Max
Planck
Institute
for
Software Systems

Beyond Access Control: Managing Online Privacy via Exposure

Mainack Mondal[†] Peter Druschel[†] Krishna Gummadi[†] Alan Mislove[‡]

[†]MPI-SWS [‡]Northeastern University

USEC, February 2014

Privacy concerns in Online Social networking sites (OSNs)

" Privacy is the ability for people to determine for themselves when, how, and to what extent, information about them is communicated to others" - A. Westin. Privacy and Freedom, 1970

~1 B users

~4.75 B daily pieces of content

How to ensure privacy of this content ?

Source: http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats

Privacy concerns with access of OSN content

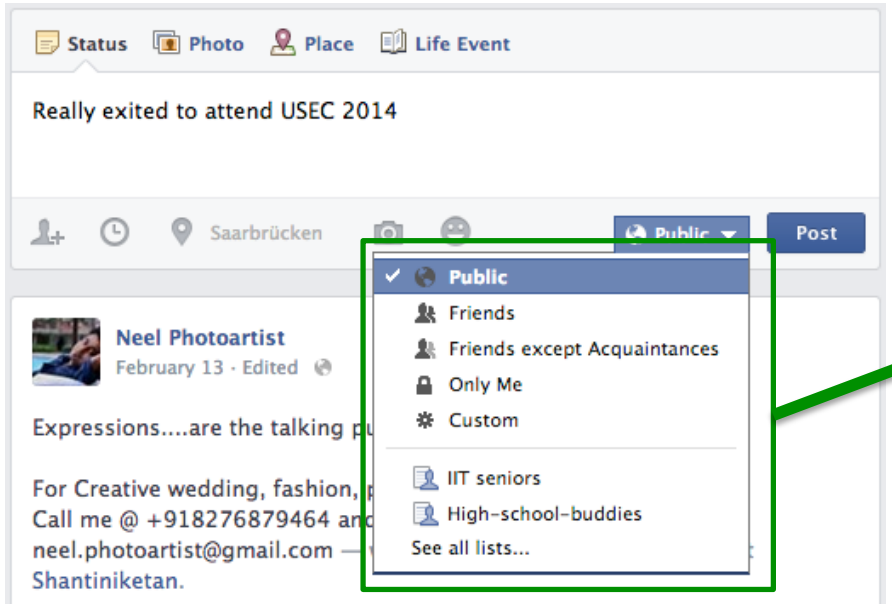
~~1. Ensure privacy from OSN operators~~

[Guha et al.] [Baden et al.]

[Shakimov et al.]

2. Our concern: Ensure privacy from other users

Managing privacy with Access Control Lists (ACLs)

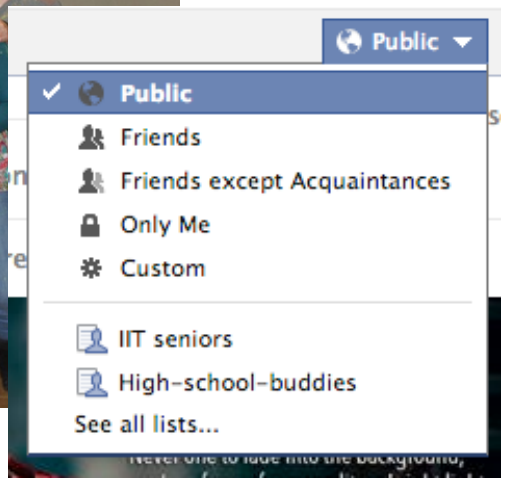


Allow others access to content

Privacy violation from ACL point of view:

If someone accesses content who the user did not allow

Privacy violations in the real world



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

Scenario 1: Facebook newsfeed

Facebook pushes your content as updates

Others automatically get your content
when they login to their Facebook page

After Newsfeed: More people actually saw the content

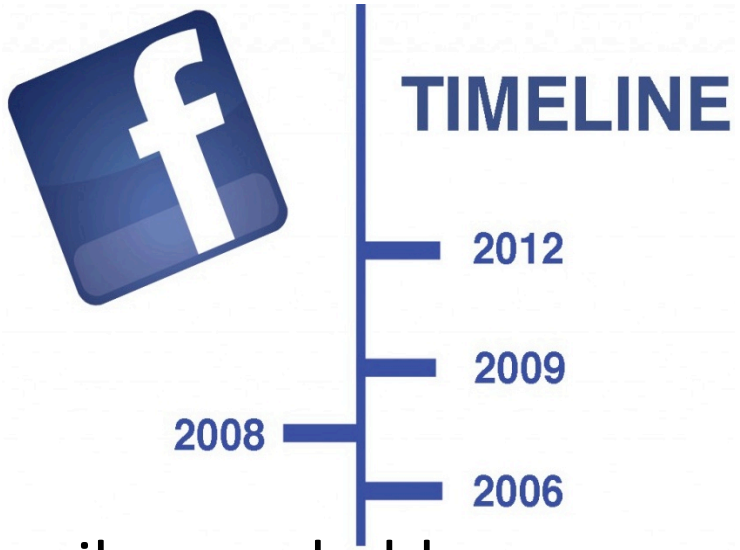
Users complained of privacy violation [Boyd et al. '08]

Before and after Newsfeed: **access control did not change**!

Scenario 2: Facebook timeline

Sort your content by upload time

Others can search by time



After timeline: Old content became easily searchable

Users felt privacy was violated

Before and after Timeline: **access control did not change**!

Scenario 3: Spokeo

Service aggregating public data from web

Others get all of this data by searching Spokeo

After aggregation: Inferring non public data become easier

Users complained of privacy violation

Before and after aggregation: **access control did not change!**

Summary

User reaction suggests each of the cases violated privacy

However access control was not violated in any of the cases

Take away 1: Access control is inadequate to capture user intention

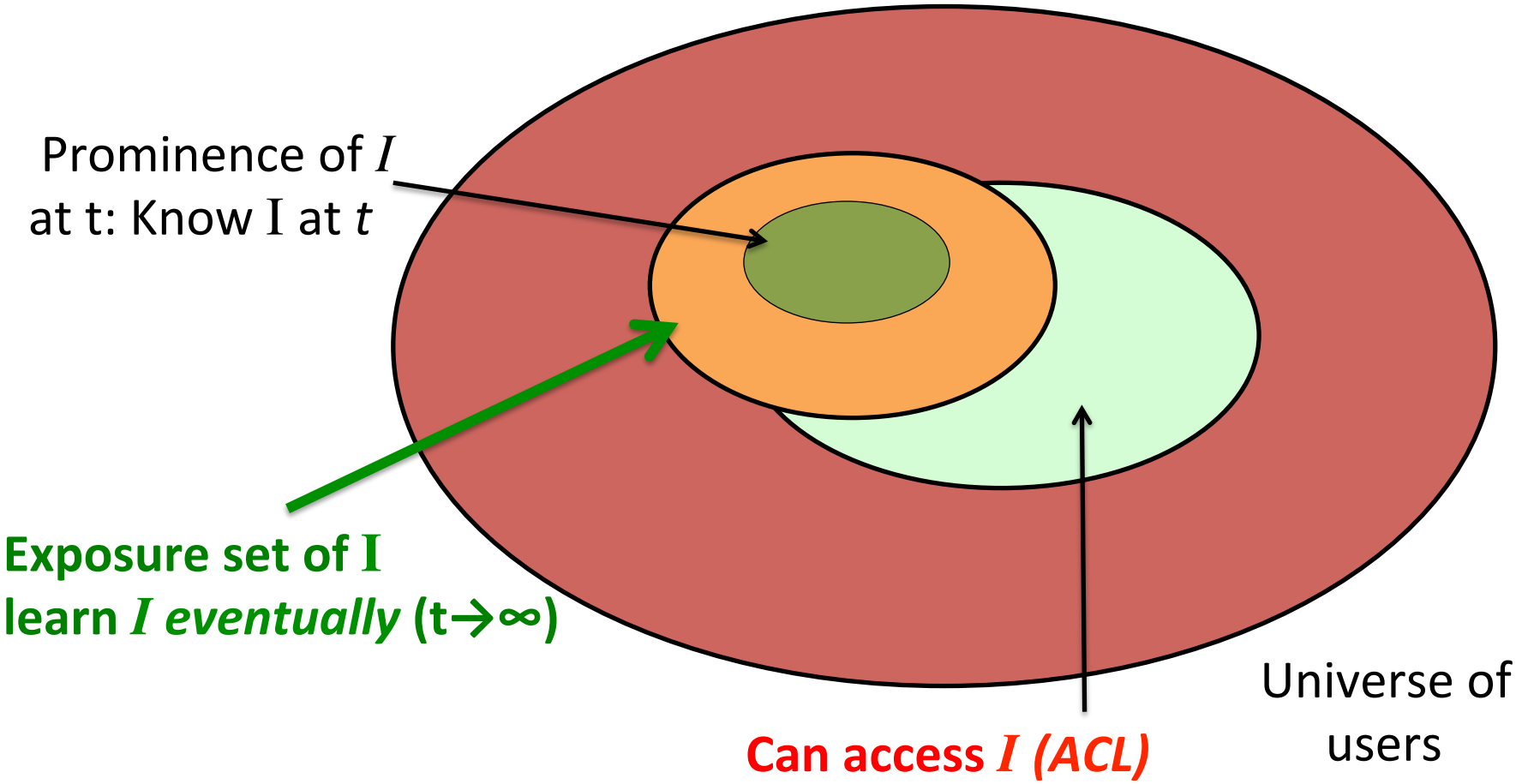
Outline

Access control is inadequate to capture privacy

Exposure: A different concept to capture information privacy

Discussion: How to manage privacy via exposure

Exposure : Definition



Exposure for content I

The set of people who will learn I eventually

How accurately do users estimate exposure?

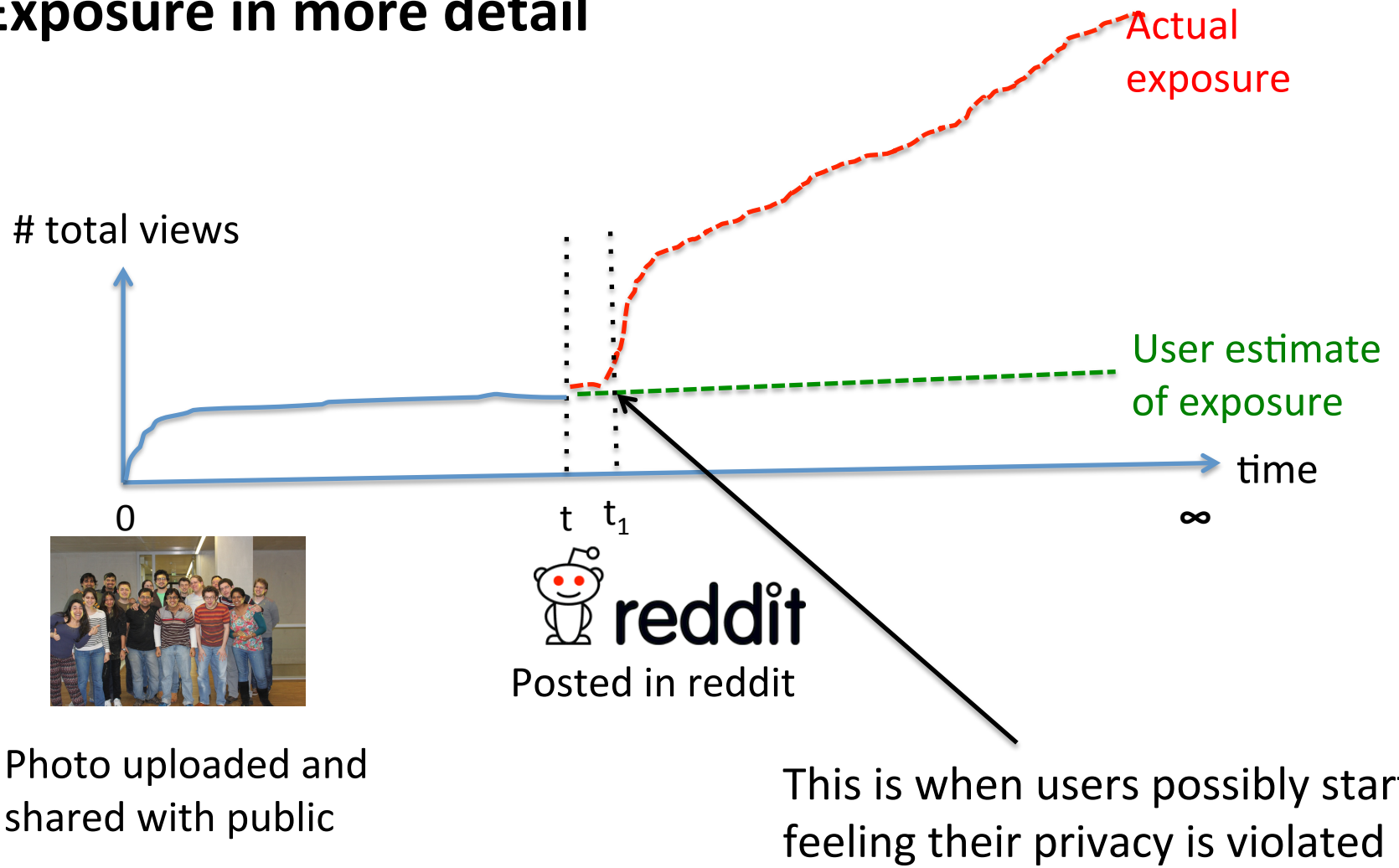
Facebook researchers did a study with 589 users

[Bernstein et al. 2013]

Perceived exposure grossly underestimates actual exposure

There may be a feeling of privacy violation when actual exposure is different from perceived exposure

Exposure in more detail



Revisiting scenario 1: Facebook newsfeed

Exposure before newsfeed

Friends who visit profile

Exposure after newsfeed

All the friends who are logged into Facebook

Exposure of uploaded information **after newsfeed** > **Exposure** of uploaded information **before newsfeed**

Revisiting scenario 2: Facebook timeline

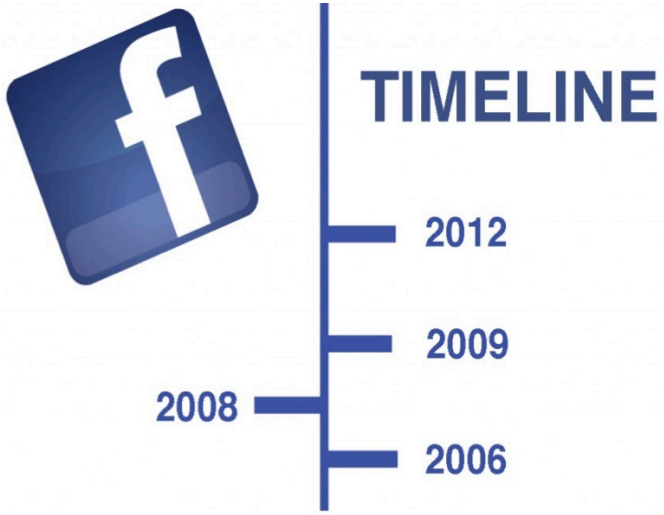
Exposure of old content before timeline

Users who will scroll down

thousands of content

Exposure of old content after timeline

All users who search by time



Exposure of old information **after timeline** > **Exposure** of old information **before timeline**

Revisiting scenario 3: Spokeo

Exposure before aggregation

Users who collect content themselves from multiple sources

Exposure after aggregation

Any user who searches in Spokeo

Exposure of inferred information **after aggregation** > **Exposure** of inferred information **before aggregation**

Take away 2: Exposure based privacy model can capture violations which are not captured by access control

Outline

Access control is inadequate to capture privacy

Exposure: A different concept to capture information privacy

Discussion: How to manage privacy via exposure

Discussion: Managing privacy via exposure

Challenge 1:

How to estimate exposure for a content?

Challenge 2:

How to make users aware of the estimated exposure?

Challenge 3:

How to allow users more control over exposure?

Challenge 1: Estimating exposure

Situations where predicting exposure is very hard

Cross site prediction, exposure of inferred information

Situations where predicting exposure is possible

Predicting exposure of content in a site

Lots of research in content popularity growth

[Borghol et al] [Figueiredo et al.]
[Hong et al.] [Zaman et al]
[Bernstein et al.]

Challenge 1: Who can best estimate exposure

OSN operators are in the **best position to predict** exposure accurately with the data they collect

- They log who is accessing what content
- They collect historical data for content access

OSN operators can also **control** exposure

- They decide which content to show other users

Challenge 2: How to make users aware of the exposure?

Prediction can be shown to users at different granularity

- ✔ **List** of predicted people for a content
- ✔ **Number** of predicted people for a content
- ✔ Showing the prediction for a certain **time period**
- ✔ Showing the prediction with **error bounds**
- ✔ Showing how a **specific dissemination mechanism** changes the prediction

  e.g., 200 more people are likely to see your content due to newsfeed

Challenge 3: How to allow users more control over exposure?

Different "knobs" can be provided to the user

- ✔ Change access control to a more restrictive setting
- ✔ Disabling particular dissemination mechanisms, e.g. search
- ✔ Enabling tripwires
  - Take content offline if more than 50 people view
  - Take content offline after two months

Take away 3: There are lots of open challenges and substantial research opportunities in how to design and deploy exposure based systems

Conclusion

Take away 1: Access control is inadequate to capture user intention

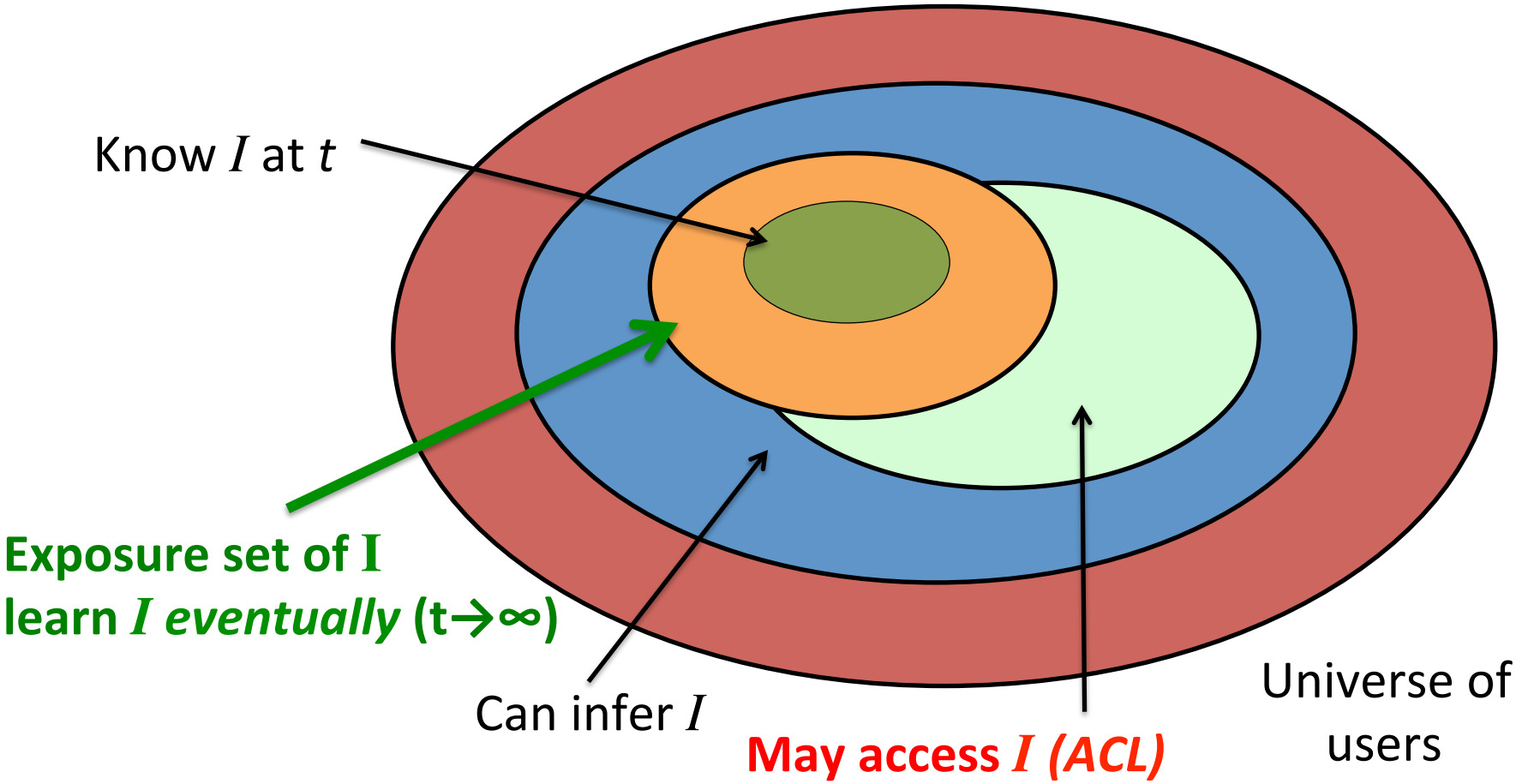
Take away 2: Exposure based privacy model can capture violations which are not captured by access control

Take away 3: Lots of open challenges to design systems which can manage privacy by controlling exposure

Thank you!

Backup slides

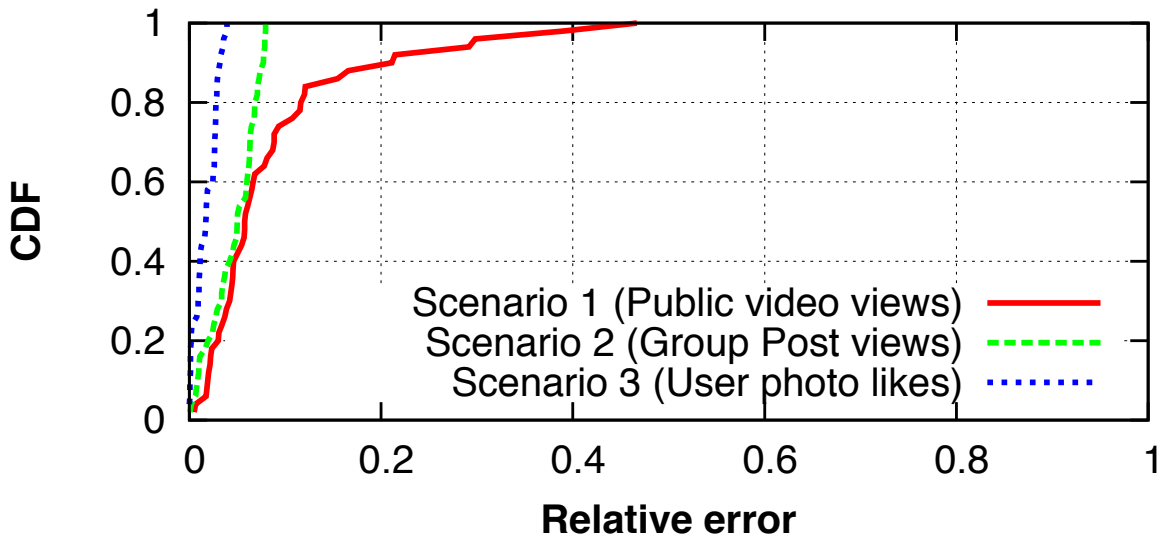
Exposure : Definition



Exposure for content I

The set of people who will learn I eventually

- How accurately can you predict future exposure?



Relative error is less than 0.1 in 75% of Scenario 3!

- Can predict exposure with high accuracy

Extra slides

Access control is inadequate, scenario 1: Facebook newsfeed

- Facebook introduced News feed in 2006
  - Involved pushing new information to friends' Facebook page
- Information became almost involuntarily accessible
- Users strongly objected stating violation of privacy

Access control was not changed !

Access control is inadequate, scenario 2: Facebook timeline

- Facebook introduced timeline in 2011 end
  - Chronologically order all the information on your profile
  - Make them easily searchable for other users
- Easier to search Potentially embarrassing older content
- Users were afraid of privacy violation

Access control was not changed !

Access control is inadequate, scenario 3: Spokeo

- Service aggregating information about individuals
  - Each individual information is public content
  - E.g., your Facebook profile, address
- One can infer new non public information
  - Estimating wealth using address and public property records
- Users complain of privacy violation

Access control was not changed !

Modeling user privacy using exposure

- For each content users have an expected exposure
  - How many other users are likely to access the content
- We can model privacy violation for an information as
  - Large deviation of actual exposure from expected exposure

Revisiting scenario 1: Facebook newsfeed

- Before newsfeed was introduced
  - Expected exposure: Friends who will visit user's profile
  - Actual exposure was same as expected exposure
- After newsfeed was introduced
  - Actual exposure: All friends to whom the information is pushed
  - Actual exposure is much higher than the expected exposure

Revisiting scenario 2: Facebook timeline

- Before timeline was introduced
  - Expected exposure for older data: Friends who will scroll to find a old content
  - Actual exposure for older data was same as expected exposure
- After timeline was introduced
  - Actual exposure for older data: All friends who visit the profile
  - Actual exposure is much higher than the expected exposure

Revisiting scenario 3: Spokeo

- Before spokeo aggregated data
  - Expected exposure for new inferred data: Users who dig up each individual pieces of content form different sources
  - Actual exposure for older data was same as expected exposure
- After spokeo aggregated data
  - Actual exposure for new inferred data: All users who visit public spokeo website
  - Actual exposure is much higher than the expected exposure

Key challenge: Predicting future exposure

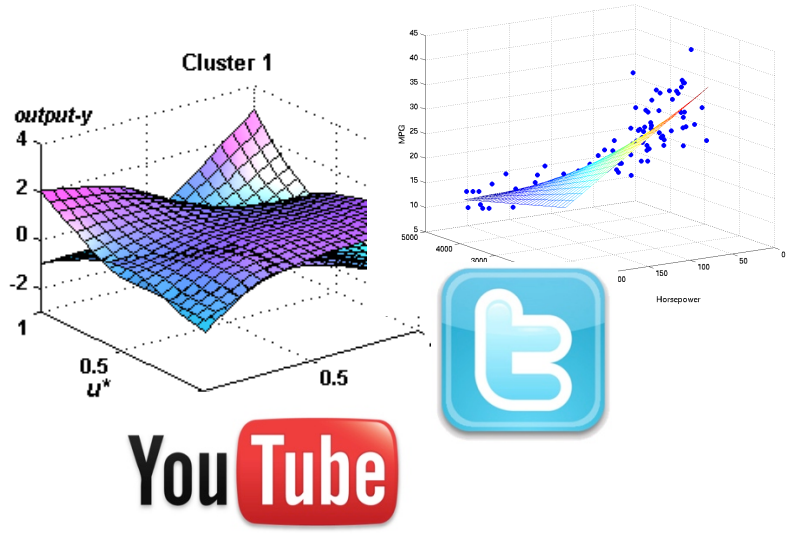
- Huge existing work for predicting growth in content popularity
  - Future YouTube views, Facebook likes, Retweets
  - Use machine learning, regression techniques
  - We can leverage advances in those fields to predict exposure

- OSN operators are best positioned to do the predictions
  - Empirical data on how information disseminates in their sites
  - Facebook or Youtube already provide number of likes or views

Change in exposure can capture the privacy violations not covered by access control

Key challenge: Predicting future exposure

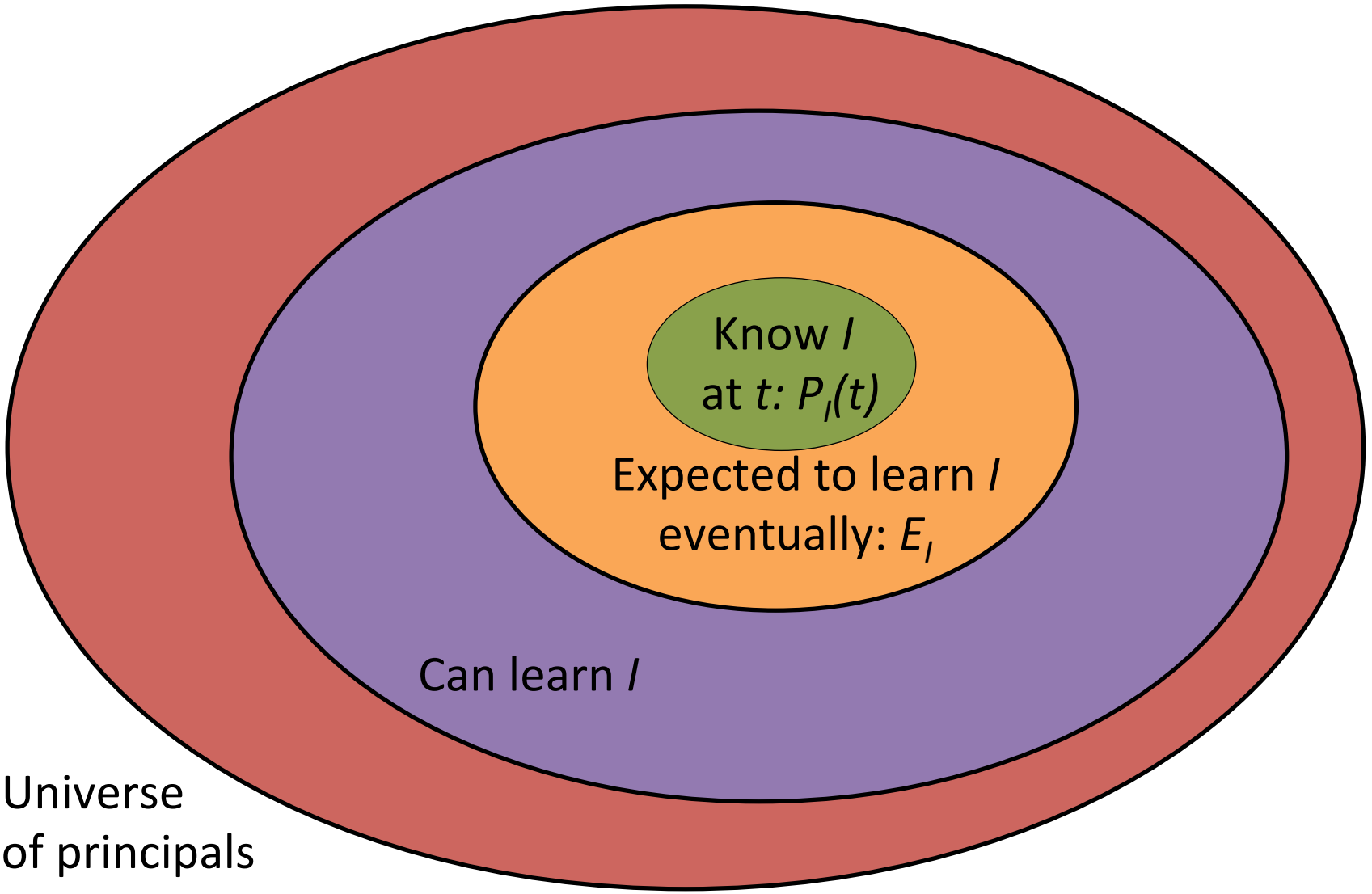
- Leverage advances in predicting popularity growth and information propagation



- Easiest to predict for OSNs by virtue of huge empirical data

Limitations of our model

- Privacy violation by inference using available data
  - It is extremely hard to enumerate all possible inference
- Privacy violation using cross site prediction
  - Prediction across multiple systems
  - E.g., posting a picture taken from Facebook in tweeter


Know I at t: $P_I(t)$
Expected to learn I eventually: E_I
Can learn I
Universe of principals

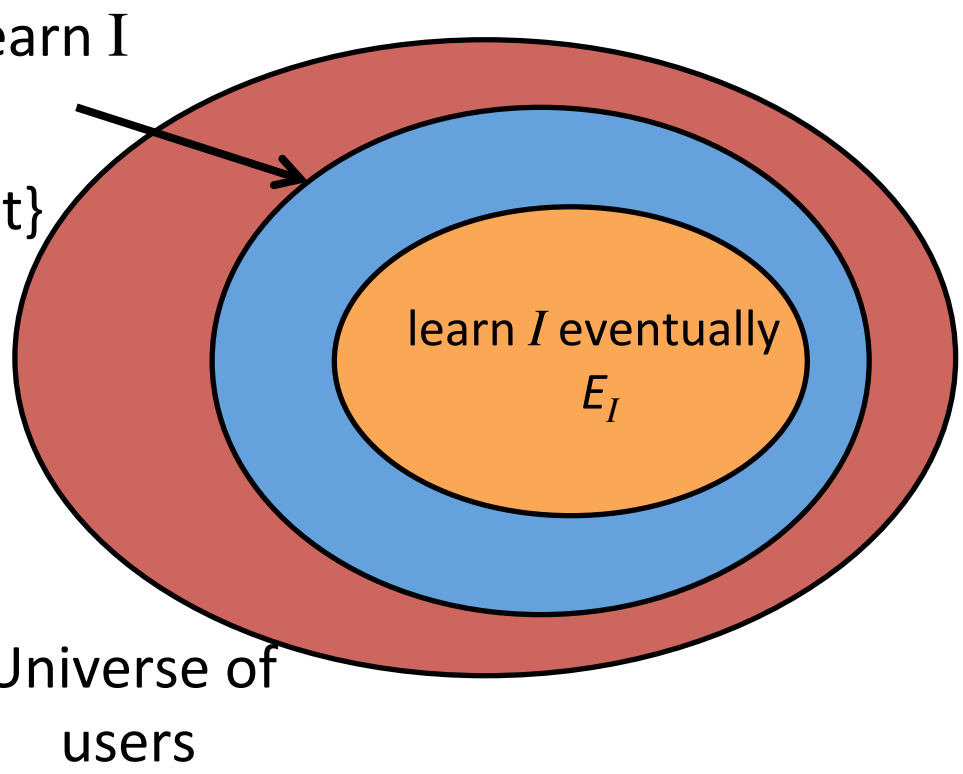
Exposure : Definition

Exposure for content I

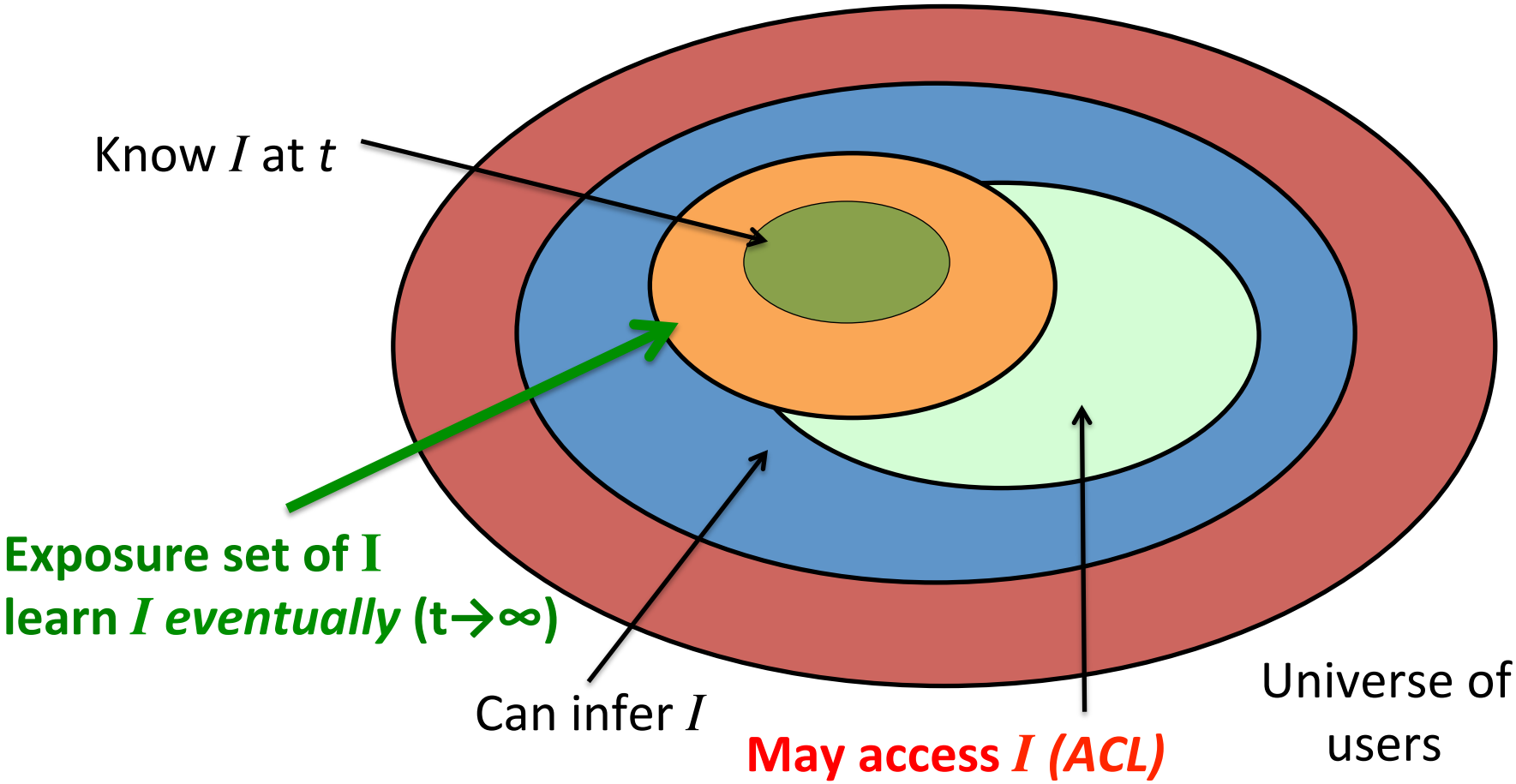
The set of people who will learn I eventually

$P_I(t)=\{U \mid U \text{ learn } I \text{ at time } t\}$

$E_I = \lim_{t\to\infty} P_I(t)$



Exposure : Definition

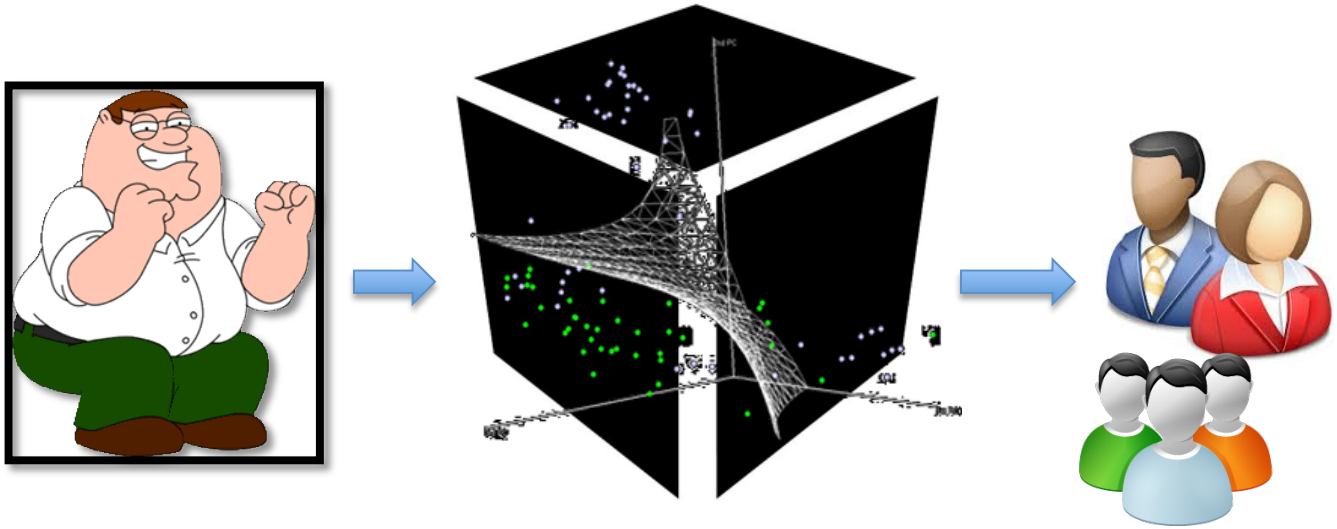


Exposure for content I

The set of people who will learn I eventually

Proposed model: managing privacy via exposure

Predicting future exposure for content



Proposed model: managing privacy via exposure

Predicting future exposure for content

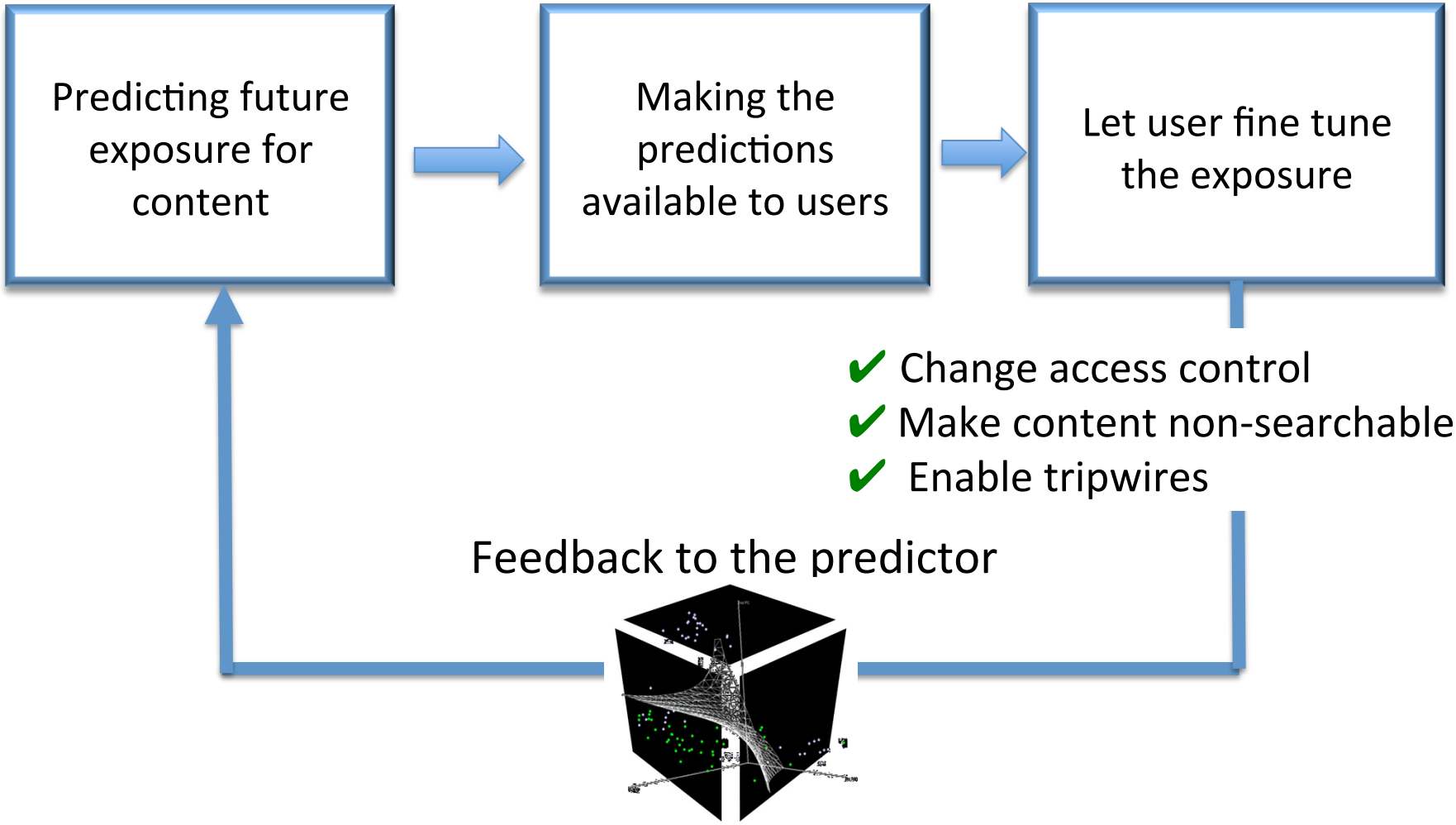
Making the predictions available to users

Like · Share · January 26 at 2:17am near Exton, PA, United States

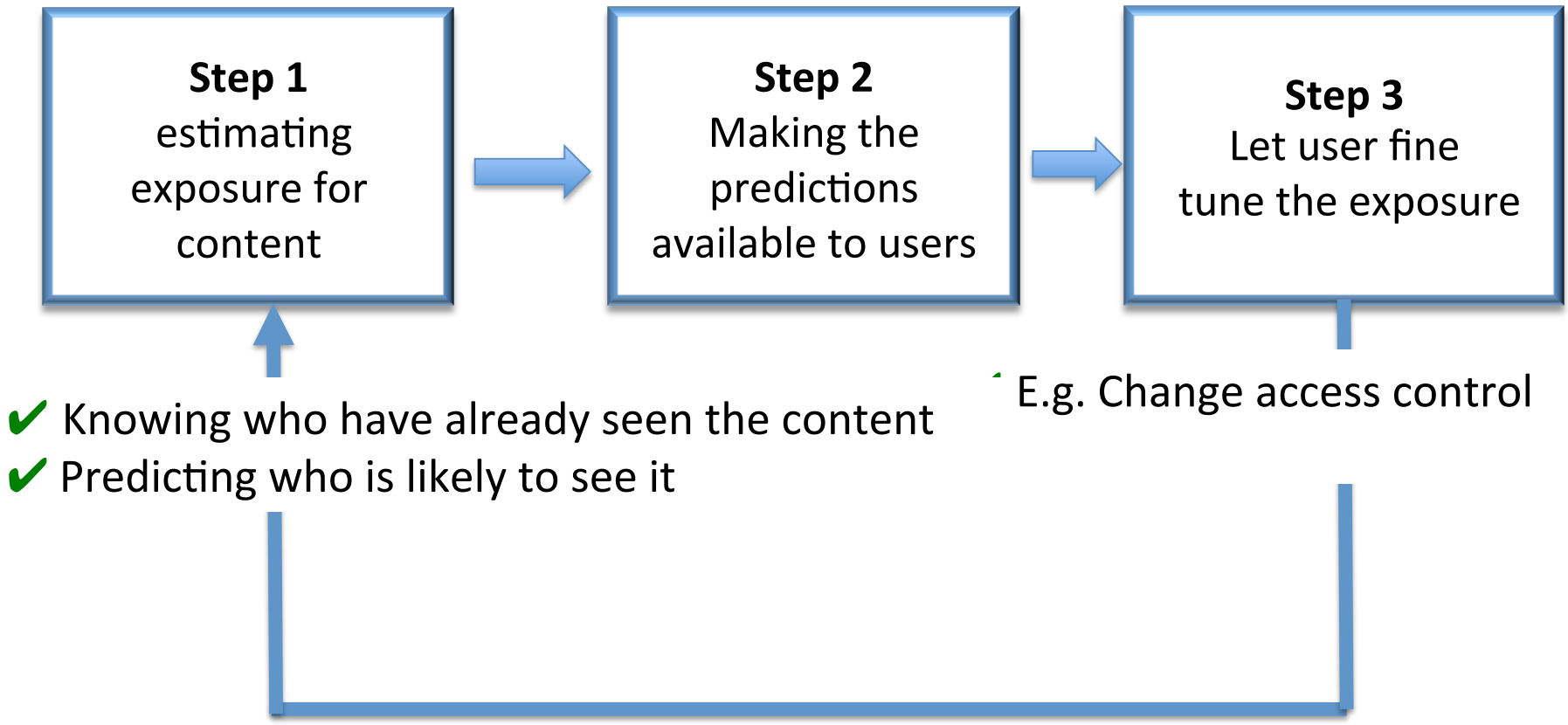
Will possibly be seen by around 20 friends (click to see who they are)

Will possibly be seen by around 20 people (click to see who they are)

Proposed model: managing privacy via exposure



Managing privacy via exposure



Step 1: Estimating future exposure

Key challenge: Predicting future exposure

Situations where predicting future exposure is very hard

Cross site prediction, e.g., exposure after re-sharing

exposure of inferred information: inferring wealth

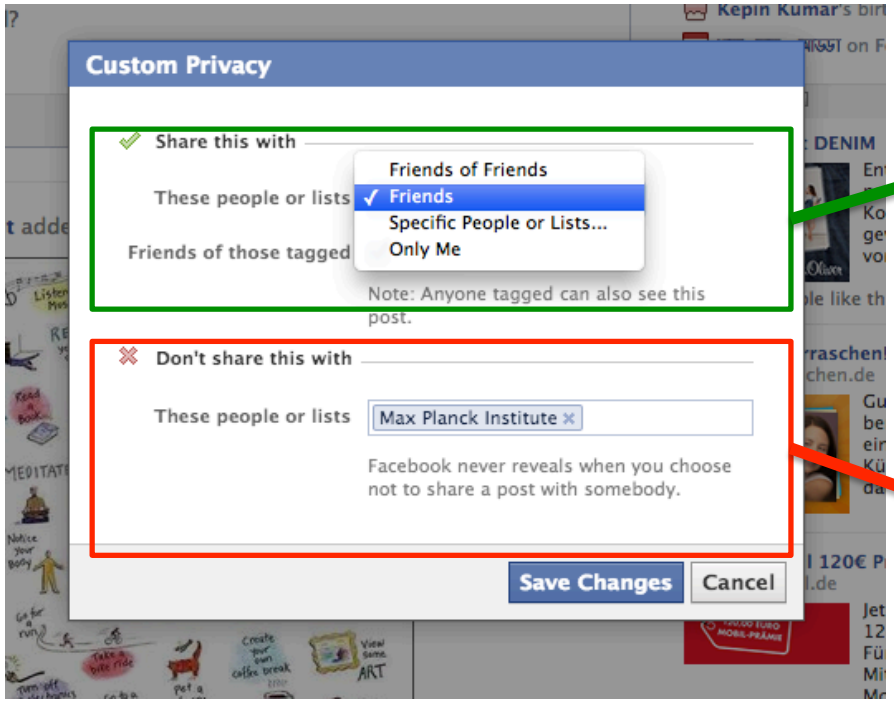
Situations where predicting exposure is possible

Predicting exposure of content in a site

Lots of research in content popularity growth

[Borghol et al] [Figueiredo et al.]
[Hong et al.] [Zaman et al]
[Bernstein et al.]

Managing privacy with Access Control Lists (ACLs)



Allow others access to content

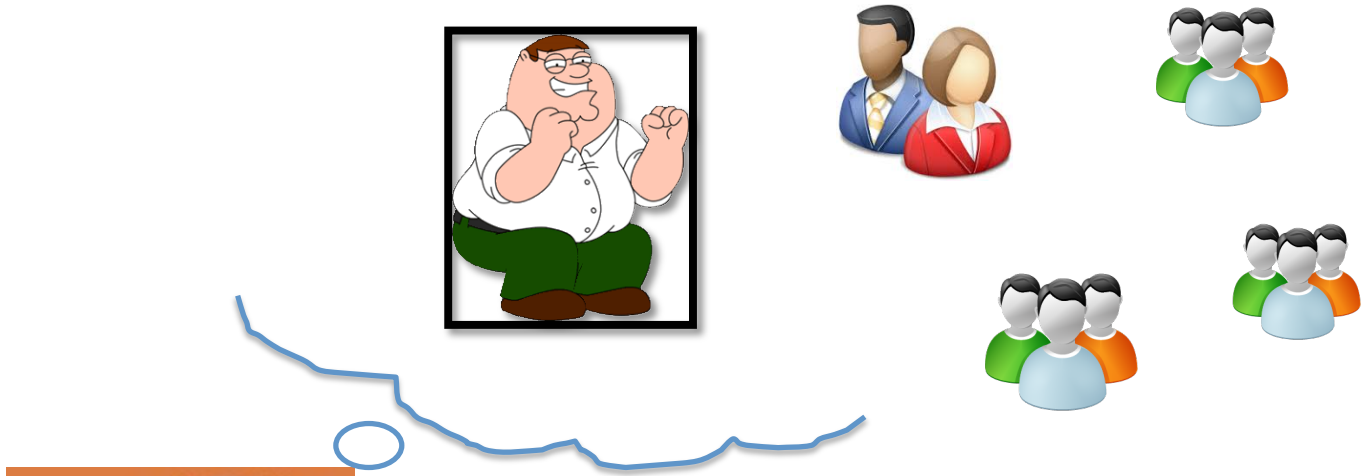
Deny others access to content

Privacy violation:

If someone accesses content who the user did not intend

ACLs don't capture many privacy violation scenarios

Modeling user privacy using exposure

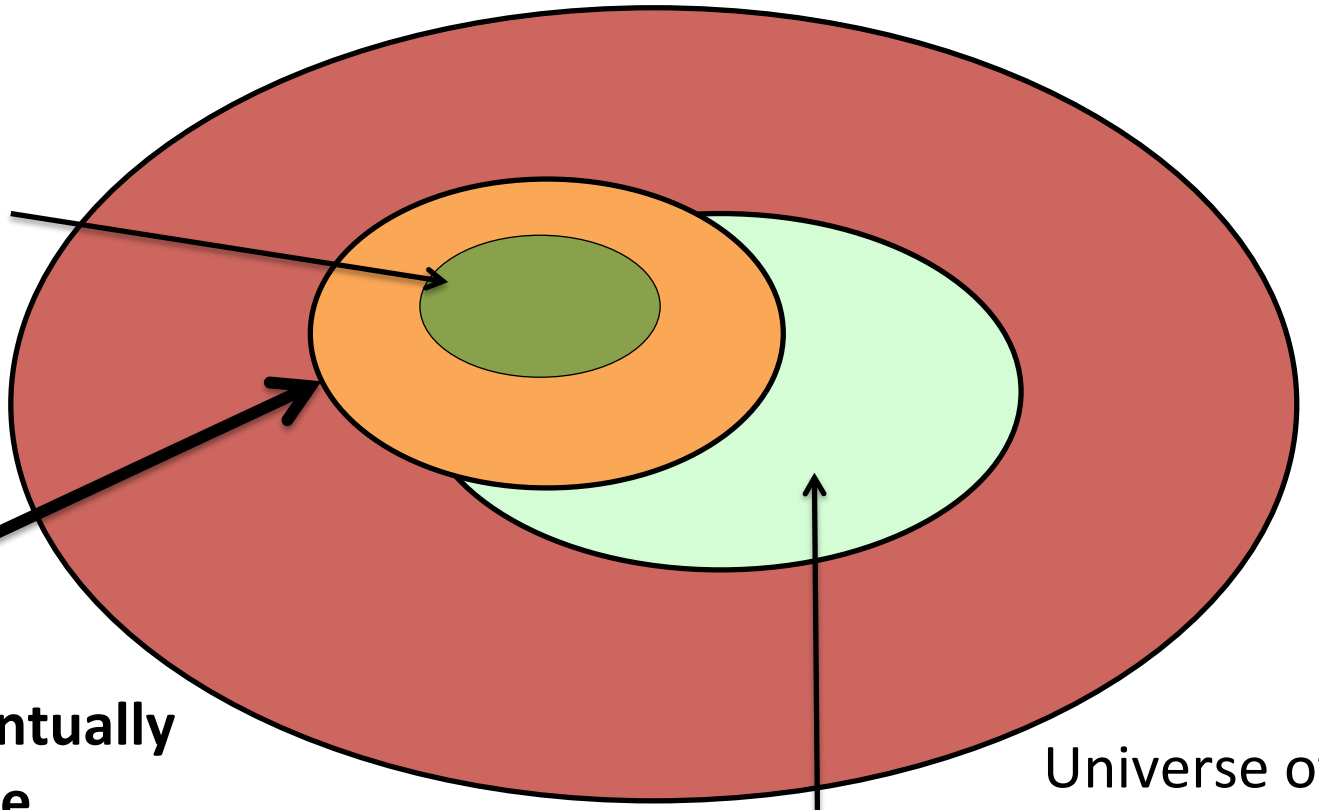


Deviation from desired exposure captures Privacy violation

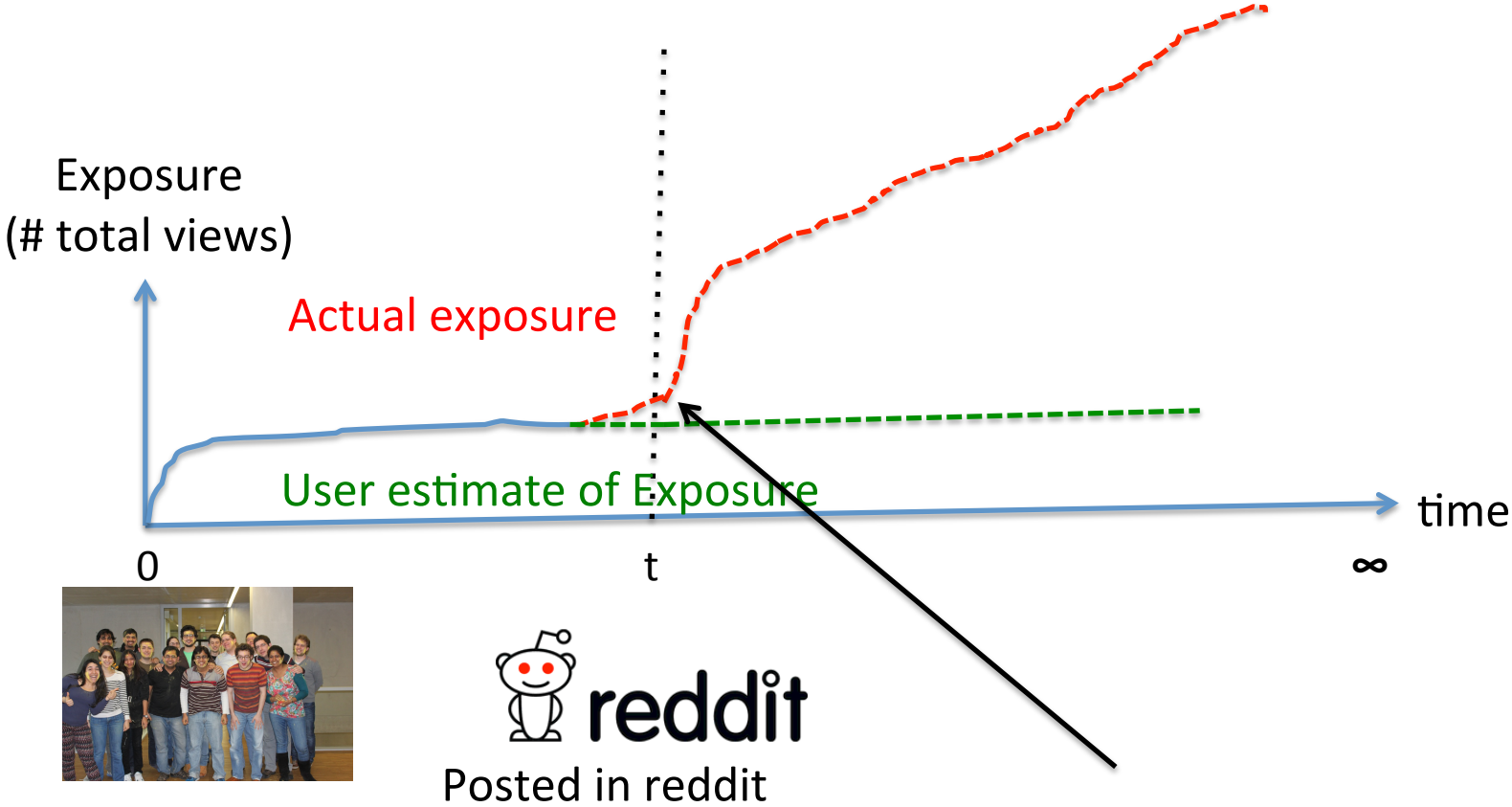
Change in exposure ⇒ Privacy violation

Exposure : Illustration

$\mathbf{I}$: Birthday of a user in Facebook



Exposure in more detail



A change in the exposure ⇒ chance of privacy violation

How accurately do users estimate exposure?

Facebook researchers did a study with 589 people

[Bernstein et al. 2013]

Question:

"How many people do you think saw it?" (i.e., a content)

Answer:

Desired exposure (median): 20

Actual exposure (median): 78

There may be a feeling of privacy violation when actual exposure is different from desired exposure