 12,523 |
| Average social contacts / person | 134 |

# Inference results

| | |
|---|---:|
| Total inferred attributes | 1,673 |
| Total verifiable inferences | 918 |
| Total attributes correctly inferred | 546 |
| Correctly inferred | 60% |


Percentages for attributes correctly inferred
100
90
80
70
60
50
40
30
20
10
0
Affiliation
Age
Country
Degree
Employer
High School
Grad Year
Political View
Relationship
State
University
Zip

# Inference prevention

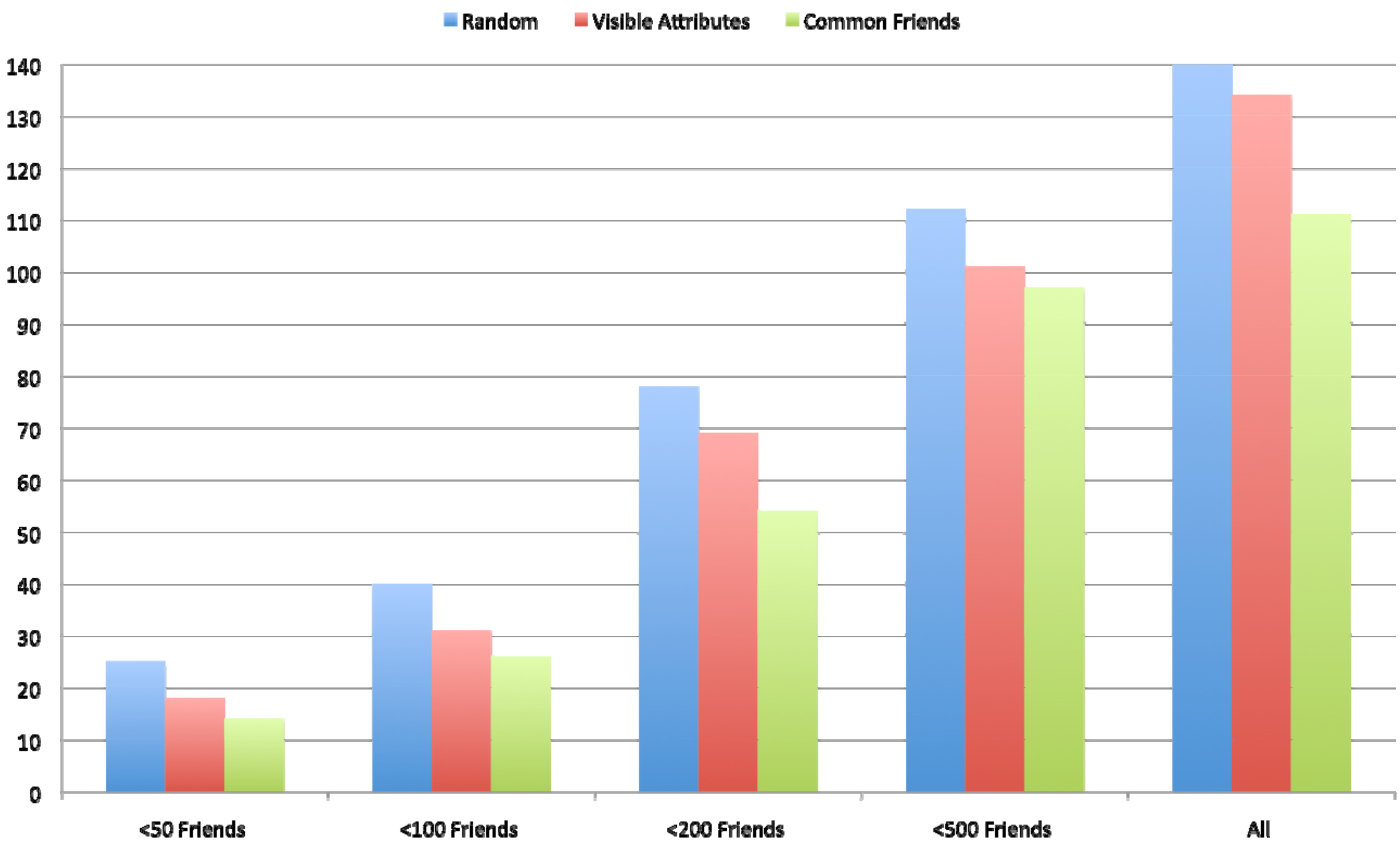
- Goals
  - Minimize the number of inferable attributes
  - Maximize the number of friends
- Approaches
  - Move risky friends into private groups
  - Delete risky friends

# Inference prevention

- Optimal solution
  - Derive privacy scores for each permutation of friends, select permutation with the lowest score
  - Runtime complexity: $O(2^n)$

# Inference prevention

- Heuristic approaches
  - Remove friends randomly
  - Remove friends with most attributes
  - Remove friends with most common friends


Random
Visible Attributes
Common Friends
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
<50 Friends
<100 Friends
<200 Friends
<500 Friends
All

# Related work

- *To join or not to join: The illusion of privacy in social networks…* [www2009]
- *On the need for user-defined fine-grained access control…*[CIKM 2008]
- *Link privacy in social networks* [SOSOC 2008]
- *Privacy Protection for Social Networking Platforms* [W2SP 2008]

# Future work

- Improve existing algorithms
  - NLP techniques
  - Data mining applications
- Include additional threat models
  - User updates
  - Friends tagging content
  - Fan pages
- Expand into domains other than social networks
  - Email
  - Search

# Conclusion

- Measure privacy risks caused by friends
- Improve privacy by identifying risky friends

On average, using the common friend heuristic, users need to delete or group **19 less users**, to meet their desired privacy level, **than randomly deleting** friends