Social privacy; Large-scale internet measurement to understand usable security and privacy

Mainack Mondal

CS 60081
Autumn 2022

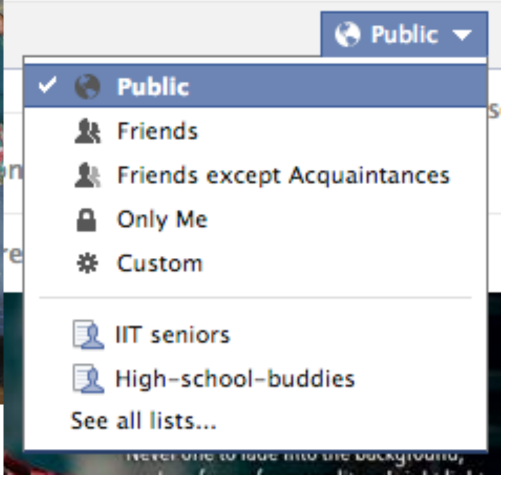
Doubt clearing session

Roadmap

- Privacy in social media; privacy in public
- Measuring usability via large scale internet measurement studies
    - Pros and Cons

Privacy violations in social media

reddit



Privacy violation in real world from user's point of view:

If someone accesses content who the user did not intend

ACLs are inadequate to capture many such privacy violations

Privacy in public (social media)

Understanding	solution
1. Information Revelation and Privacy in Online Social Networks, Acquisti and Gross, WPES’05 2. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook, Acquisti and Gross, PETS’06 3. Analyzing Facebook Privacy Settings: User Expectations vs. Reality, Liu et al. , IMC’2011 4. Quantifying the Invisible Audience in Social Networks, Bernstein et. al., CHI’2013	1. Privacy Wizards for Social Networking Sites, Fang et. al., WWW'2010
1. Understanding and Specifying Social Access Control Lists, Mondal et. al. SOUPS’14	

General methodology for each of these papers

- Step 1: Collect behavioral data by passively observing users (via collecting “public” data, sometimes consent)
- Step 2: [Optional] use the collected data to ground surveys and capture user expectations or desires about privacy
- Step 3: Analyze the data and identify patterns to answer research questions

General methodology for each of these papers

- Step 1: Collect behavioral data by passively observing users (via collecting “public” data, sometimes consent)
- Step 2: [Optional] use the collected data to ground surveys and capture user expectations or desires about privacy
- Step 3: Analyze the data and identify patterns to answer research questions

Large-scale internet measurement to understand usable security and privacy

Roadmap

- Privacy in social media; privacy in public

- Measuring usability via large scale internet measurement studies
    - Pros and Cons

Slide to describe papers often borrowed from presentations of respective authors

Privacy in public (social media)

Understanding	solution
1. Information Revelation and Privacy in Online Social Networks, Acquisti and Gross, WPES’05 2. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook, Acquisti and Gross, PETS’06 3. Analyzing Facebook Privacy Settings: User Expectations vs. Reality, Liu et al. , IMC’2011 4. Quantifying the Invisible Audience in Social Networks, Bernstein et. al., CHI’2013	1. Privacy Wizards for Social Networking Sites, Fang et. al., WWW'2010
1. Understanding and Specifying Social Access Control Lists, Mondal et. al. SOUPS’14	

Data gathering

- In June 2005, searched for CMU Facebook members’ profiles using advanced search feature and extracted profile IDs
    - Downloaded 4,540 profiles
    - Inferred additional information not immediately visible from profiles

Demographics

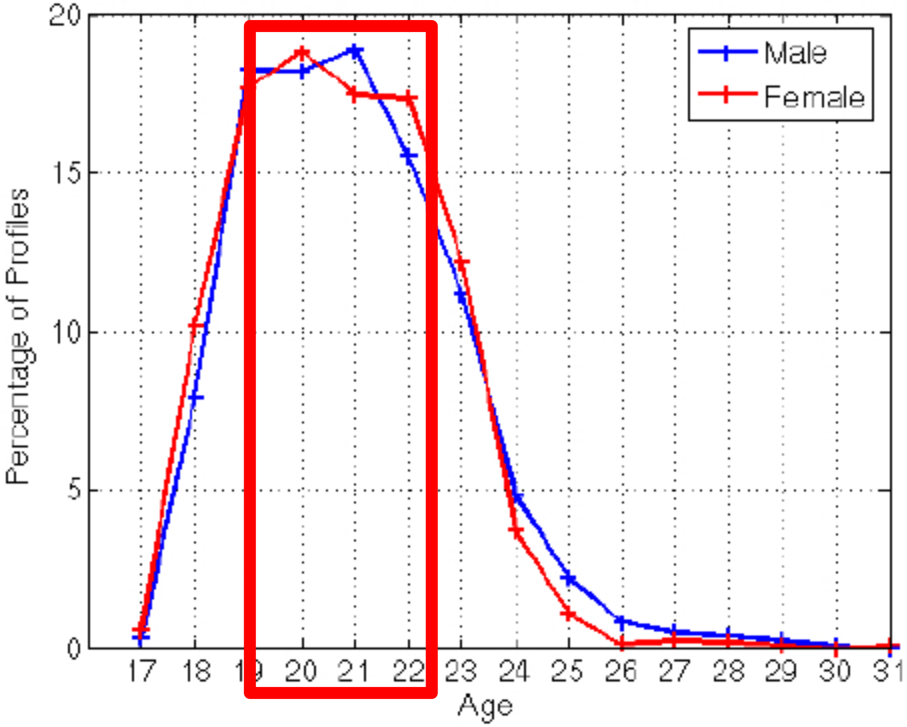


Figure 1: Age distribution of Facebook profiles at CMU. The majority of users (95.6%) falls into the 18-24 age bracket.

Demographics

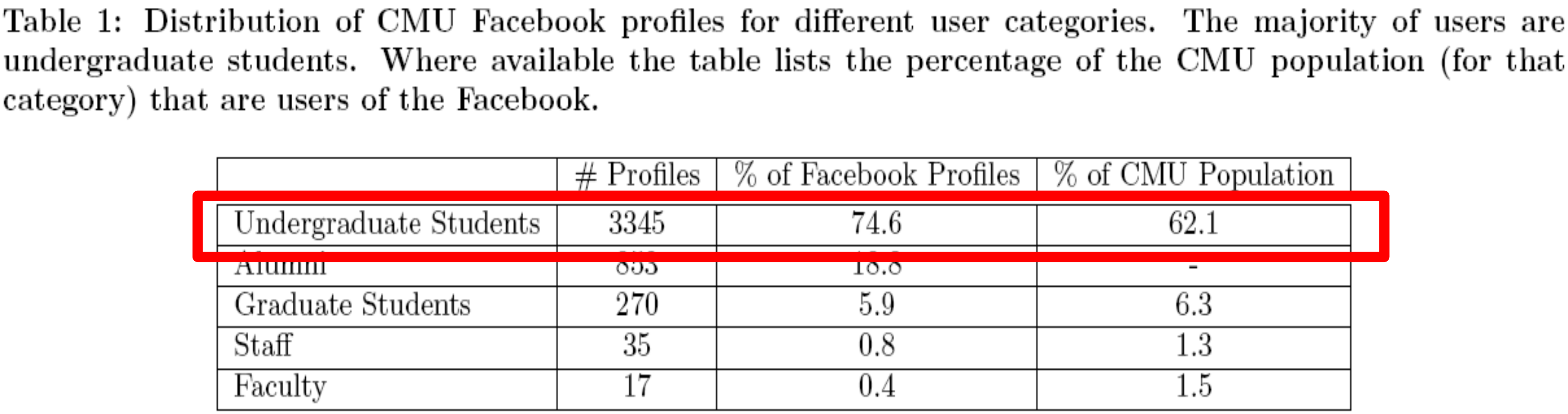


Table 1: Distribution of CMU Facebook profiles for different user categories. The majority of users are undergraduate students. Where available the table lists the percentage of the CMU population (for that category) that are users of the Facebook.

	# Profiles	% of Facebook Profiles	% of CMU Population
Undergraduate Students	3345	74.6	62.1
Alumni	853	18.8	-
Graduate Students	270	5.9	6.3
Staff	35	0.8	1.3
Faculty	17	0.4	1.5

Demographics

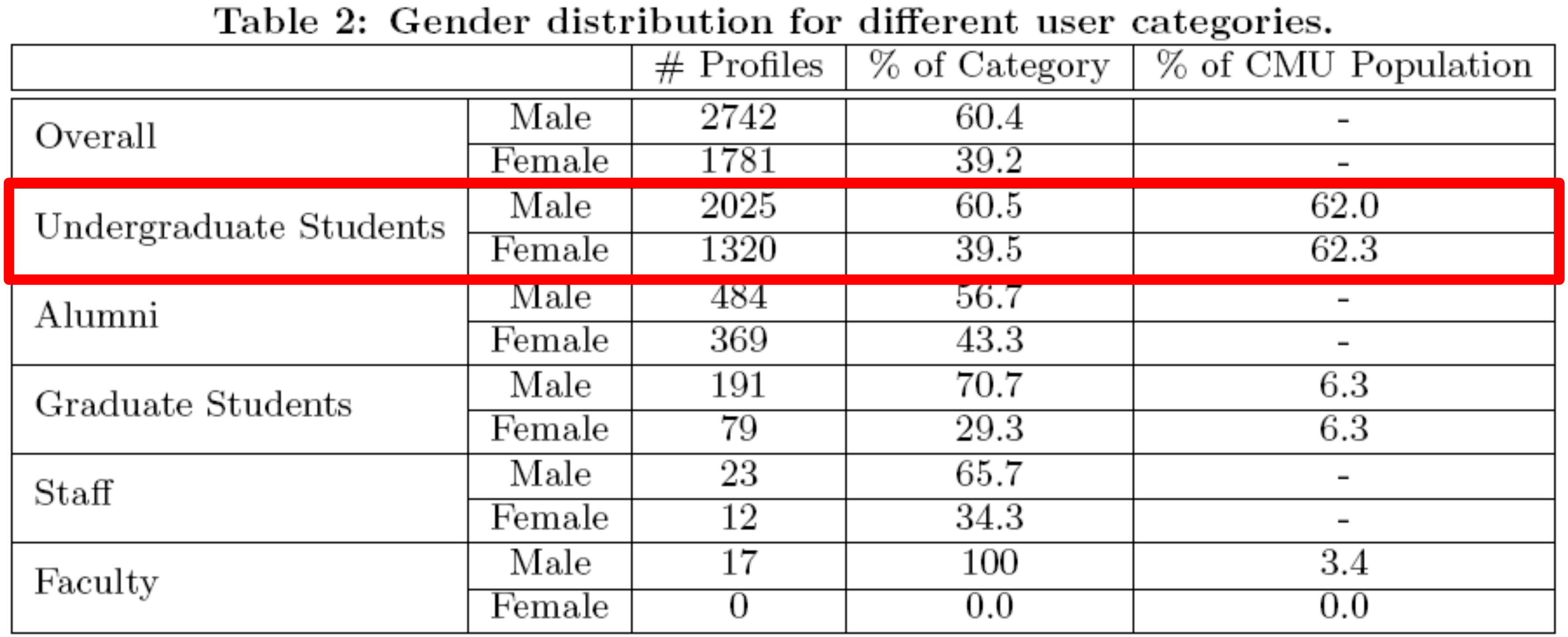


Table 2: Gender distribution for different user categories.

		# Profiles	% of Category	% of CMU Population
Overall	Male	2742	60.4	-
	Female	1781	39.2	-
Undergraduate Students	Male	2025	60.5	62.0
	Female	1320	39.5	62.3
Alumni	Male	484	56.7	-
	Female	369	43.3	-
Graduate Students	Male	191	70.7	6.3
	Female	79	29.3	6.3
Staff	Male	23	65.7	-
	Female	12	34.3	-
Faculty	Male	17	100	3.4
	Female	0	0.0	0.0

Information revelation

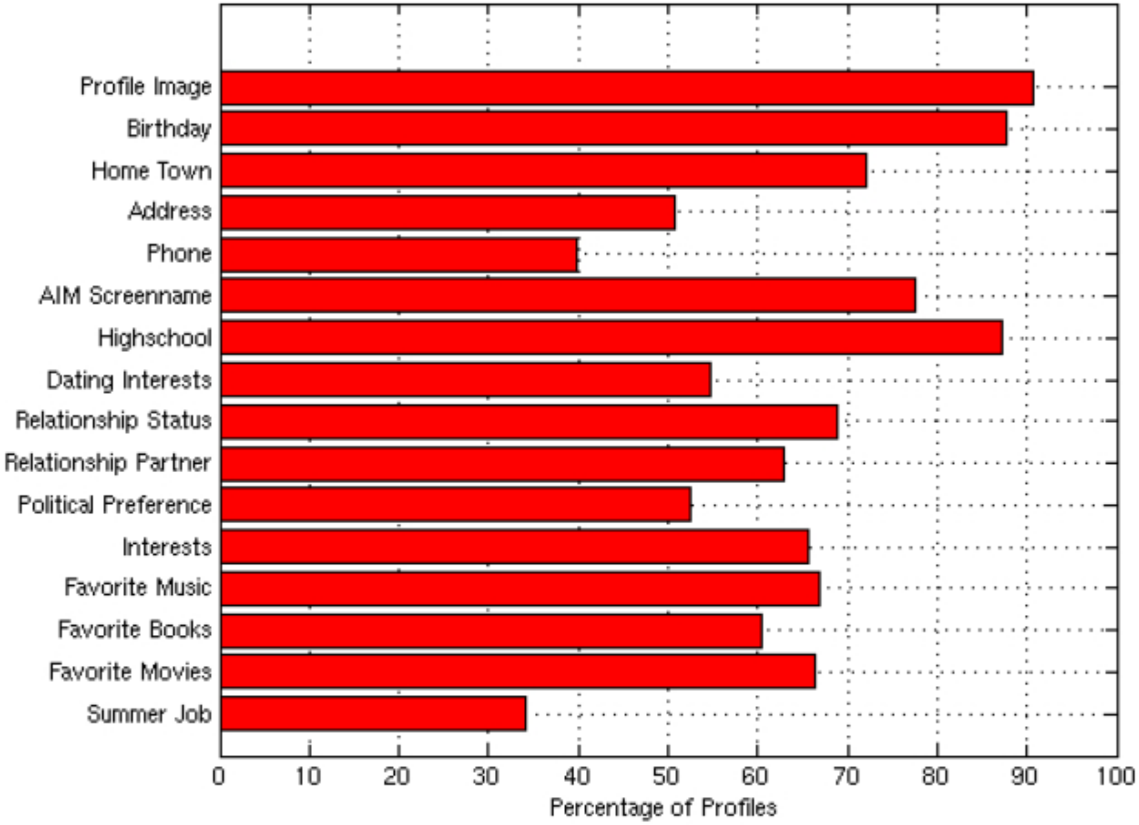


Figure 2: Percentages of CMU profiles revealing various types of personal information.

Patterns in revealed information

- Male users 63% more likely to leave phone number than female users

- Single male users tend to report their phone numbers in even higher frequencies

Verifying the data: real name

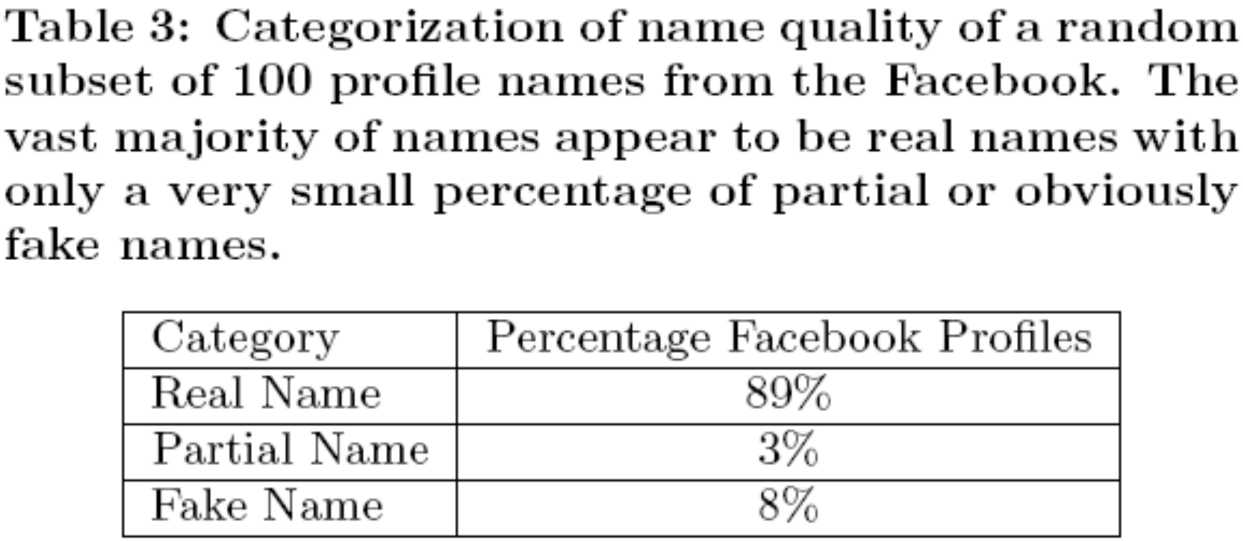


Table 3: Categorization of name quality of a random subset of 100 profile names from the Facebook. The vast majority of names appear to be real names with only a very small percentage of partial or obviously fake names.

Category	Percentage Facebook Profiles
Real Name	89%
Partial Name	3%
Fake Name	8%

Verifying the data: contain image

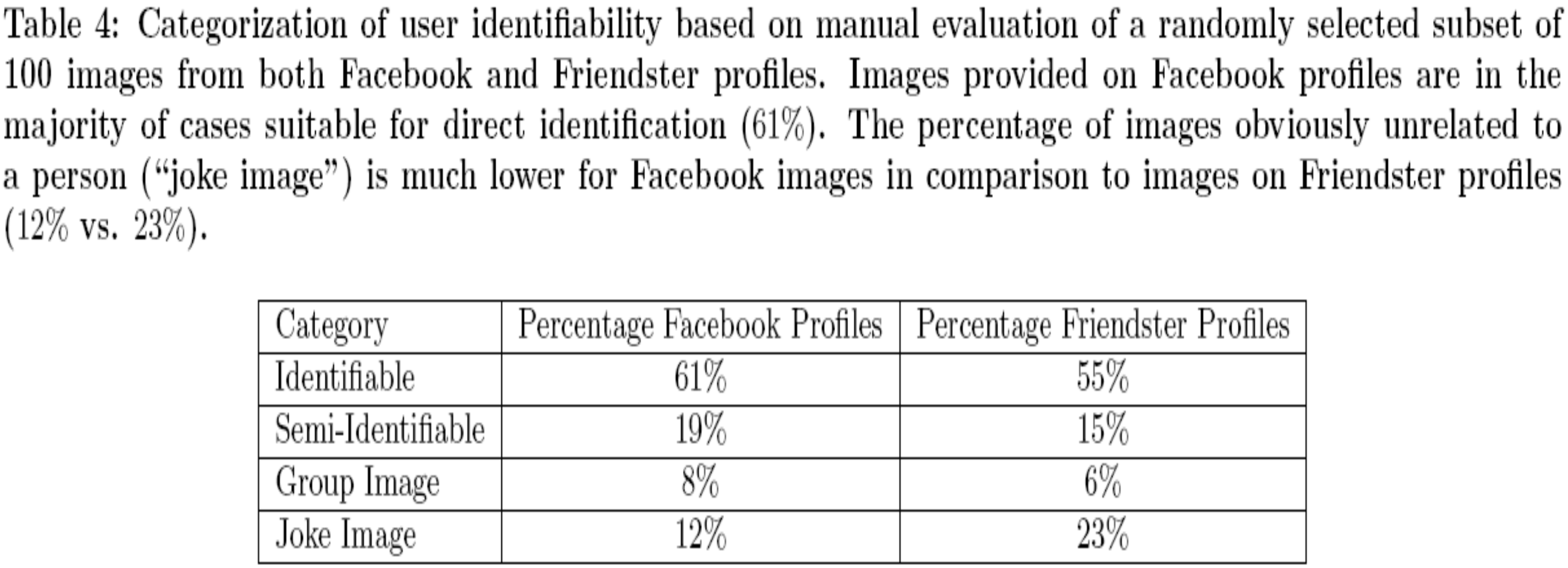


Table 4: Categorization of user identifiability based on manual evaluation of a randomly selected subset of 100 images from both Facebook and Friendster profiles. Images provided on Facebook profiles are in the majority of cases suitable for direct identification (61%). The percentage of images obviously unrelated to a person ("joke image") is much lower for Facebook images in comparison to images on Friendster profiles (12% vs. 23%).

Category	Percentage Facebook Profiles	Percentage Friendster Profiles
Identifiable	61%	55%
Semi-Identifiable	19%	15%
Group Image	8%	6%
Joke Image	12%	23%

Always need to have a baseline

Privacy risks from …

- Stalking
- Re-identification
- Digital dossier

Privacy risks: Stalking

- Real-World Stalking
  - College life centers around class attendance
  - Facebook users put home address and class list on their profiles; whereabouts are known for large portions of the day

- Online stalking
  - Facebook profiles list AIM screennames
  - AIM lets users add “buddies” without notification
  - Unless AIM privacy settings have been changed, adversary can track when user is online

Privacy risks: Re-identification

- Demographics re-identification
  - 87% of US population is uniquely identified by *{gender, ZIP, date of birth}* (Sweeney, 2001)
  - Facebook users that put this information up on their profile could link them up to outside, de-identified data sources
- Face re-identification
  - Facebook profiles often show high quality facial images
  - Images can be linked to de-identified profiles on e.g. Match.com or Friendster.com using face recognition
- Social Security Number re-identification
  - Anatomy of a social security number: *xxx yy zzzz*
  - Based on hometown and date of birth *xxx* and *yy* can be narrowed down substantially

Privacy risks: Digital Dossier

- Users reveal sensitive information (e.g. current partners, political views) in profiles

- Simple script programs allow adversaries to continuously retrieve and save all profile information

- Cheap hard drives enable essentially indefinite storage

Privacy risks

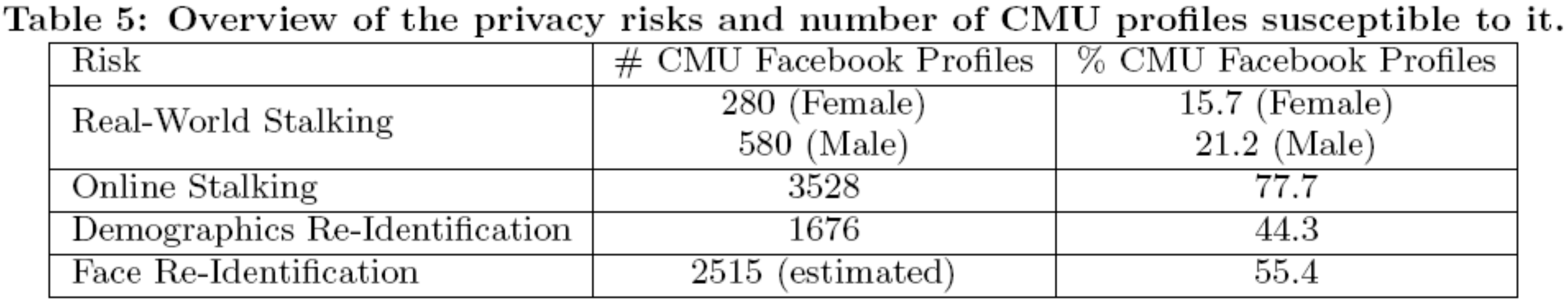


Table 5: Overview of the privacy risks and number of CMU profiles susceptible to it.

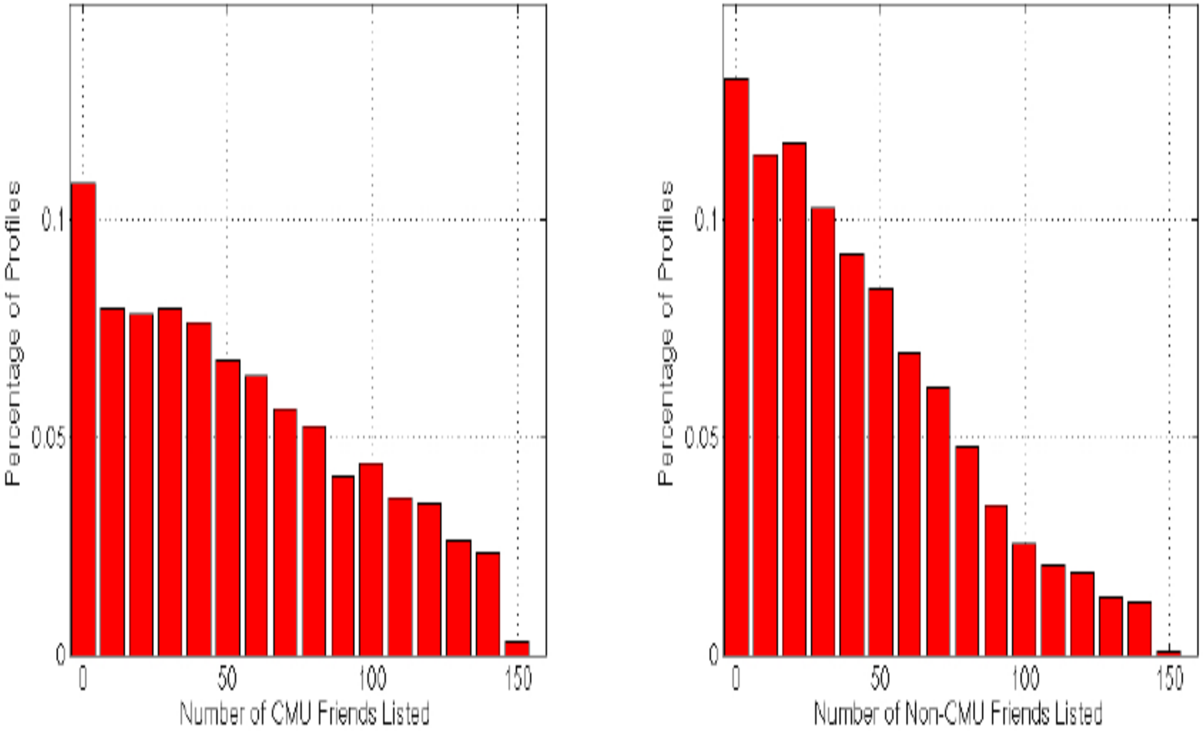
Risk	# CMU Facebook Profiles	% CMU Facebook Profiles
Real-World Stalking	280 (Female) 580 (Male)	15.7 (Female) 21.2 (Male)
Online Stalking	3528	77.7
Demographics Re-Identification	1676	44.3
Face Re-Identification	2515 (estimated)	55.4

Data accessibility

- Profile Searchability
  - Measured the percentage of users that changed search default setting away from being searchable to everyone on the Facebook to only being searchable to CMU users
  - 1.2% of users (18 female, 45 male) made use of this privacy setting
- Profile Visibility
  - Evaluated the number of CMU users that changed profile visibility by restricting access from unconnected users
  - Only 3 profiles (0.06%) in total fall into this category

- *Caveat: They would not detect users who had made themselves both unsearchable and invisible within CMU network (safe to assume their number is very low)*

Data accessibility



(a) Network of CMU friends

(b) Network of Non-CMU friends

Figure 3: Histogram of the size of networks for both CMU friends (a) and non-CMU friends (b). Users maintain large networks of friends with the average user having 78.2 friends at CMU and 54.9 friends elsewhere.

How many of the students were aware of the problems?

Actual data accessibility: An imagined community?

- Extensive, uncontrolled social networks
- Fragile protection:
  - Fake email addresses
  - Manipulating users
  - Geographical location
  - Advanced search features
    - Using advanced search features various profile information can be searched for, e.g. relationship status, phone number, sexual preferences, political views and (college) residence
    - By keeping track of the profile IDs returned in the different searches a significant portion of the previously inaccessible information can be reconstructed
- *Facebook profiles are, effectively, public data*

Actual data accessibility vs. perceived: An imagined community

- “What a great illustration of how things you might not mind being public in one context can cause all sorts of problems when they wind up globally public.”
  - CMU student

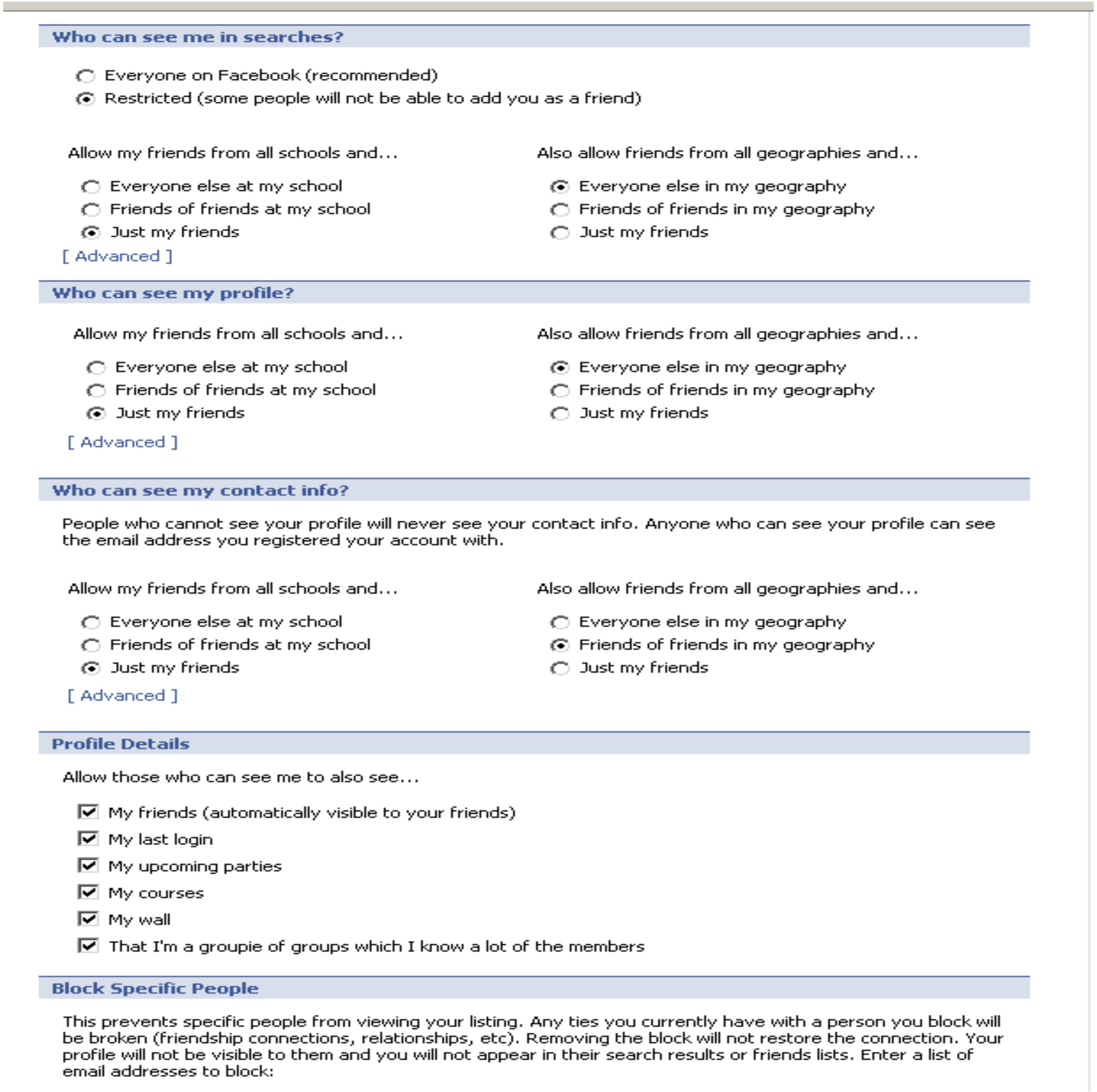
Initial hypotheses/Research Q

- Default settings (Mackay 1991)/ Myopic discounting?
  - Less than 2% make their profiles less searchable
  - Less than 1% make their profiles less visible
- Peer pressure
- Incomplete information and biased perspectives
  - An *imagined* community

- Or simply:
  - Low privacy concerns
  - Signaling
    - Single males list phone number with highly significant more frequency than females

Survey set up

- Goals
  - Understand CMU Facebook’s users degree of awareness about the site and its information revelation patterns; understand their privacy attitudes and expectations
- Thirty-six online questions
- Anonymous, paid
- Set up
  - 294 respondants
  - Focused on Facebook users



Who can see me in searches?

○ Everyone on Facebook (recommended)
● Restricted (some people will not be able to add you as a friend)

Allow my friends from all schools and...

○ Everyone else at my school
○ Friends of friends at my school
● Just my friends

[Advanced]

Also allow friends from all geographies and...

● Everyone else in my geography
○ Friends of friends in my geography
○ Just my friends

Who can see my profile?

Allow my friends from all schools and...

○ Everyone else at my school
○ Friends of friends at my school
● Just my friends

[Advanced]

Also allow friends from all geographies and...

● Everyone else in my geography
○ Friends of friends in my geography
○ Just my friends

Who can see my contact info?

People who cannot see your profile will never see your contact info. Anyone who can see your profile can see the email address you registered your account with.

Allow my friends from all schools and...

○ Everyone else at my school
○ Friends of friends at my school
● Just my friends

[Advanced]

Also allow friends from all geographies and...

○ Everyone else in my geography
● Friends of friends in my geography
○ Just my friends

Profile Details

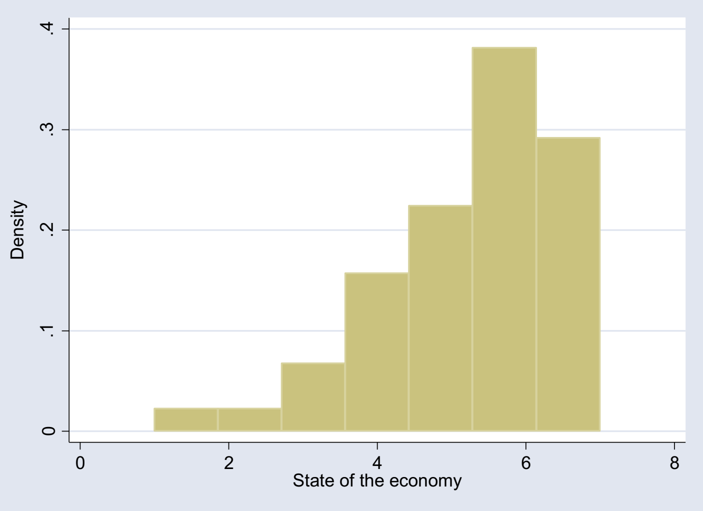
Allow those who can see me to also see...

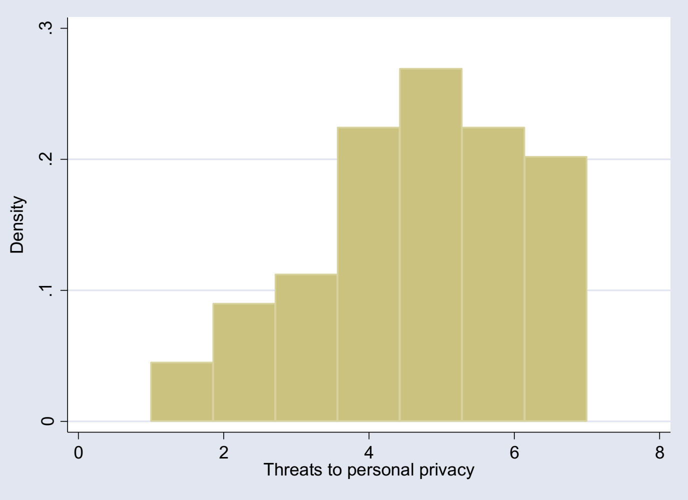
☑ My friends (automatically visible to your friends)
☑ My last login
☑ My upcoming parties
☑ My courses
☑ My wall
☑ That I'm a groupie of groups which I know a lot of the members

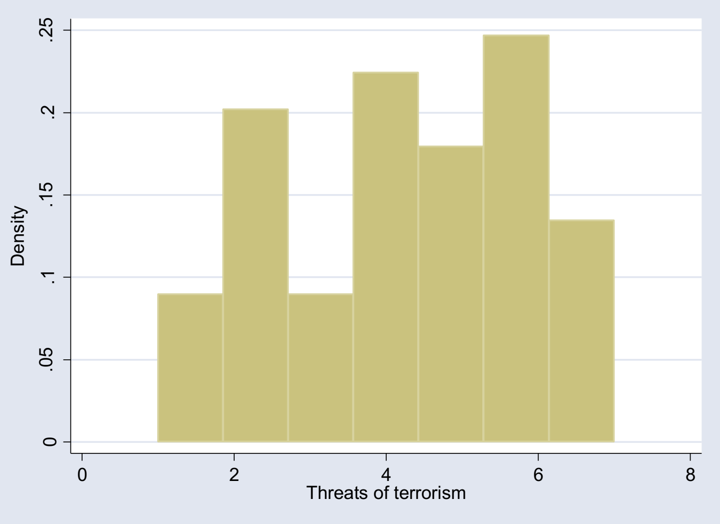
Block Specific People

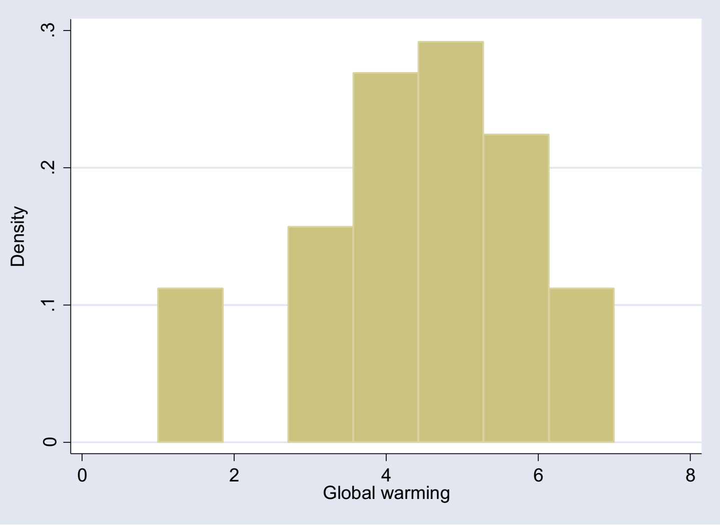
This prevents specific people from viewing your listing. Any ties you currently have with a person you block will be broken (friendship connections, relationships, etc). Removing the block will not restore the connection. Your profile will not be visible to them and you will not appear in their search results or friends lists. Enter a list of email addresses to block:

Generic concerns (7-point Likert scale)

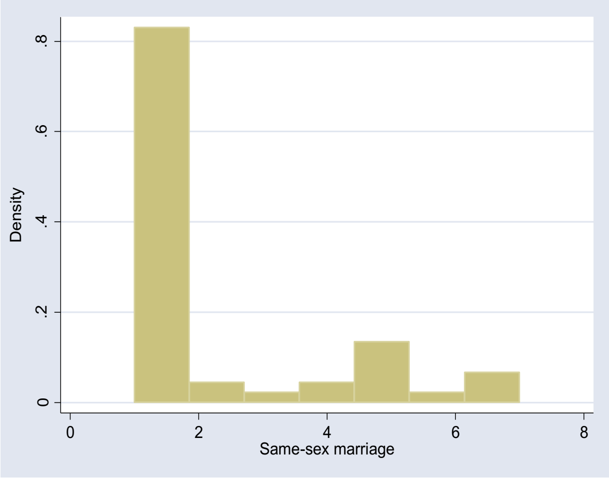


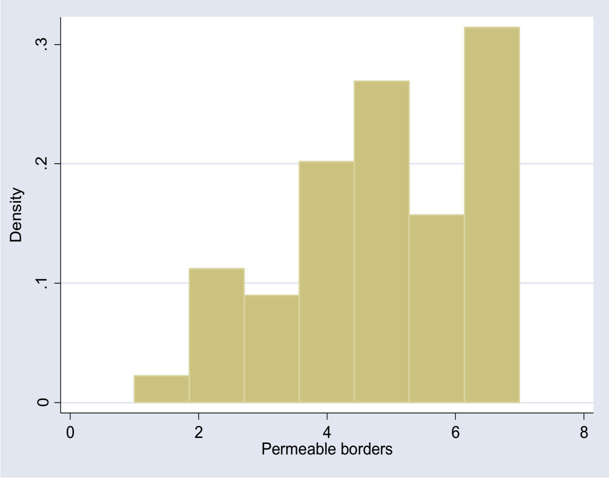


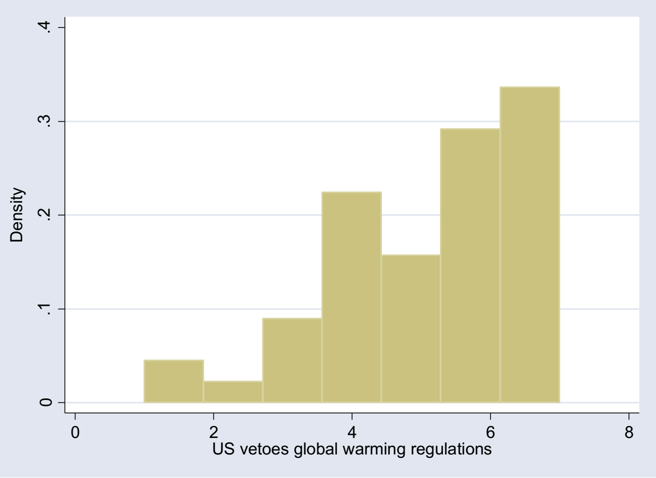


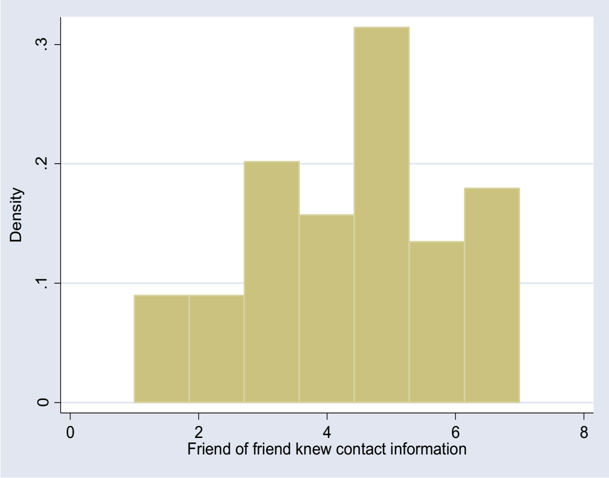


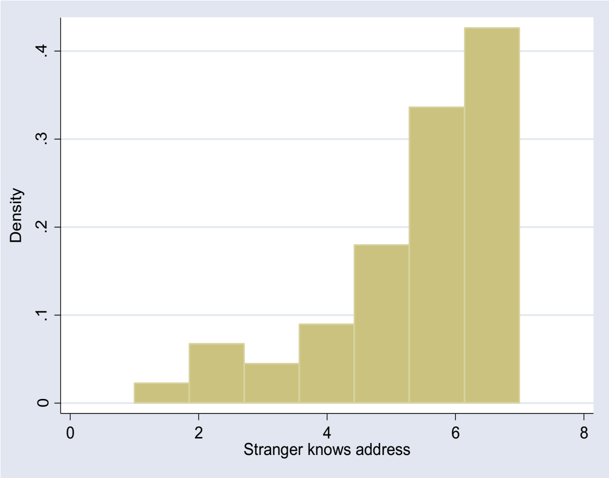
Specific concerns (7-point Likert scale)

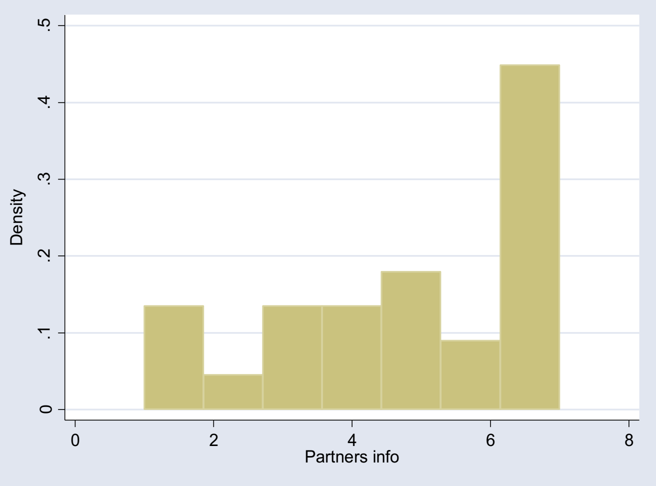












Attitudes vs. behavior

- Share of users with high sensitivity (Likert >5) to partner/sexual orientation information who provide it on Facebook: ~70%
- Share of users with high sensitivity (Likert >5) to home location and class schedule information who provide it on Facebook: ~32%
- Share of users with high sensitivity (Likert >5) to contact information who provide it on Facebook: ~42%

Awareness: Visibility and searchability

- 21% incorrectly believe only CMU users can search their profiles
- 71% do not realize that everybody at can search their profiles
- 40% do not realize that anybody on Facebook can search their profiles
- 31% do not realize that everybody at CMU can read their profiles
- On the other side, 23% incorrectly believe that everybody on Facebook can read their profiles

Facebook‘s privacy policy: Perception vs. relaity

“Facebook also collects information about you from other sources, such as newspapers and instant messaging services. This information is gathered regardless of your use of the Web Site.”

- 85% believe that is *not* the case

“We use the information about you that we have collected from other sources to supplement your profile unless you specify in your privacy settings that you do not want this to be done.”

- 87% believe that is *not* the case

“In connection with these offerings and business operations, our service providers may have access to your personal information for use in connection with these business activities.”

- 60% believe that is *not* the case
- Control: perusal of privacy policy does *not* improve awareness

Information revelation

- Reasons to provide more personal information (in order of importance):

1. No factor in particular, it's just fun
2. No factor in particular, but the amount of information I reveal is necessary to me and other users to benefit from the *FaceBook*
3. No factor in particular, rather I am following the norms and habits common on the site
4. Quite simply, expressing myself and defining my online persona
5. Showing more information about me to "advertise" myself

…..

- Getting more potential dates

Summary of Gross-acquisti studies

- Facebook users claim, in general, to be concerned about their privacy but
  - Publish plenty of personal information
  - Do not use privacy enhancing features
- However, they are both
  - …uninformed about specific information revelation patterns
  - … aware of generic possibilities
- Suggestive evidence pointing towards:
  - Signaling, but also
  - Myopic discounting
  - Incomplete information

Privacy in public (social media)

Understanding	solution
1. Information Revelation and Privacy in Online Social Networks, Acquisti and Gross, WPES’05 2. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook, Acquisti and Gross, PETS’06 3. Analyzing Facebook Privacy Settings: User Expectations vs. Reality, Liu et al. , IMC’2011 4. Quantifying the Invisible Audience in Social Networks, Bernstein et. al., CHI’2013	1. Privacy Wizards for Social Networking Sites, Fang et. al., WWW'2010
1. Understanding and Specifying Social Access Control Lists, Mondal et. al. SOUPS’14	

Motivation

- Our perception of audience size (how many people are watching) affect our behavior
  - We guide our audience's impression of us [Goffman'59]
  - We manage the boundaries of when to engage [Altman'75]
  - On social media, we speak to the audience that we expect is listening [Marwick and Boyd'11]

Do social media users know how many people are watching?

Goals

- Quantify the difference between users' estimated and actual audience
- Measure audience size uncertainty for 220,000 Facebook users

Two contributions

- Measuring perceived audience vs. reality
    - Survey
    - folk theories of audience
    - desired audience size

- Predicting audience size
    - using friend count
    - using feedback

Two contributions

- Measuring perceived audience vs. reality
    - Survey
    - folk theories of audience
    - desired audience size
- Predicting audience size
    - using friend count
    - using feedback

Survey setup

- 220,000 U.S. Facebook users who share with friends-only privacy
    - Collected audience information for their status updates and link shares over 30 days
    - 150,000,000 viewer-story pairs

- Measuring audience size using code
    - Javascript tracking whether a post stays in browser viewport for at least 900 ms
    - Approximate measure of attention – users remember 70% of the posts they see by 900 ms (Counts and Fisher 2011)

Recruitment

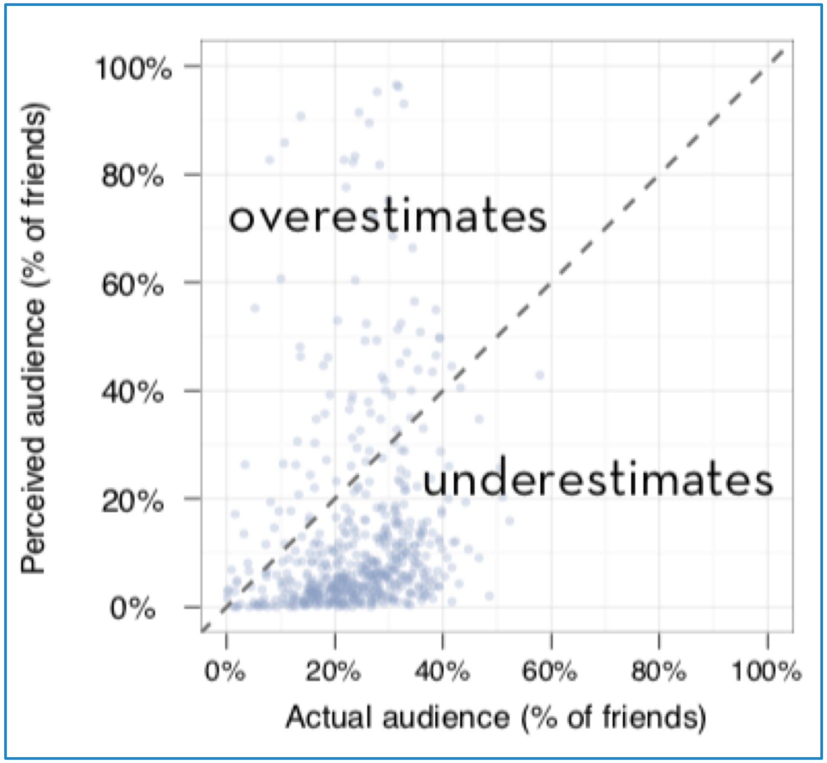
- Recruited users with recent content (2-90 days ago) via a request at the top of news feed
  - N=589; 61% female; mean age 33

Key survey question

- Perceived audience size survey: Show participants their most recent post and ask:
  - “How many people do you think saw it?”
  - “Describe how you came up with that number.”
  - “How many people do you wish saw this content?”

Consider your own most recent status update: What percentage of your social network do you think saw it?

Result: Underestimation by 4x



- **Estimated:** 20 friends = 6% of network
- **Actual:** 78 friends = 24% of network

Folk theories of audience

- Inductive coding on participants’ reasons for how they estimated their audience
  - Random guess 23%
  - Feedback — likes and comments 21%
  - Fraction of friend count 15%
  - Login timing 9%
  - Friends seen active on the site 5%
  - Number of close friends and family 3%
  - Who might be interested in the topic 2%
  - Other 10%

Summary of understanding

- Users underestimate their audience by 4x
- Common folk theories use feedback and friend count
- Users want larger audiences, but already have them

Privacy in public (social media)

Understanding	solution
1. Information Revelation and Privacy in Online Social Networks, Acquisti and Gross, WPES'05 2. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook, Acquisti and Gross, PETS'06 3. Analyzing Facebook Privacy Settings: User Expectations vs. Reality, Liu et al. , IMC'2011 4. Quantifying the Invisible Audience in Social Networks, Bernstein et. al., CHI'2013	1. Privacy Wizards for Social Networking Sites, Fang et. al., WWW'2010
1. Understanding and Specifying Social Access Control Lists, Mondal et. al. SOUPS'14	

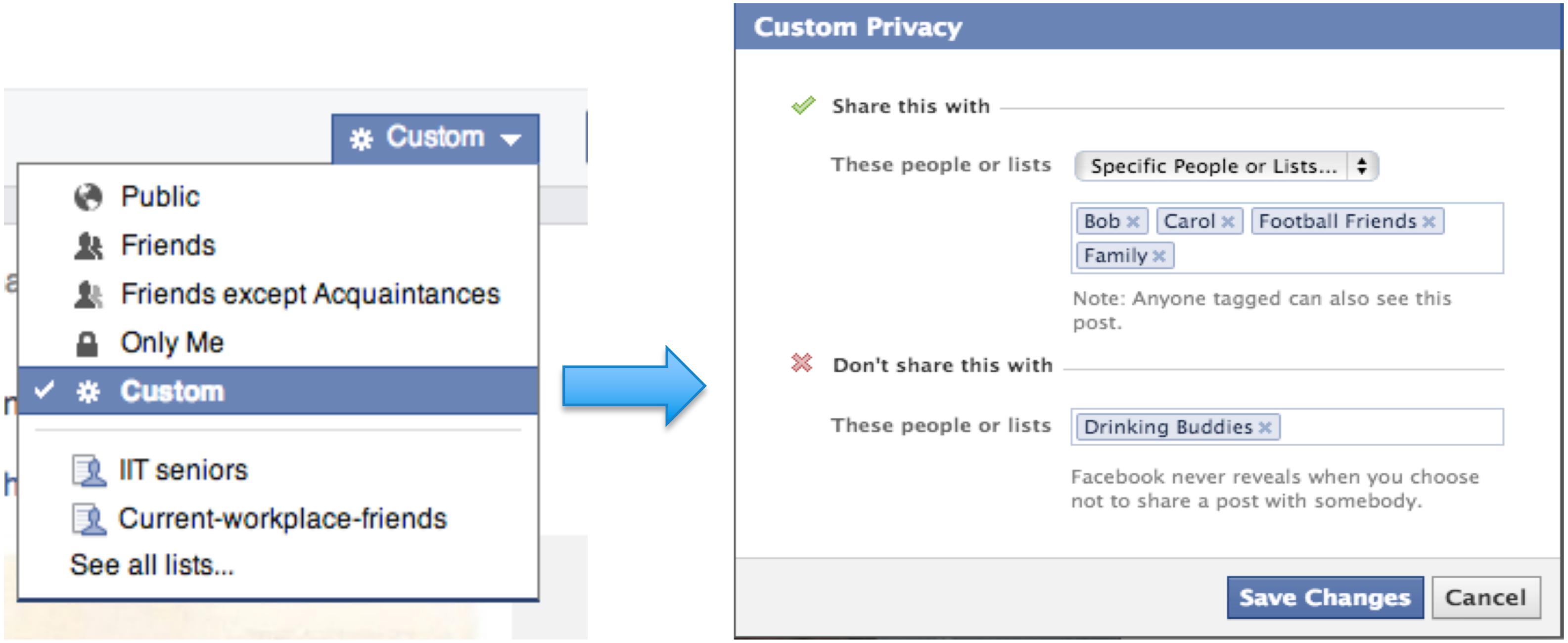
Privacy sensitive content in OSNs

Non privacy sensitive content: all friends should be able to access

Privacy sensitive Content: only select friends should be able to access

How do users manage access currently?

- Users specify Social Access Control Lists (SACLs)
- SACLs: Share with a subset of friends



Custom Privacy

✓ Share this with

These people or lists: Specific People or Lists...

Bob × Carol × Football Friends × Family ×

Note: Anyone tagged can also see this post.

✗ Don't share this with

These people or lists: Drinking Buddies ×

Facebook never reveals when you choose not to share a post with somebody.

State of the art for helping SACL specification

- Provide users **automatically detected groups**
  - **Network community** based group detection
  - **User profile attribute** based group detection
  - **User activity** based group detection
- Assumption by existing work:
  - **Automatically detected groups are similar to SACLs**

They evaluated their proposals based on small scale user interviews

Goal of this paper

To better understand real world SACL usage and specification

Friendlist Manager: Functionality

- We built and deployed Friendlist Manager (FLM) Facebook app
  - Help users to group their Facebook friends in **friendlists**
  - Available at: https://apps.facebook.com/friendlist_manager/
- 1,200+ users have installed FLM in 2 years!
- We asked consent from users to access their data
  - 1,100+ users gave consent
  - Collected a snapshot of all their profile and SACLs
  - All data collected and analyzed under IRB approval
- First large scale dataset of real in-use SACLs

User consent 👍 for data

How diverse is our user base?

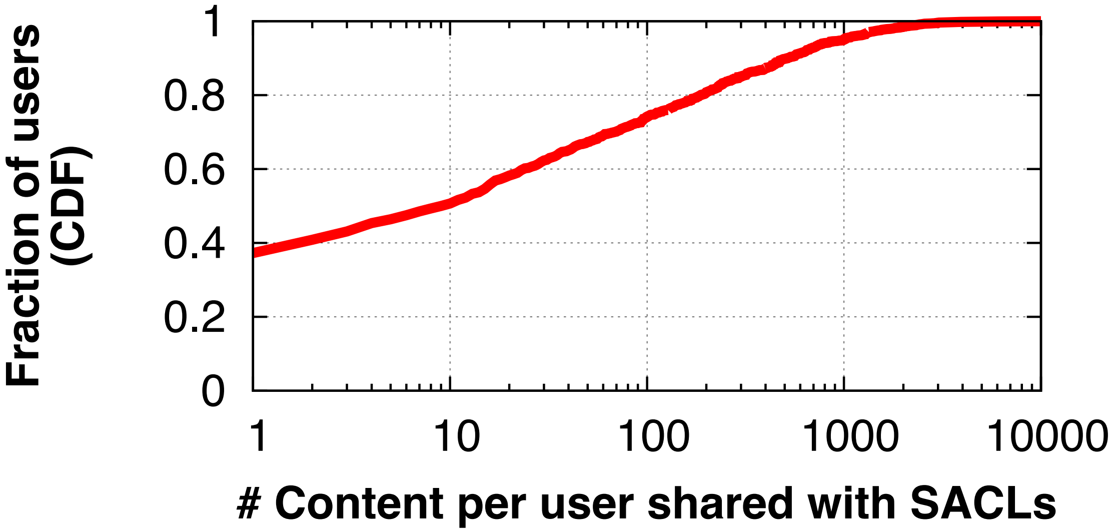
- Key demographic statistics
  - We have users from over 75 countries
  - Median age of our users: 29
  - A male bias (76% male)

- Our users are quite active on Faceboo
  - Median number of contents shared: 506

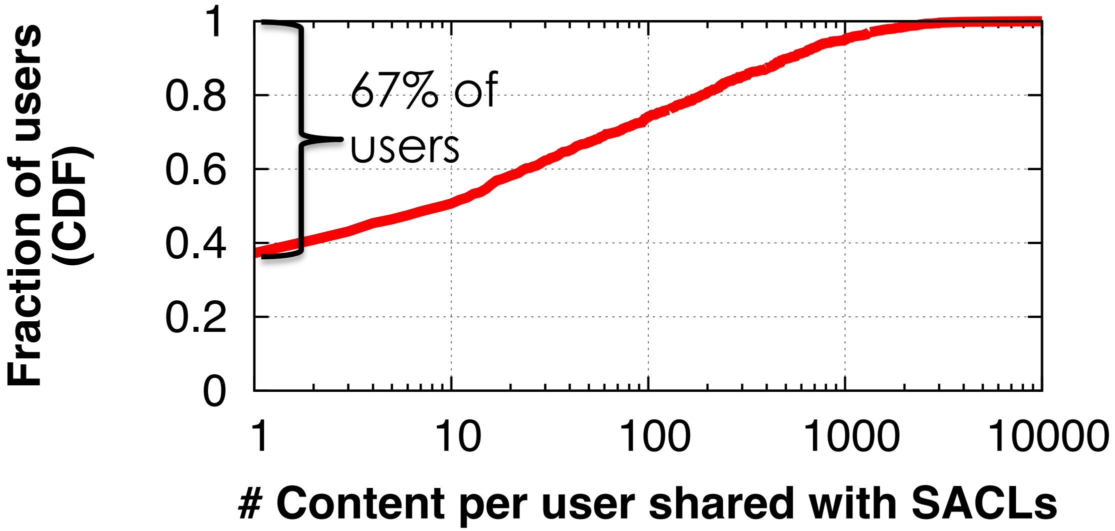
- These users are aware of Facebook friendlists
  - May be more privacy aware than a random user

Privacy
Aware

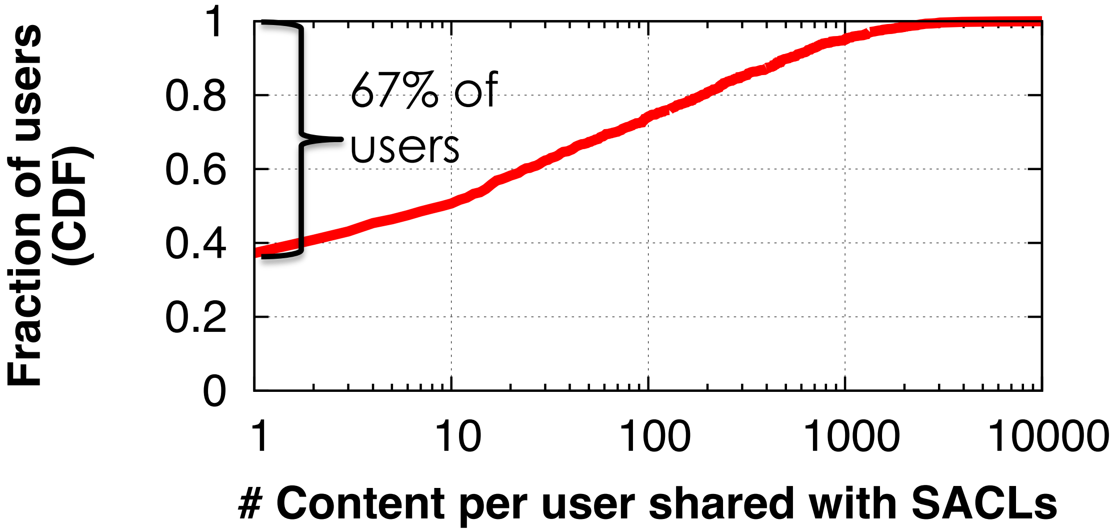
Do users use SACLs to share content?



Do users use SACLs to share content?

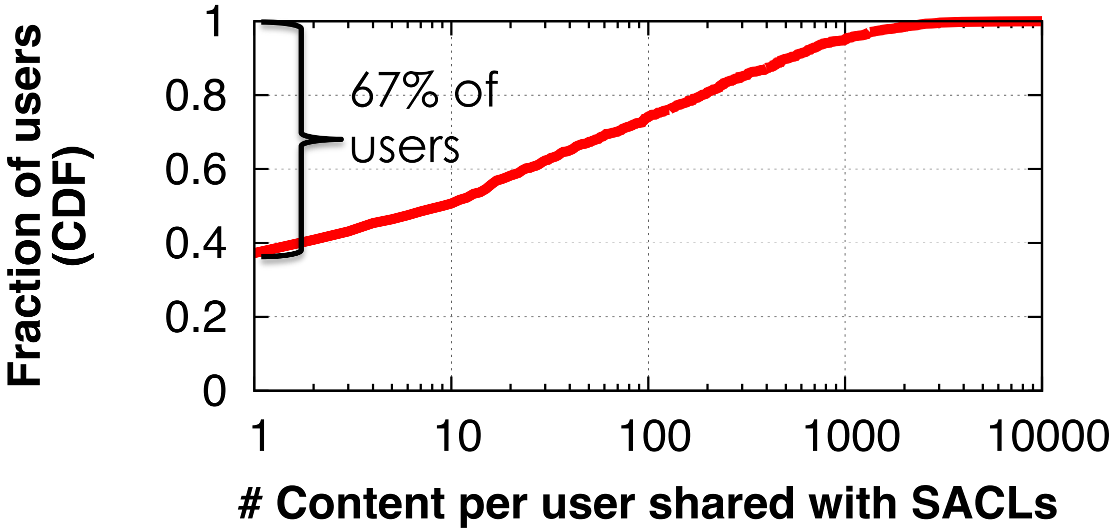


Do users use SACLs to share content?



- Total 200K content is shared with 7.6k unique SACLs!
- Majority of users used SACLs for at least one of their content

Do users use SACLs to share content?



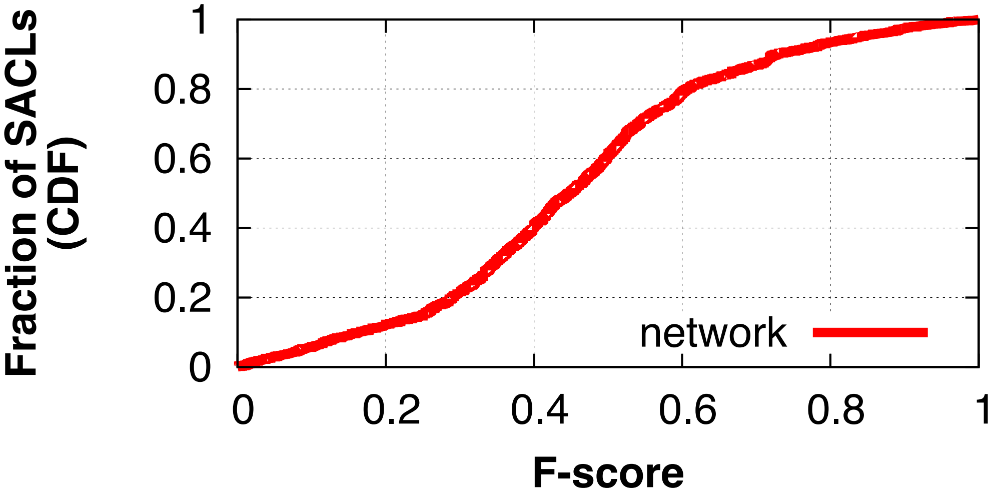
- Total 200K content is shared with 7.6k unique SACLs!
- Majority of users used SACLs for at least one of their content
- It is important to look for ways to simplify SACL specification

Do network groups correlate with SACLs?

- Ideal Case:
  - Groups by network community detection are highly similar to SACLs
- We measure F-score to check similarity

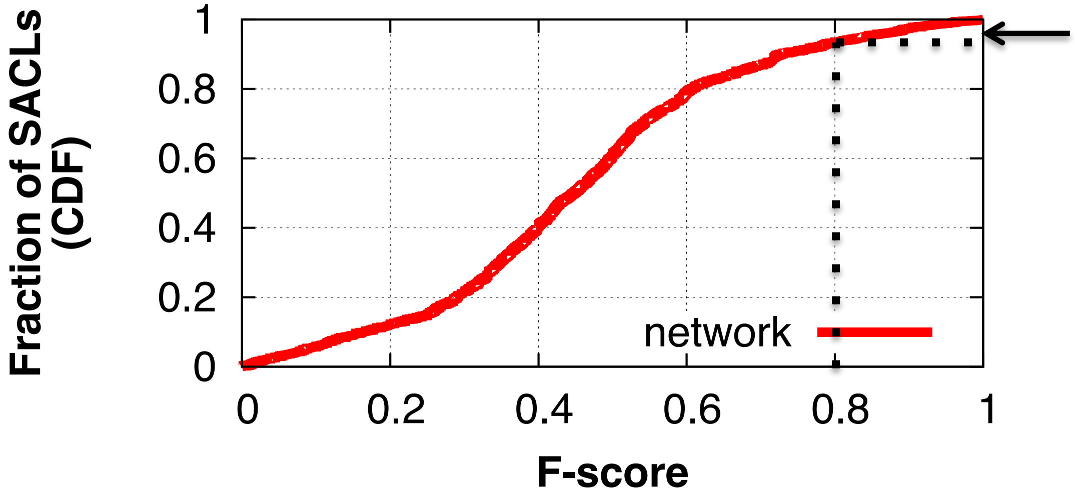
Do network groups correlate with SACLs?

- Ideal Case:
  - Groups by network community detection are highly similar to SACLs
- We measure F-score to check similarity



Do network groups correlate with SACLs?

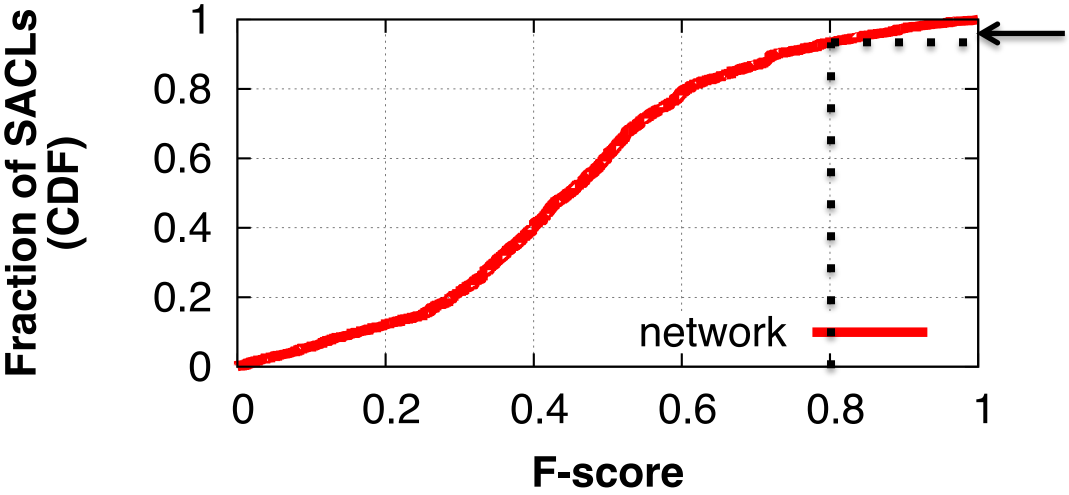
- Ideal Case:
  - Groups by network community detection are highly similar to SACLs
- We measure F-score to check similarity



Only 7% of SACLs Have F-score > 0.8

Do network groups correlate with SACLs?

- Ideal Case:
  - Groups by network community detection are highly similar to SACLs
- We measure F-score to check similarity



Only 7% of SACLs Have F-score > 0.8

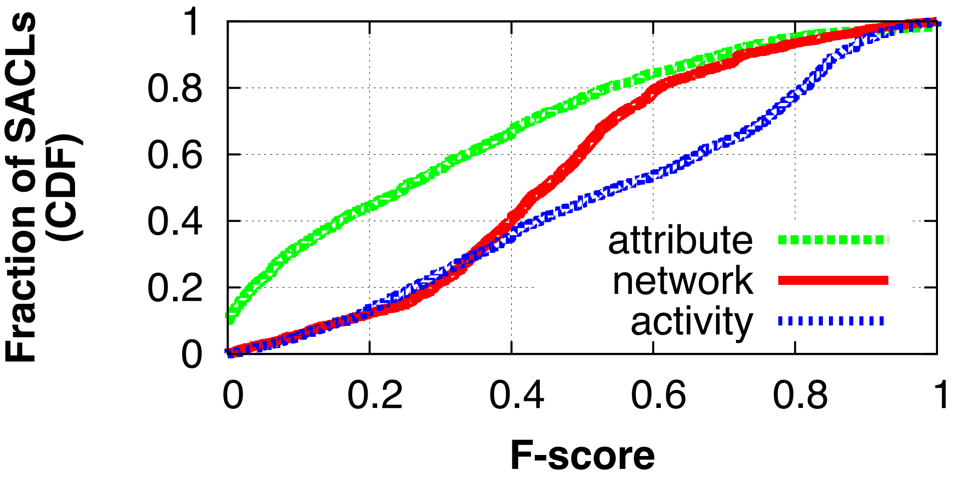
- Most SACLs do not correlate highly with network groups

Do other types of automatically detected groups correlate better with SACLs?

- In the existing work there are two more types of group detection:
  - Groups detected based on user profile attributes
  - Group detection based on user activity

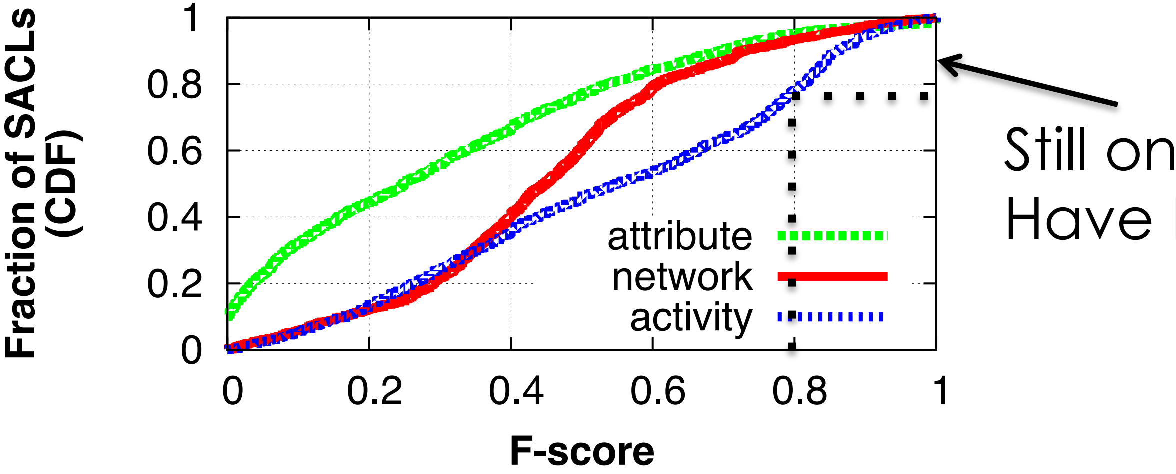
Do other types of automatically detected groups correlate better with SACLs?

- In the existing work there are two more types of group detection:
  - Groups detected based on user profile attributes
  - Group detection based on user activity



Do other types of automatically detected groups correlate better with SACLs?

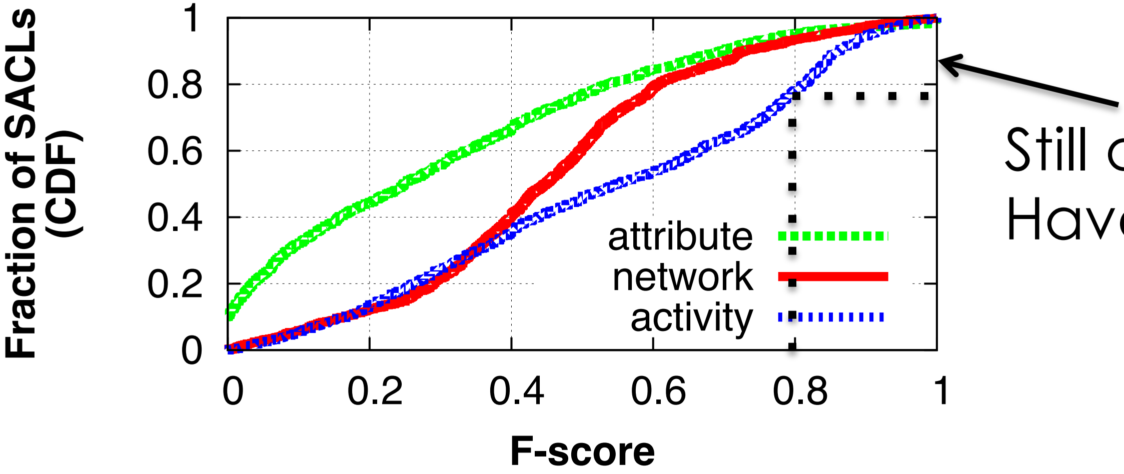
- In the existing work there are two more types of group detection:
  - Groups detected based on user profile attributes
  - Group detection based on user activity



Still only 20% of SACLs Have F-score > 0.8

Do other types of automatically detected groups correlate better with SACLs?

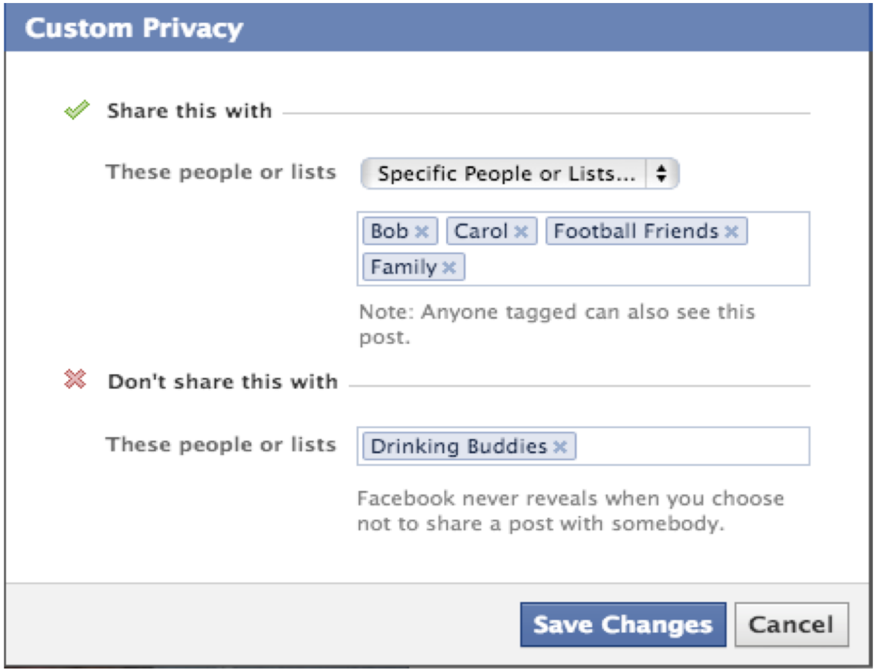
- In the existing work there are two more types of group detection:
  - Groups detected based on user profile attributes
  - Group detection based on user activity



Still only 20% of SACLs Have F-score > 0.8

- Most SACLs are NOT highly correlated with automatically detected groups

How to quantify user overhead of SACL specification?



✓ Share this with

These people or lists: Specific People or Lists...

Bob ✕ Carol ✕ Football Friends ✕ Family ✕

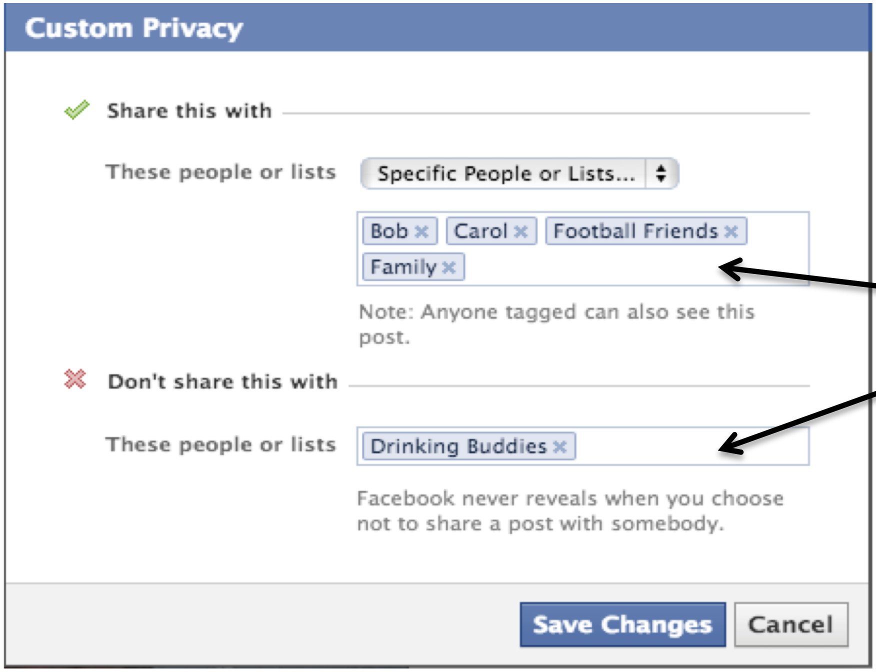
Note: Anyone tagged can also see this post.

✕ Don't share this with

These people or lists: Drinking Buddies ✕

Facebook never reveals when you choose not to share a post with somebody.

How to quantify user overhead of SACL specification?



Custom Privacy

✓ Share this with

These people or lists: Specific People or Lists...

Bob ✕ Carol ✕ Football Friends ✕ Family ✕

Note: Anyone tagged can also see this post.

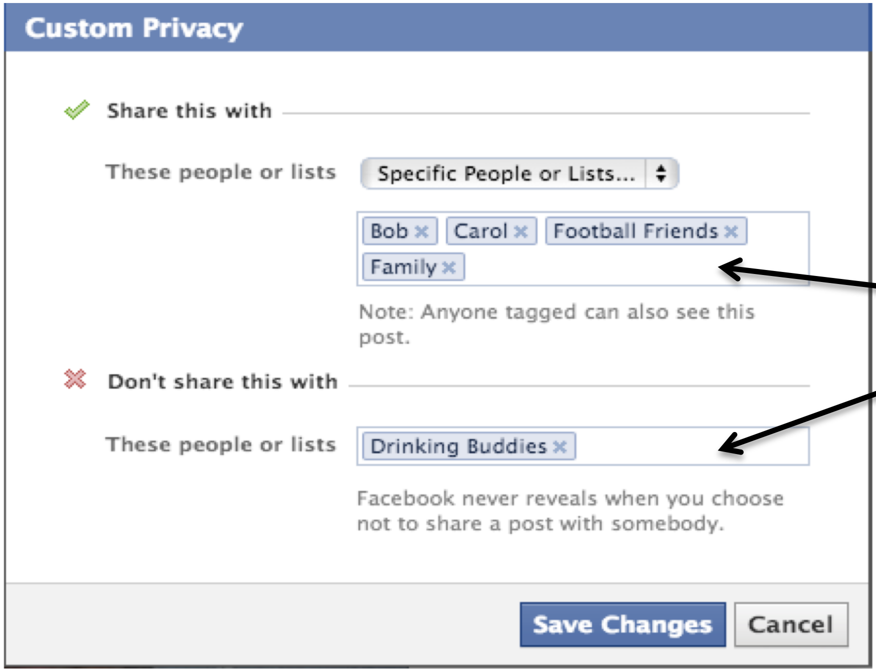
✕ Don't share this with

These people or lists: Drinking Buddies ✕

Facebook never reveals when you choose not to share a post with somebody.

Total **terms**: 5

How to quantify user overhead of SACL specification?



Custom Privacy

✓ Share this with

These people or lists: Specific People or Lists...

Bob, Carol, Football Friends, Family

Note: Anyone tagged can also see this post.

✗ Don't share this with

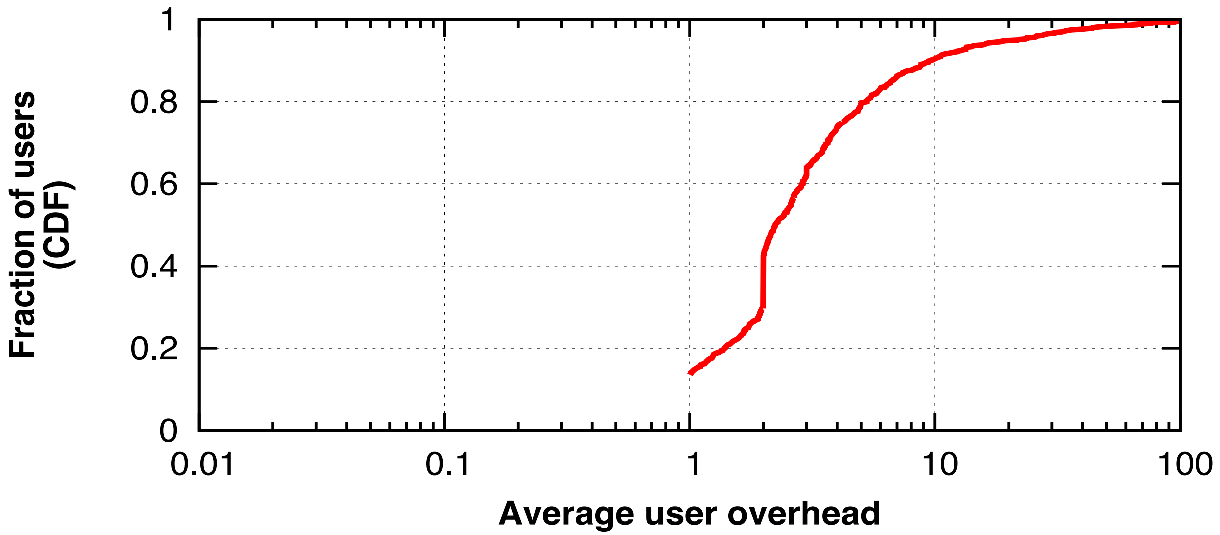
These people or lists: Drinking Buddies

Facebook never reveals when you choose not to share a post with somebody.

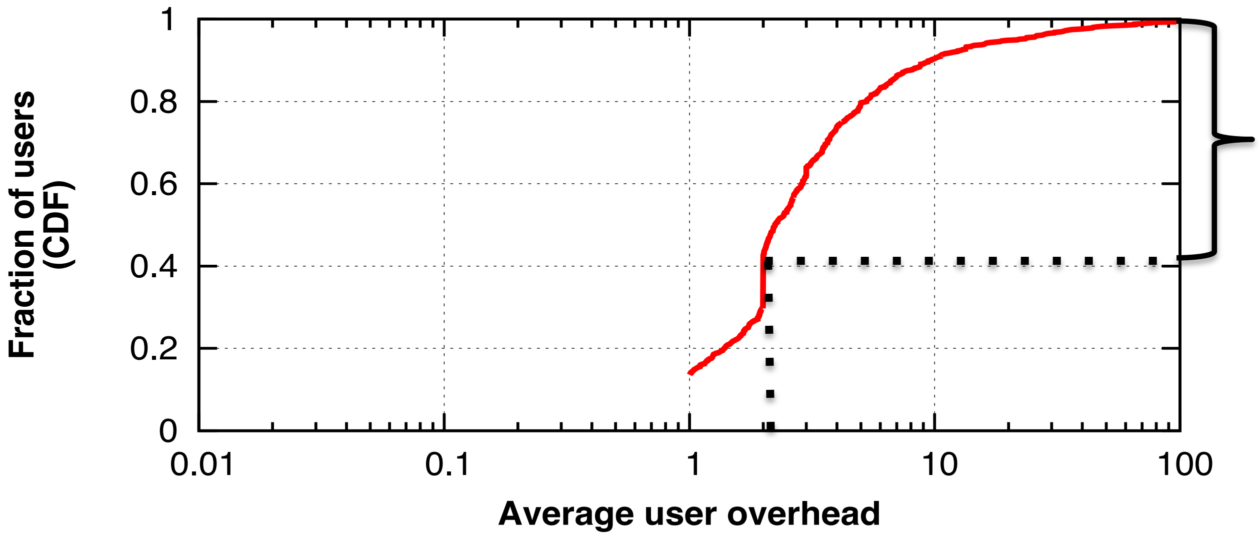
Total **terms**: 5

- Average user overhead = average #terms per content

What is the current overhead for users?

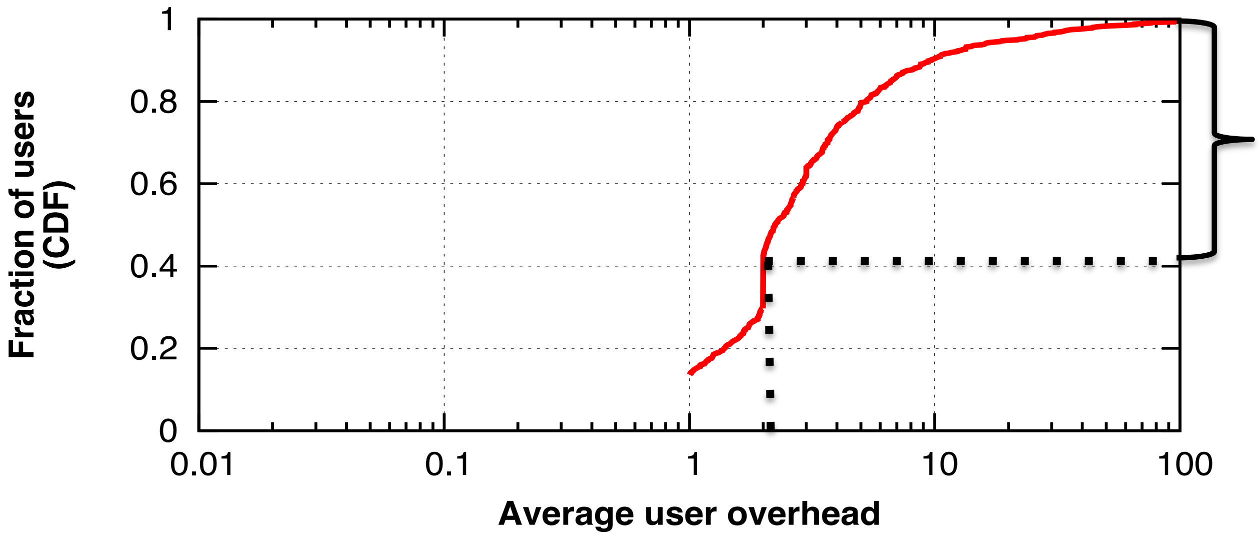


What is the current overhead for users?



58% of users
Have average
overhead > 2

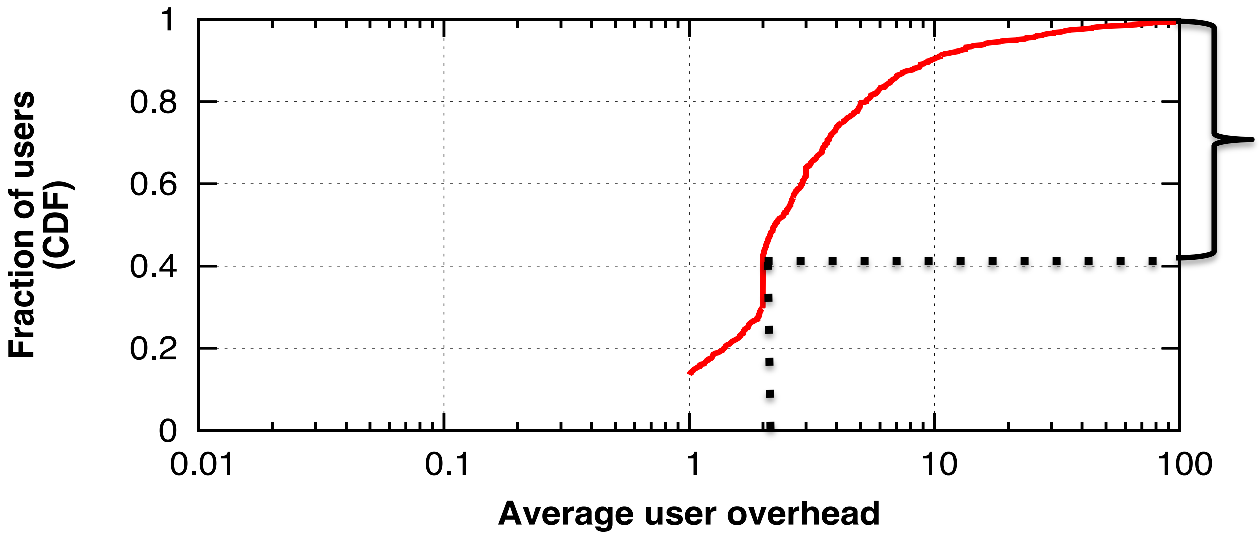
What is the current overhead for users?



58% of users
Have average
overhead > 2

- More than 150 users have overhead more than 5

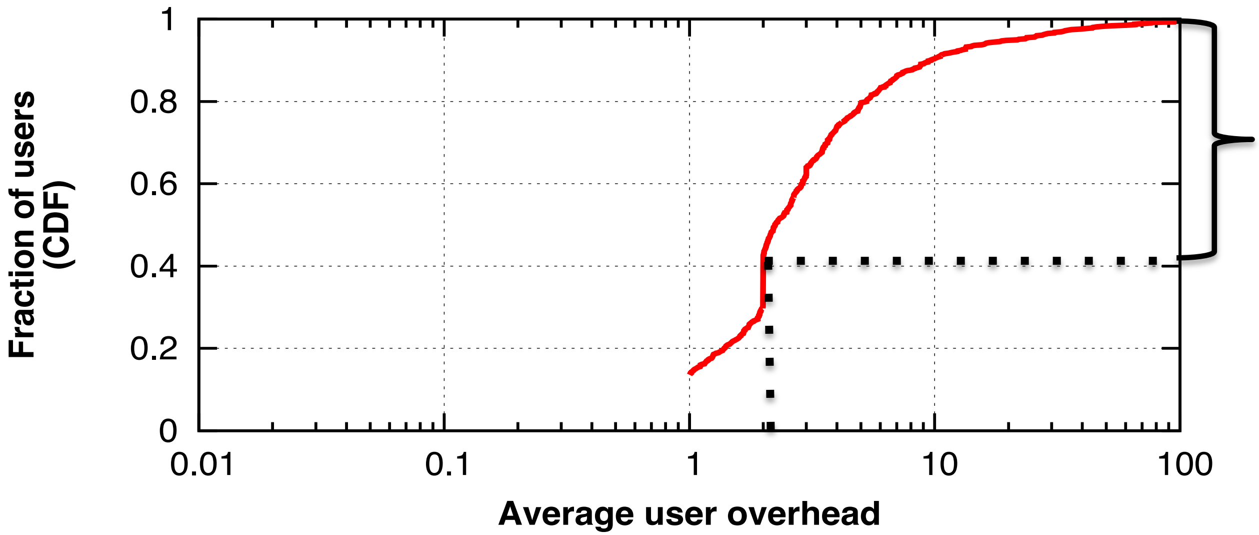
What is the current overhead for users?



58% of users Have average overhead > 2

- More than 150 users have overhead more than 5
- User overhead is high for specifying SACLs!

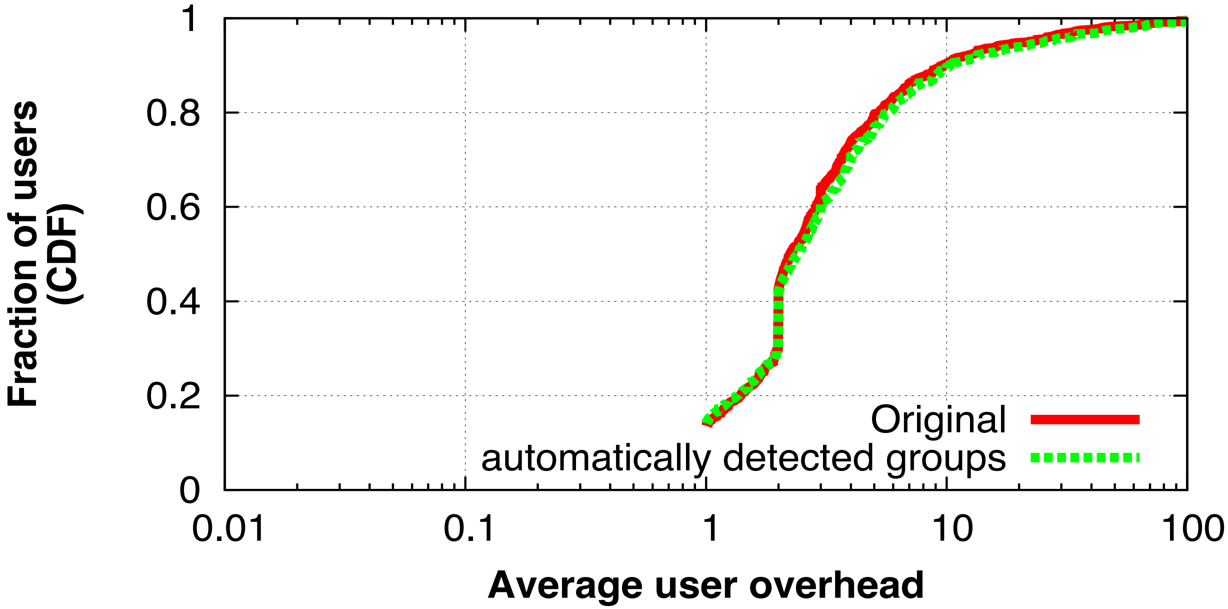
What is the current overhead for users?



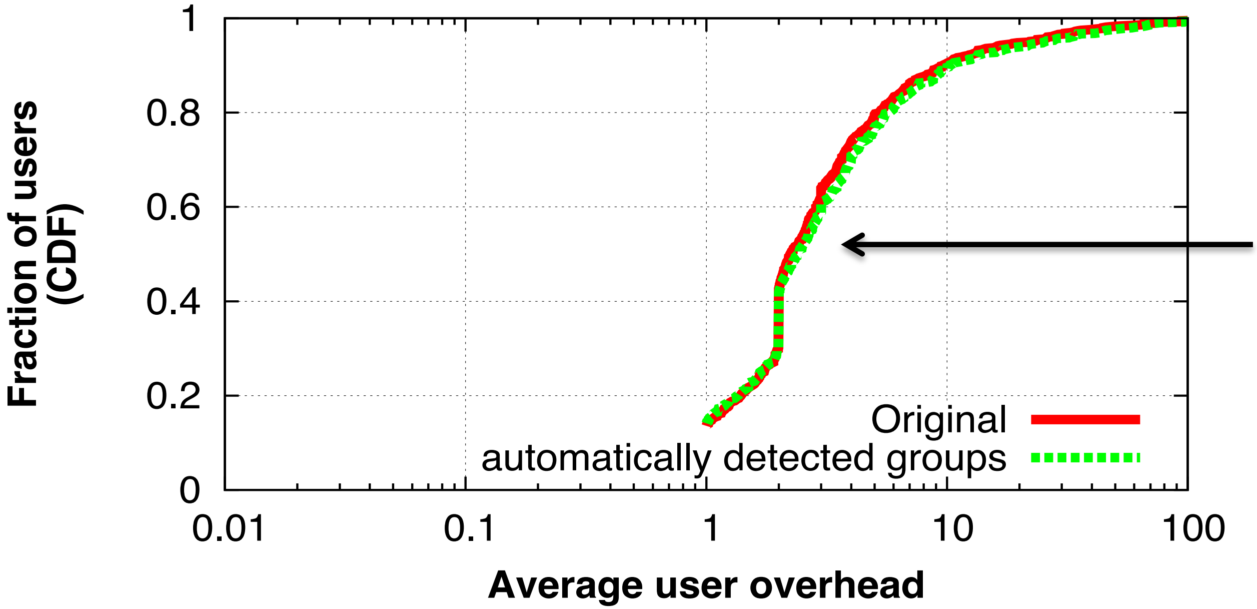
58% of users
Have average
overhead > 2

- More than 150 users have overhead more than 5
- User overhead is high for specifying SACLs!
- Can we use automated groups to reduce this overhead?

Can we reduce overhead using automatically detected groups?

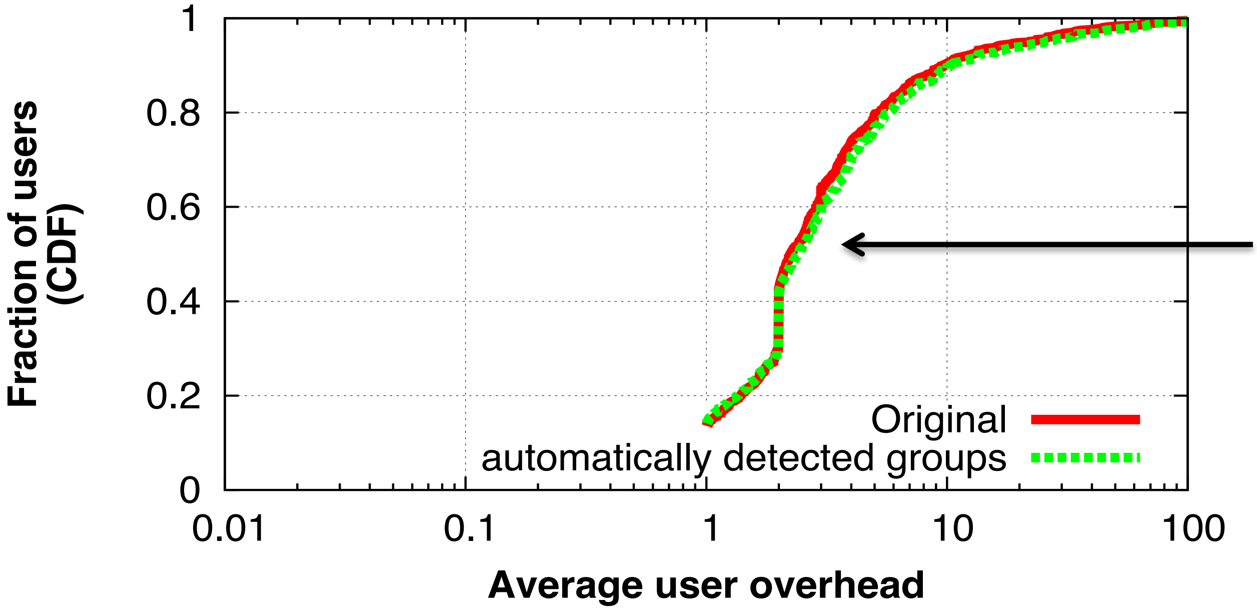


Can we reduce overhead using automatically detected groups?



No significant reduction

Can we reduce overhead using automatically detected groups?

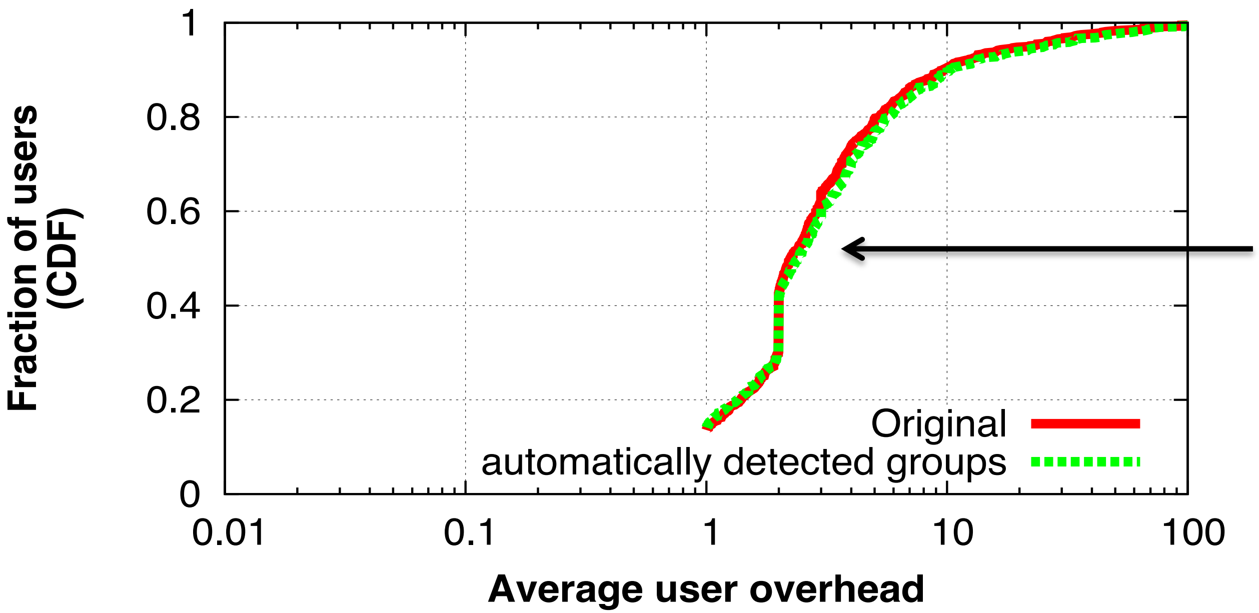


No significant reduction

Automated groups **do not** significantly reduce overhead

Consistent with our previous result

Can we reduce overhead using automatically detected groups?



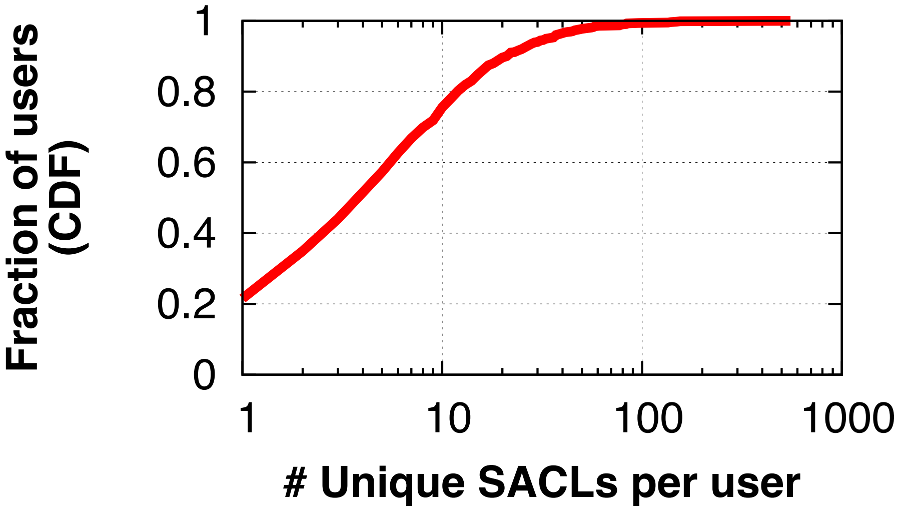
No significant reduction

Automated groups **do not** significantly reduce overhead

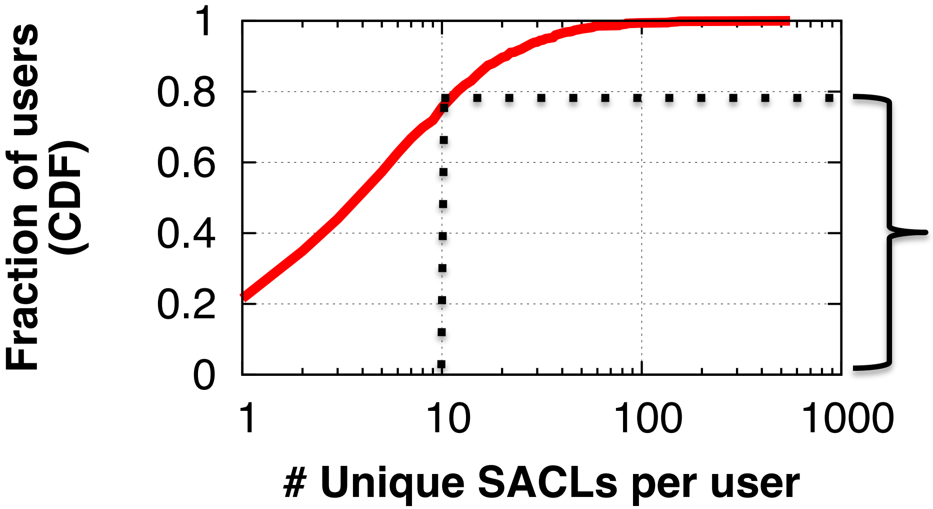
Consistent with our previous result

Can we reduce SACL specification overhead in some other way?

Insight: Users reuse SACLs repeatedly

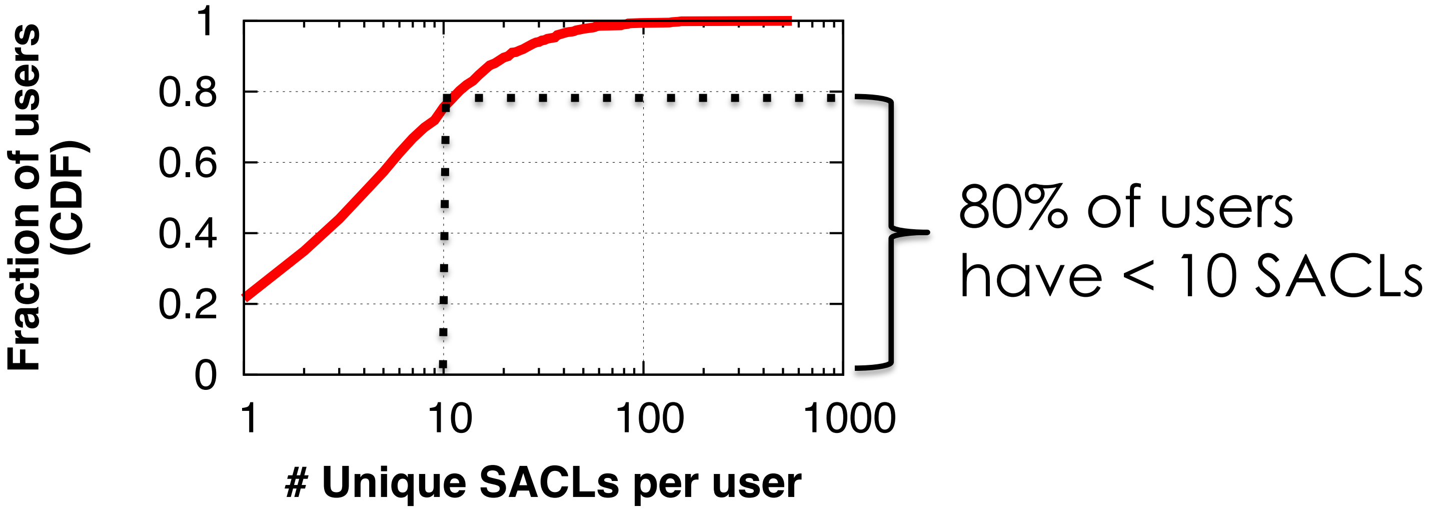


Insight: Users reuse SACLs repeatedly



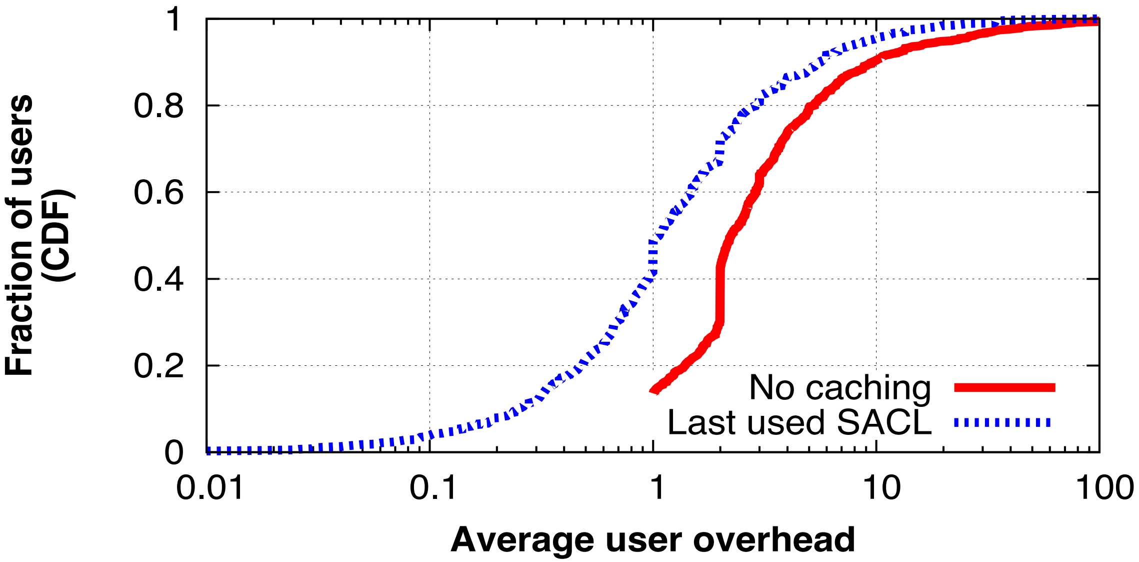
80% of users have < 10 SACLs

Insight: Users reuse SACLs repeatedly



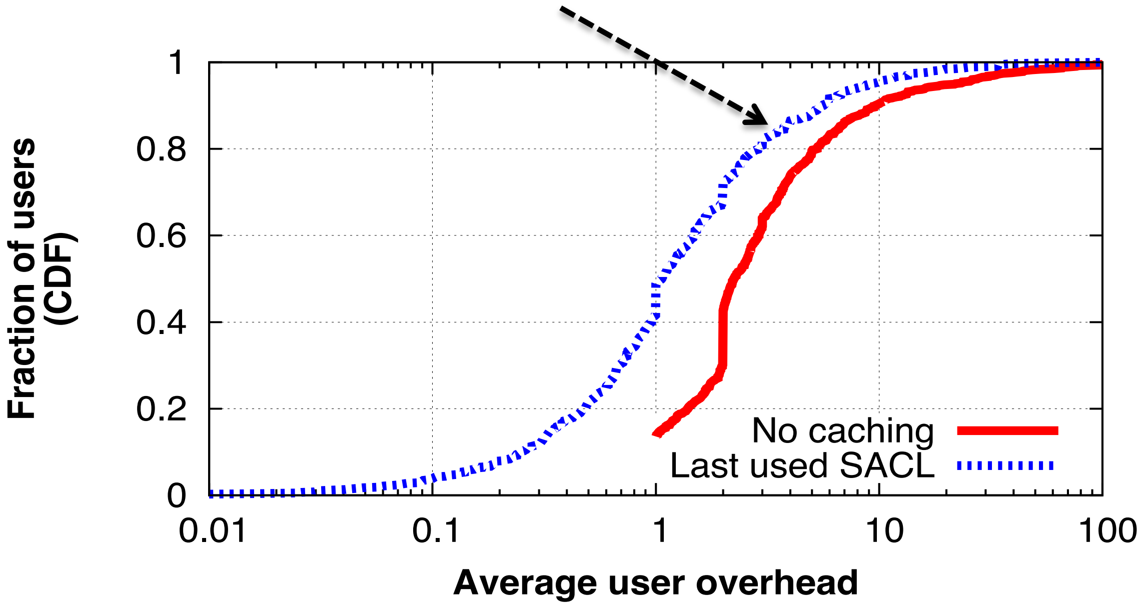
- Moreover on average a SACL is reused for 28 contents
- Most users reuse their SACLs repeatedly
- Idea: How about caching a few of the past SACLs to reduce overhead?

Does caching reduce SACL specification overhead?



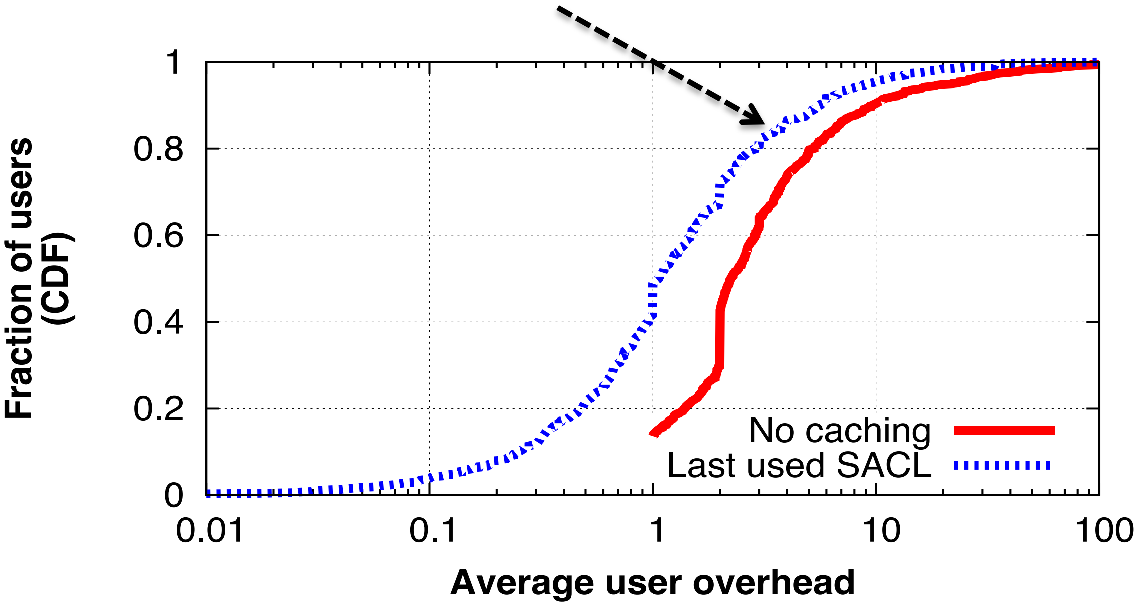
Does caching reduce SACL specification overhead?

Already used by **Facebook**



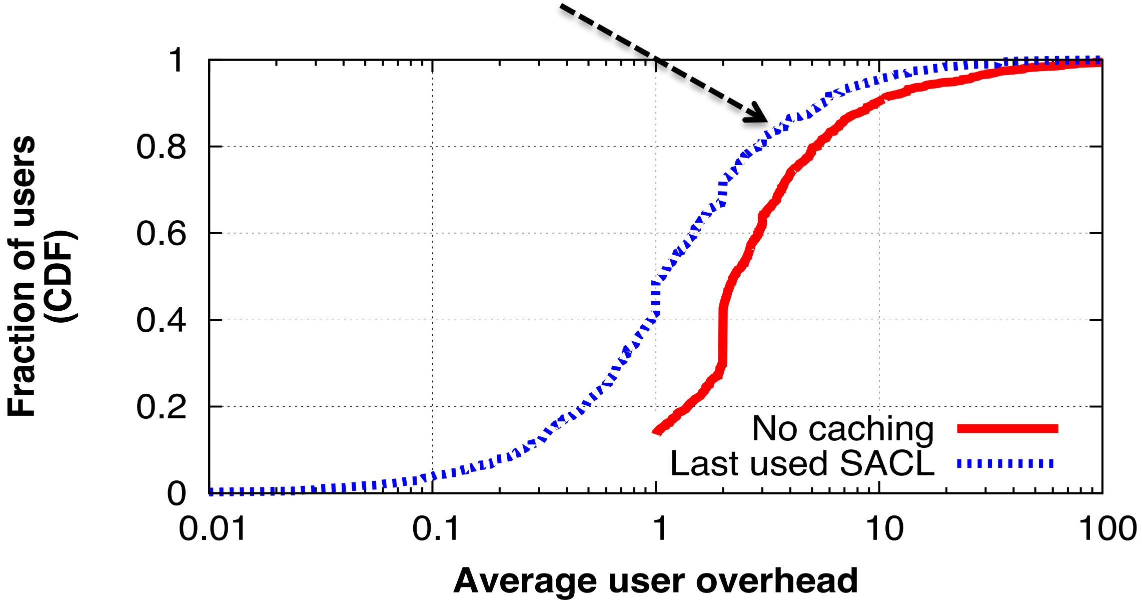
Does caching reduce SACL specification overhead?

Already used by **Facebook**
reduces overhead for **48% of users**



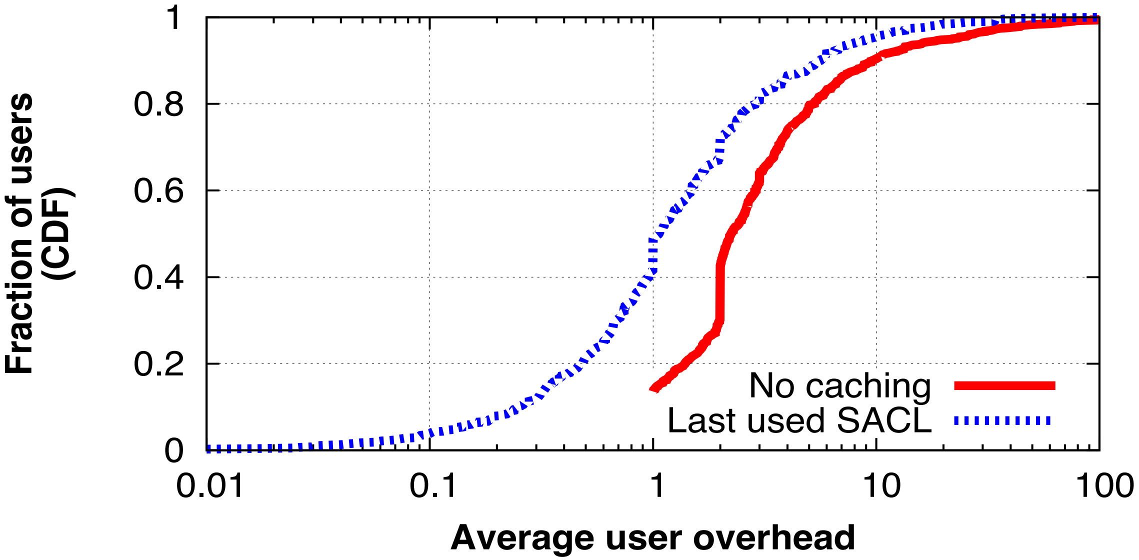
Does caching reduce SACL specification overhead?

Already used by **Facebook**
reduces overhead for **48% of users**



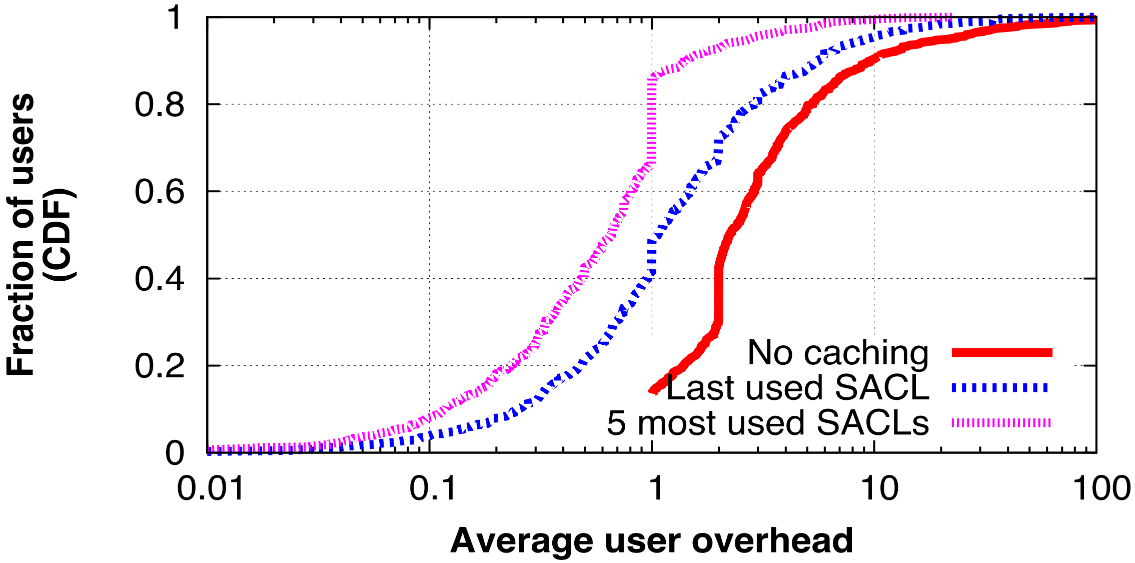
Can we improve by caching a few SACLs instead of one SACL?

Does caching reduce SACL specification overhead?



Can we improve by caching a few SACLs instead of one SACL?

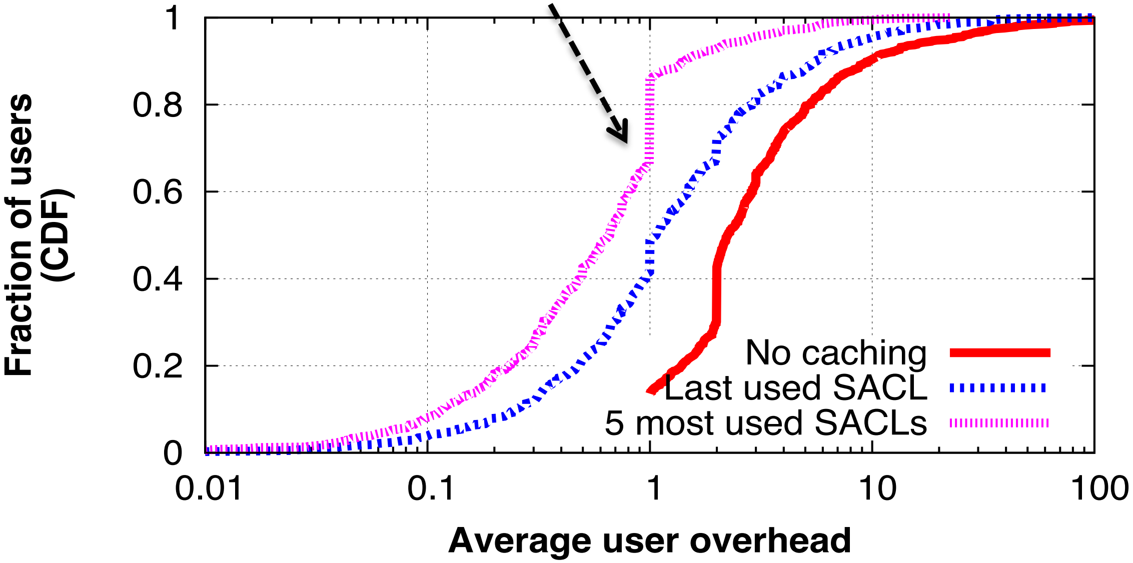
Does caching reduce SACL specification overhead?



Can we improve by caching a few SACLs instead of one SACL?

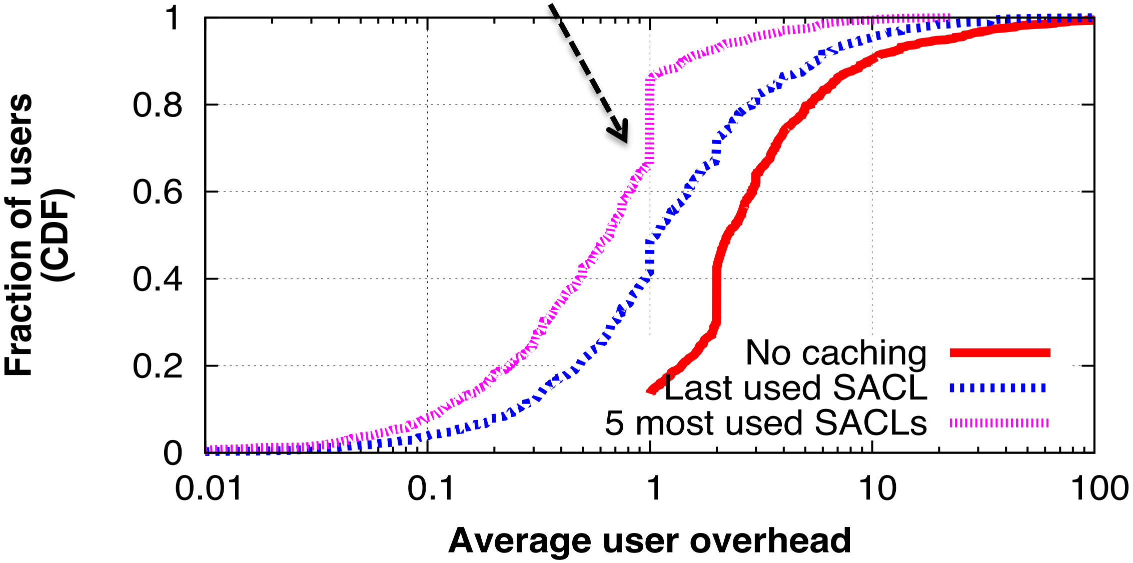
Does caching reduce SACL specification overhead?

Caching 5 most used SACLs reduces overhead for **86% of users**



Does caching reduce SACL specification overhead?

Caching 5 most used SACLs reduces overhead for **86% of users**



- SACL specification overhead significantly reduces if we cache just a few SACLs!

Summary of SACL study

- First study to collect and analyze data about real-world SACL usage
- Usage of SACLs are surprisingly common
- SACLs show little correlation with automatically detected groups
- Caching past few SACLs is a very promising direction to greatly reduce the user overhead of SACL specification

Privacy in public (social media)

Understanding	solution
1. Information Revelation and Privacy in Online Social Networks, Acquisti and Gross, WPES’05 2. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook, Acquisti and Gross, PETS’06 3. Analyzing Facebook Privacy Settings: User Expectations vs. Reality, Liu et al. , IMC’2011 4. Quantifying the Invisible Audience in Social Networks, Bernstein et. al., CHI’2013	1. Privacy Wizards for Social Networking Sites, Fang et. al., WWW'2010
1. Understanding and Specifying Social Access Control Lists, Mondal et. al. SOUPS’14	

General methodology for each of these papers

- Step 1: Collect behavioral data by passively observing users (via collecting “public” data, sometimes consent)
- Step 2: [Optional] use the collected data to ground surveys and capture user expectations or desires about privacy (SACL study did not use it)
- Step 3: Analyze the data and identify patterns to answer research questions

Large-scale internet measurement : Pros / Cons

- Pros
  - Just requires coding to collect data
  - Don’t need to always take consent, Better if you do
  - Sometimes you cannot (example?)
  - Gives you statistically valid facts due to large volume
- Cons
  - What data to collect? How? What infrastructure (often requires non-traditional approach)
  - You will not have user feedback: Whatever pattern you saw if that what REALLY users are doing?
- Hybrid studies are often better
  - You do your measurement
  - Also get some feedback from users on why and how