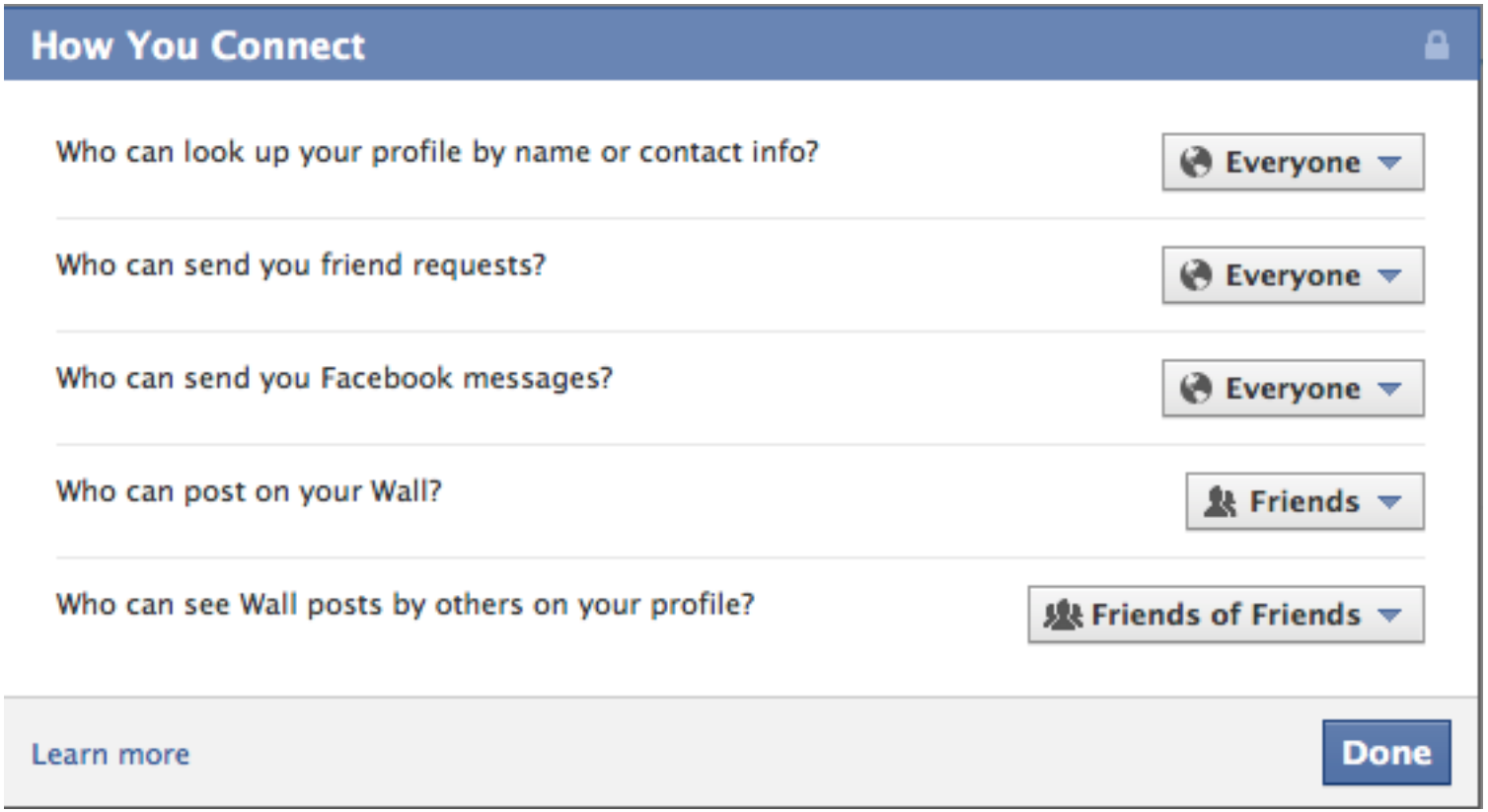
Max
Planck
Institute
for
Software Systems

Longitudinal Privacy Management in Social Media: The need for better controls

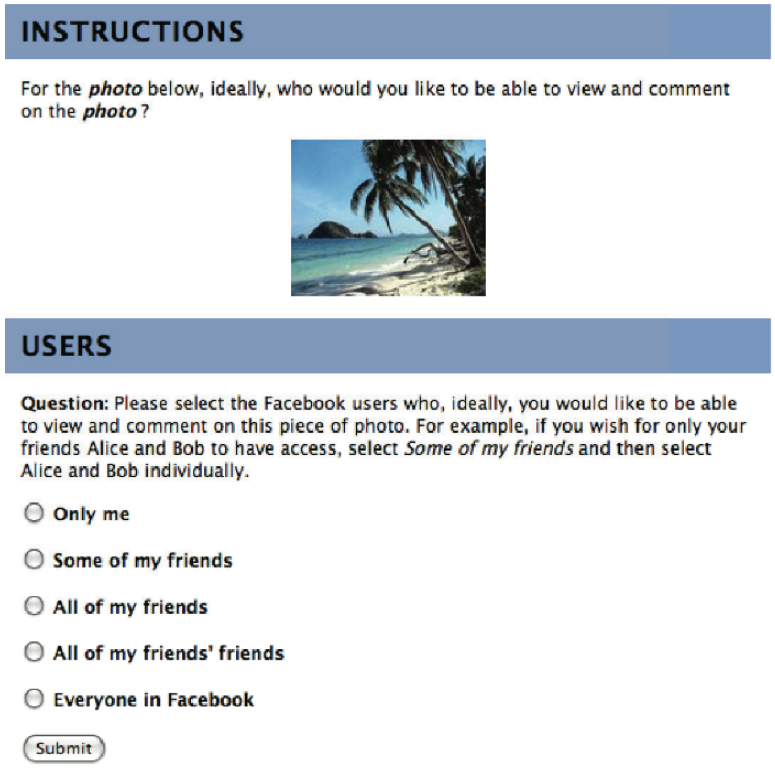
Mainack Mondal†

How users manage privacy settings today



These are the default settings in Facebook

Measuring the difficulty in managing privacy



Designed a Facebook privacy survey application *[IMC '11]*

Recruited 200 users with Amazon's Mechanical Turk

gathered demographic information, past history on Facebook

Survey results

Key findings:

a majority do not change their default settings

a majority do not understand their default settings

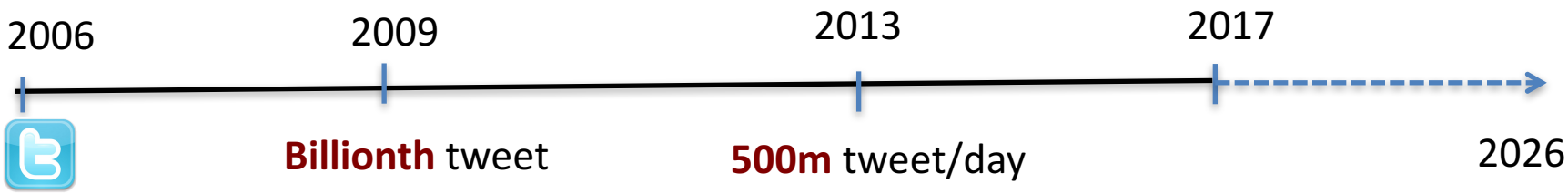
a majority of content is exposed to more users than desired

even when users' changed their defaults!

Because people changed their privacy preferences over time!

Online Social Media sites (OSMs) are aging

OSMs are already around for a decade



In sites like Twitter

- Users are **content creators and managers**
- They might even need **to change privacy preferences over time**

Users change privacy preferences over time

2009

Content posted in freshman year: shared with everybody on internet

2012

3 years later: Hiring manager and colleagues **should not** see this

They need to control **longitudinal exposure**: control **who can see old content**

Understanding longitudinal exposure control

[WPES 2013]
[SOUPS 2013]

Recent studies found via user surveys

- Users' willingness to share content drops as the content become old
- Willingness to share further decreases with a life-change

A large scale study on tweets posted within a week reported

- 2.4% of those tweets are deleted by users within their week of observation
- However **they only considered** content posted in **very recent past**

[CSCW 2013]

No investigation so far about

- Do users **change privacy preferences** to control **longitudinal exposure**
- How **effective** are **current mechanisms** to control longitudinal exposure?

Goal

To better understand and control longitudinal exposure in OSMs

Rest of the talk

- ✔ Do users change privacy preferences over time?
- ✔ How effective are these exposure control mechanisms?
- ✔ How can these mechanisms be improved?

Rest of the talk

✔ **Do users change privacy preferences over time?**

✔ How effective are these exposure control mechanisms?

✔ How can these mechanisms be improved?

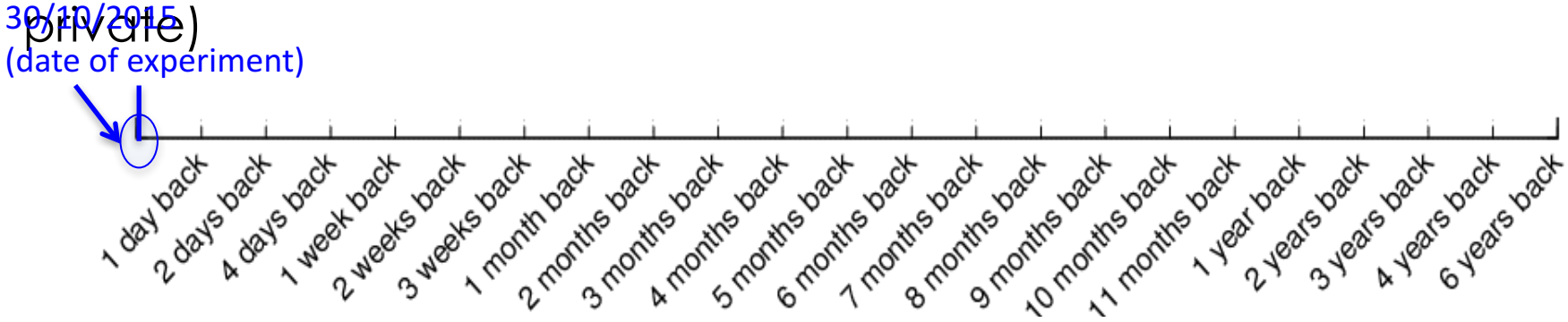
Collecting data on users changing privacy preferences

In this study we focus on Twitte

Simple privacy preferences

Either publicly visible to everyone

Or withdrawn from public domain (by deletion or making account private)



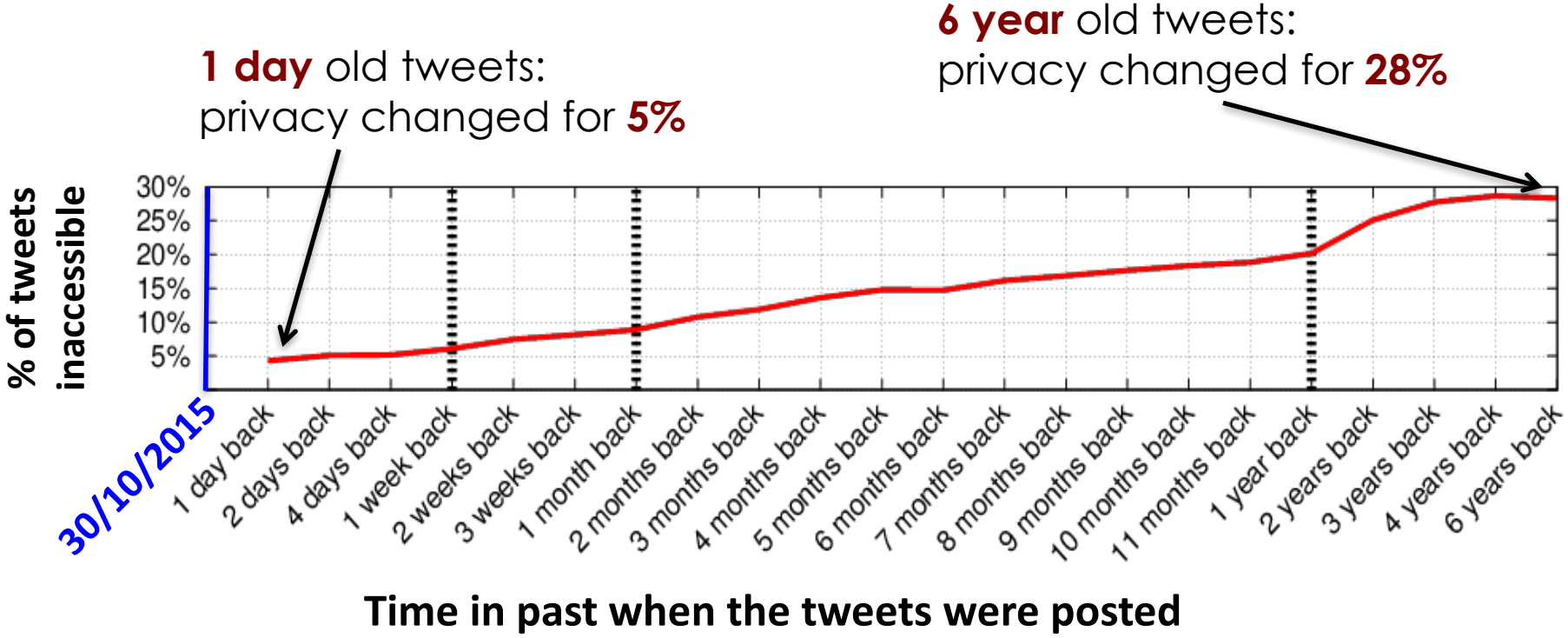
Time in past when the tweets were posted (relative to the date of experiment**)**

ranges from tweets posted **1 day back** to **6 years back**

All of these past tweets were **public when they were posted**

If **inaccessible** on experiment date, privacy preferences **changed** over time

Do users change privacy preferences over time?



Users change privacy for increasing amount of old data with time

How do these users change privacy of this content?

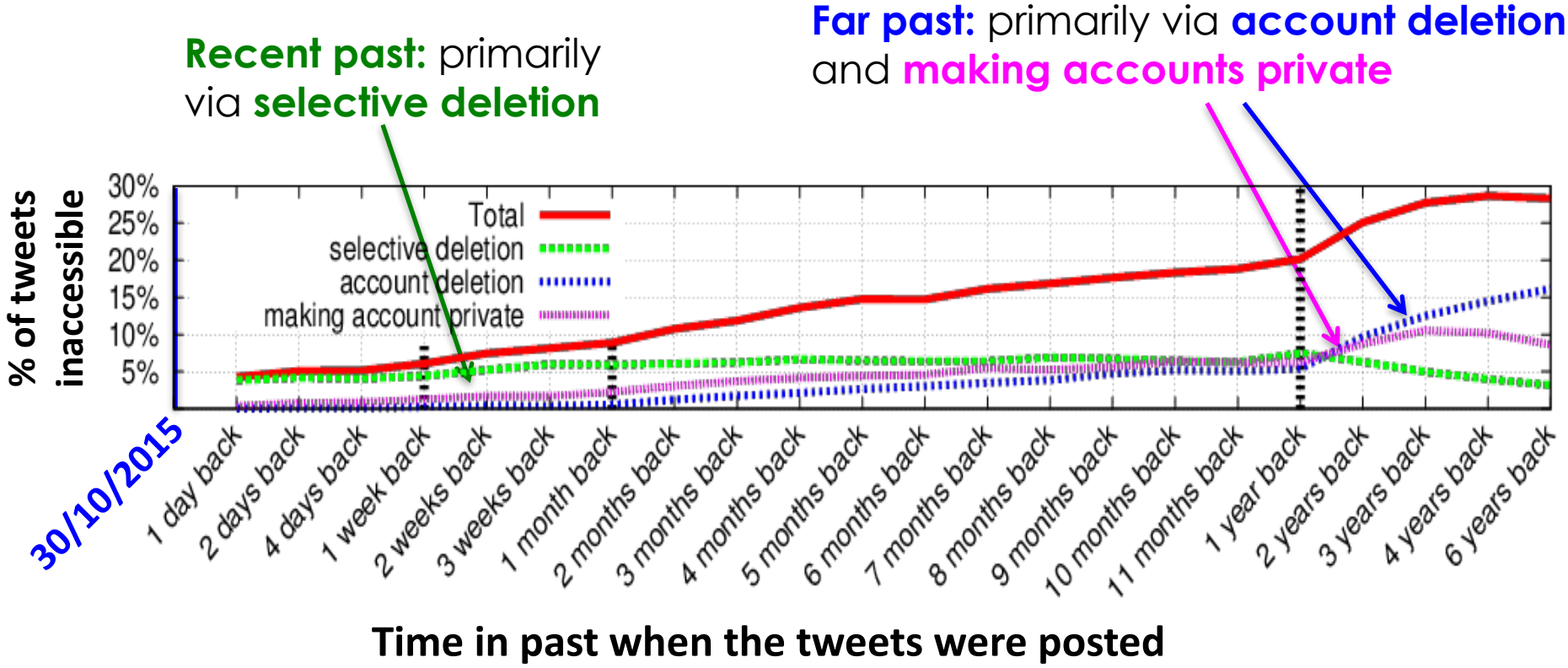
Mechanisms to change privacy preferences in Twitter

Three ways users change privacy of old content in Twitter

They are the longitudinal exposure control mechanisms

Mechanism	Description
Selective deletion	Selectively withdraw some old tweets to control exposure
Account deletion	Withdraw all old tweets to control exposure in bulk
Making account private	Withdraw all old tweets to control exposure in bulk

How do users change privacy preferences?



Very different mechanisms to change privacy for content from far past compared to recent past

Do many users change privacy of old content?

We randomly sample **100k** active users from 2009

Out of 8.9m random old tweets from these users 29.1% is inaccessible

What fraction of users change privacy of their content?

User type	% of all users
Selectively deleted tweets	8.3%
Deleted their account	15.9%
Made their account private	10.4%
Users who take actions that changes privacy of their content	**34.6%**

A significant fraction of users change privacy of their old content

Rest of the talk

✔ Do users change privacy preferences over time?

Privacy preferences are changed for significant fraction of old content

✔ **How effective are these exposure control mechanisms?**

✔ How can these mechanisms be improved?

Limitations of current exposure control

We identified two limitations of current exposure controls

Limitation 1: Retaining **residual activities**

Limitation 2: **Creating signal** to identify possibly sensitive content

Limitations of current exposure control

We identified two limitations of current exposure controls

Limitation 1: Retaining **residual activities**

Limitation 2: Creating signal to identify possibly sensitive content

What are residual activities?

Limitation 1: Retaining residual activities

These **conversations from** other users **remain public** even **after** a user **remove** her tweets/account

We call these **conversations residual activities**

Residual activities **contain information** about **withdrawn old content**

Anybody online can collect and analyze them by a username search

Residual activities might breach longitudinal exposure control

What information can we recover from residual activities?

Sensitive user interests revealed by residual activities

We checked user interests revealed for deleted/private accounts from 2009

Deleted/private accounts	Topics of interest from hashtags	Hashtags revealed by residual activities
Account 1	Politics, Sports, Technology	#iranelection, #prisoners, #strike, #frenchopen, #tech
Account 2	Sports, LGBTQ issues	#daviscup, #samesexsunday, #india, #lgbt, #followfriday
Account 3	Sports	#grandrapids, #nascar

Some of these hashtags can be considered **sensitive**

Residual activities also reveal

- Demographics of accounts
- Meaning of deleted tweets -- Check out our SOUPS 2016 paper for details

Residual activities can leak information about withdrawn accounts/tweets and breach longitudinal exposure control

We developed a web app for users to check residual activities

Check out the app is at: **http://twitter-app.mpi-sws.org/footprint/**

Limitations of current exposure control

We identified two limitations of current exposure controls

Limitation 1: Retaining residual activities

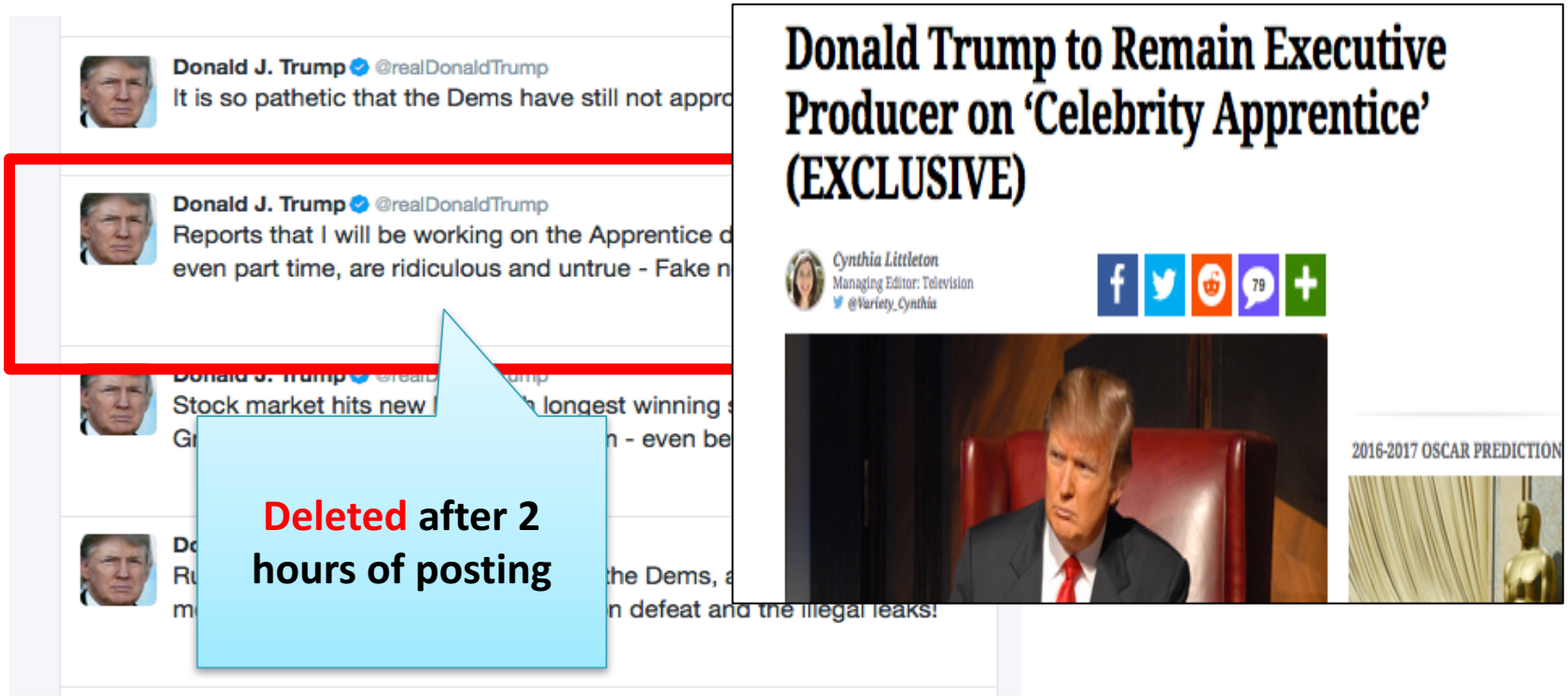
Limitation 2: **Creating signal** to identify possibly sensitive content

What do we mean by "creating signal"?

Limitation 2: Creating signals to identify possibly sensitive content

Which one is possibly sensitive?

Limitation 2: Creating signals to identify possibly sensitive content



An attacker can detect when your content is withdrawn

She can just compare snapshots

Withdrawal of a content **signals** an attacker to investigate the content

Controlling longitudinal exposure of content today helps an attacker to identify and investigate possibly sensitive content

There is already a web app for detecting deleted tweets of politicians

Check out Politwoops:

https://projects.propublica.org/politwoops/

Rest of the talk

✔ Do users change privacy preferences over time?

Privacy preferences are changed for significant fraction of old content

✔ How effective are these exposure control mechanisms?

Current mechanisms do not take care of information leakage by residual activities and creates signal to identify withdrawn content

✔ **How can these mechanisms be improved?**

Dealing with the limitations is difficult

Straw man:

Withdraw all the residual activities with original tweet/account

Problem

Residual activities are **not "owned"** by the original poster

Emerging OSMs deal residual activities by **age based withdrawal**

Withdraw all content after a preset time T (e.g. 24 hour

Snapchat, Cyber dust

Both original post as well as residual activities are withdrawn

No signal for specific sensitive tweets

Problem with age based withdrawal

1. Do **not facilitate interaction** with content ☑
2. **No archive** of past activities, **no long term memory**

Improved mechanism 1: Inactivity based withdrawal

Automatically withdraw content only when it is inactive

Inactive content: **no interaction** (e.g., retweets) for **time T**

In other words

No tweet receiving **recent interaction** will be withdrawn

Through simulation we discovered

Inactivity based withdrawal **allows more interactions**

Especially for popular tweets -- Details in our SOUPS 2016 paper

Even for this idea **no archive** of past activities, **no long term memory**

Improved mechanism 2: *Letheia*

Key idea: hide and unhide content periodically

- Withdrawn content is permanently hidden
- Attacker can not be sure if a content is deleted or hidden
- Provides plausible deniability to users

Time stamp 1

Time stamp 2

Time stamp 3

Each tweet is available 2 out of 3 timestamps : **67% availability**

Letheia: Design challenges

Letheia provides

Probabilistic privacy guarantees

A trade-off between availability and privacy

Letheia needs to decide

The hide and unhide time period distributions for each individual tweet

Through experimentation we discovered

Negative binomial distribution is a good choice for time distributions

Letheia **does not** take care of information leak by **residual activities**

It does raise the bar for attacker to collect data

Summary

Analyzed longitudinal exposure control from recent to very far past

Users control exposure by withdrawing **large amount of old data**

First study to analyze key limitations of current mechanisms

Residual activities **leak significant information** about **withdrawn content**

Withdrawal helps an attacker to **identify possibly sensitive** content

Inactivity based withdrawal is a mechanism **to stop information leakage** from residual activities and **facilitate interaction**

Letheia is a mechanism to **balance** between **privacy and availability**

THANKS!

Our Twitter web app to see your information leakage via residual activities: http://twitter-app.mpi-sws.org/footprint/

Our SOUPS 2016 paper -- "Forgetting in Social Media: Understanding and Controlling Longitudinal Exposure of Socially Shared Data "

Extra slides!

Leveraging data archived in the past

In this study we focus on data from Twitte 

We archived Tweets when they were posted over the past years

- All of those tweets were **public when they were posted**
- Currently we have **tweets collected continuously from 2009**!

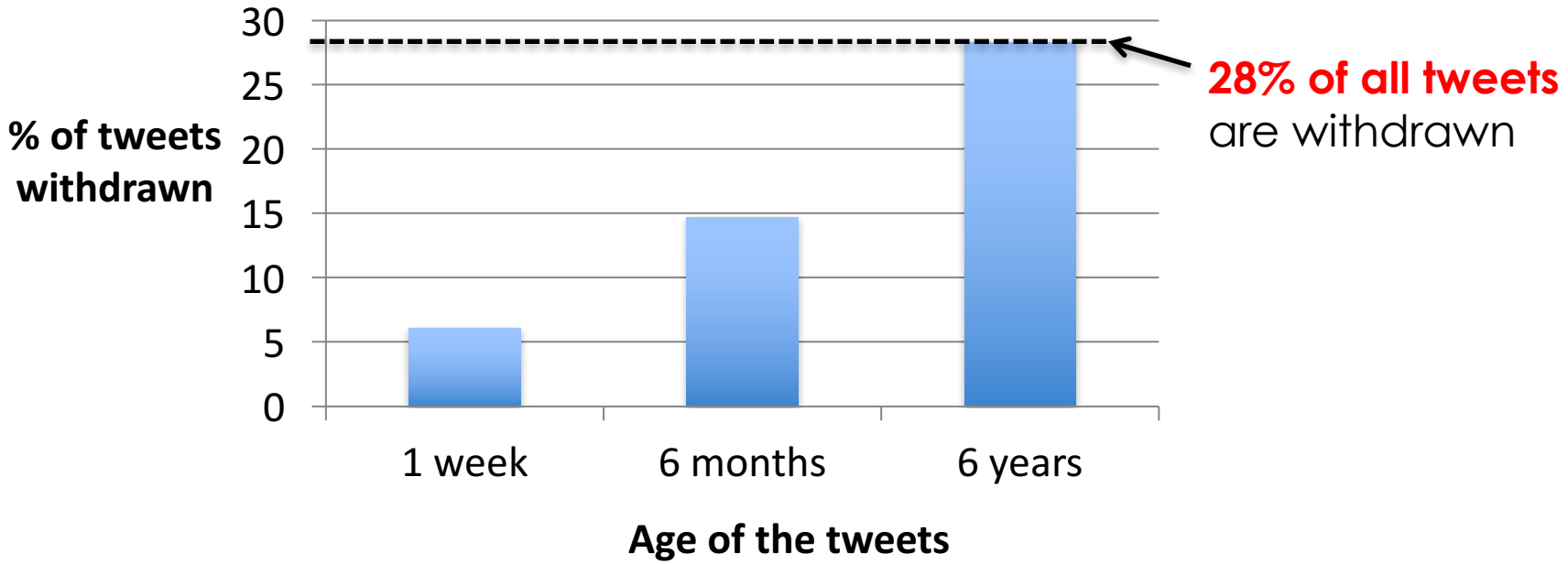
We try to **re-fetch** these archived tweets using Twitter API **today**

- If **inaccessible**, then users have **withdrawn those old tweets**

Note: We did not study the user intention behind these withdrawals

- Assumption: withdrawal of old tweets are due to privacy

Do users actually change the privacy preferences over time?

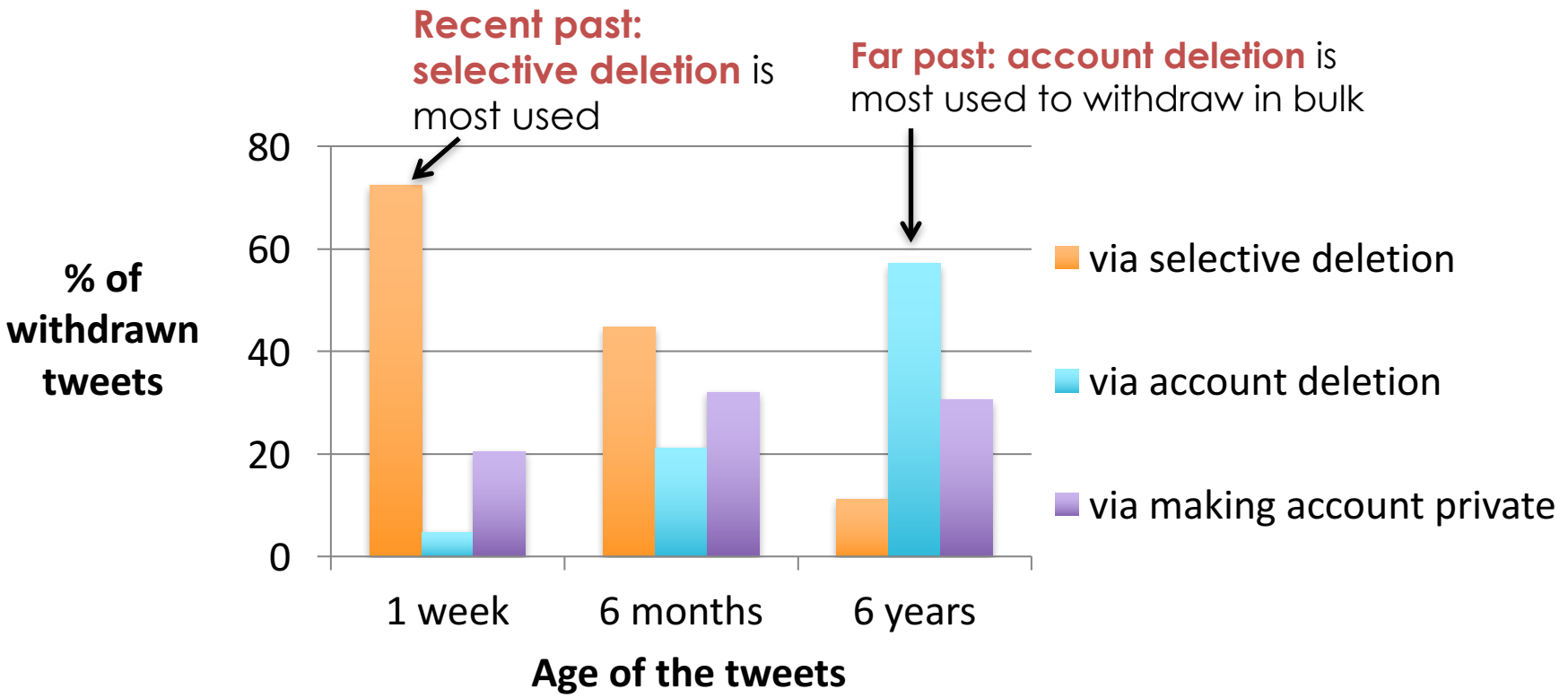


Users **change privacy preferences for 28%** of the data posted 6 years bac

Users withdrew surprisingly large amount of old data to control exposure

What mechanisms did they use to withdraw this old data?

How do the users change privacy of their old content?



Tweets posted in recent to far past were withdrawn via very different exposure control mechanisms

How do individual users control exposure of their old content?

Classifying users by exposure control behavior

We randomly sample **100k** active users from 2009

Out of 8.9m random old tweets from these users 29.1% is withdrawn

Based on their exposure control behavior three categories

user category	description	% of all users
Non-withdrawers	did not withdraw any tweet	65.4%
Partial withdrawers	selectively deleted few old tweets	8.3%
Complete withdrawers	withdrew all tweets (deleted/private account)	26.3%

A limitation in current exposure control mechanisms will effect substantial fraction of users

Information leakage via residual activities

Selectively withdrawn tweets have residual activities **in the form of replies**

We can recover part of the withdrawn tweets from these reply tweets

Withdrawn accounts have residual activities **in the form of user mentions**

91.4% of the our complete withdrawers have residual activities

From these residual activities we can recover information like

- Social connections of the complete withdrawers
- Demographics information
- Interests of the account

Recovering user interest from residual activities

Intuition:

Residual activities might reveal user interests in the form of hashtags

Are hashtags in residual activities also used by complete withdrawers?

In **25%** of the cases **all the hashtags in residual activities** are **also used** by complete withdrawers

Complete withdrawer	Topics of interest from hashtags	Hashtags revealed by residual activities
withdrawer 1	Politics, Sports, Technology	#iranelection, #prisoners, #strike, #frenchopen, #tech
withdrawer 2	Sports, LGBTQ issues	#daviscup, #samesexsunday, #india, #lgbt, #followfriday
withdrawer 3	Sports	#grandrapids, #nascar

Residual activities leak information about withdrawn tweets /accounts and breach longitudinal exposure control

Check out our Twitter app to see your information leakage via residual activities: http://twitter-app.mpi-sws.org/footprint/

Recovering user interest from residual activities

Intuition:

Residual activities might reveal user interests in the form of hashtags

Are hashtags in residual activities also used by complete withdrawers?

In **25%** of the cases **all the hashtags in residual activities** are **also used** by complete withdrawers

Complete withdrawer	Topics of interest from hashtags	Hashtags revealed by residual activities
withdrawer 1	Politics, Sports, Technology	#iranelection, #prisoners, #strike, #frenchopen, #tech
withdrawer 2	Sports, LGBTQ issues	#daviscup, #samesexsunday, #india, #lgbt, #followfriday
withdrawer 3	Sports	#grandrapids, #nascar

Existing longitudinal exposure control mechanisms

Spectrum of existing longitudinal exposure control mechanisms

Putting users in charge	**Age based withdrawal**
Users control longitudinal exposure by **individual withdrawal**	All data are **automatically deleted after time T**
Deployed in most of the OSMs like Twitter, Facebook	Deployed in few emerging OSMs like snapchat
☑ Facilitate interaction	☒ Do not facilitate interaction
☒ Do not take care of residual activity	☑ Also deletes residual activity

Can we have best of both the mechanisms?

Our proposal: Inactivity based withdrawal

Key idea: Only when a content becomes **inactive, automatically withdraw**

Inactive content: no interaction (e.g., retweets) for time T

We compared age based and inactivity based withdrawal

We simulated both the strategies for 30k tweets with retweets

	Interactions (Retweets) stopped
Withdraw tweets after **1 day of posting**	7,798
Withdraw tweets after **1 day of inactivity**	4,117

Inactivity based withdrawal **allows more interactions**

Moreover it facilitate interactions for popular tweets -- Details in the paper

Inactivity based withdrawal stops information leak via residual activities and at the same time facilitates interaction in OSMs

Rest of the talk

- ✔ **How to collect data on changed privacy preference over time?**
- ✔ Do users change privacy preferences over time?
- ✔ How effective are these exposure control mechanisms?
- ✔ How can these mechanisms be improved?

Rest of the talk

✔ How to collect data on changed privacy preference over time?

Check which past public tweets are withdrawn from public

✔ Do users actually change privacy preferences over time?

Privacy preferences are changed for significant fraction of old content

✔ **How effective are these exposure control mechanisms?**

✔ How can these mechanisms be improved?

Rest of the talk

✔ How to collect data on changed privacy preference over time?

Check which past public tweets are withdrawn from public

✔ Do users actually change privacy preferences over time?

Privacy preferences are changed for significant fraction of old content

✔ How effective are these exposure control mechanisms?

Current mechanisms do not take care of information leakage by residual activities

✔ **How can these mechanisms be improved?**

Improvement of longitudinal exposure control mechanisms

Emerging OSMs deal residual activities by **age based withdrawal**

- **Withdraw all content** after a preset time T (e.g. 24 hours)
- Snap chat, Cyber dust

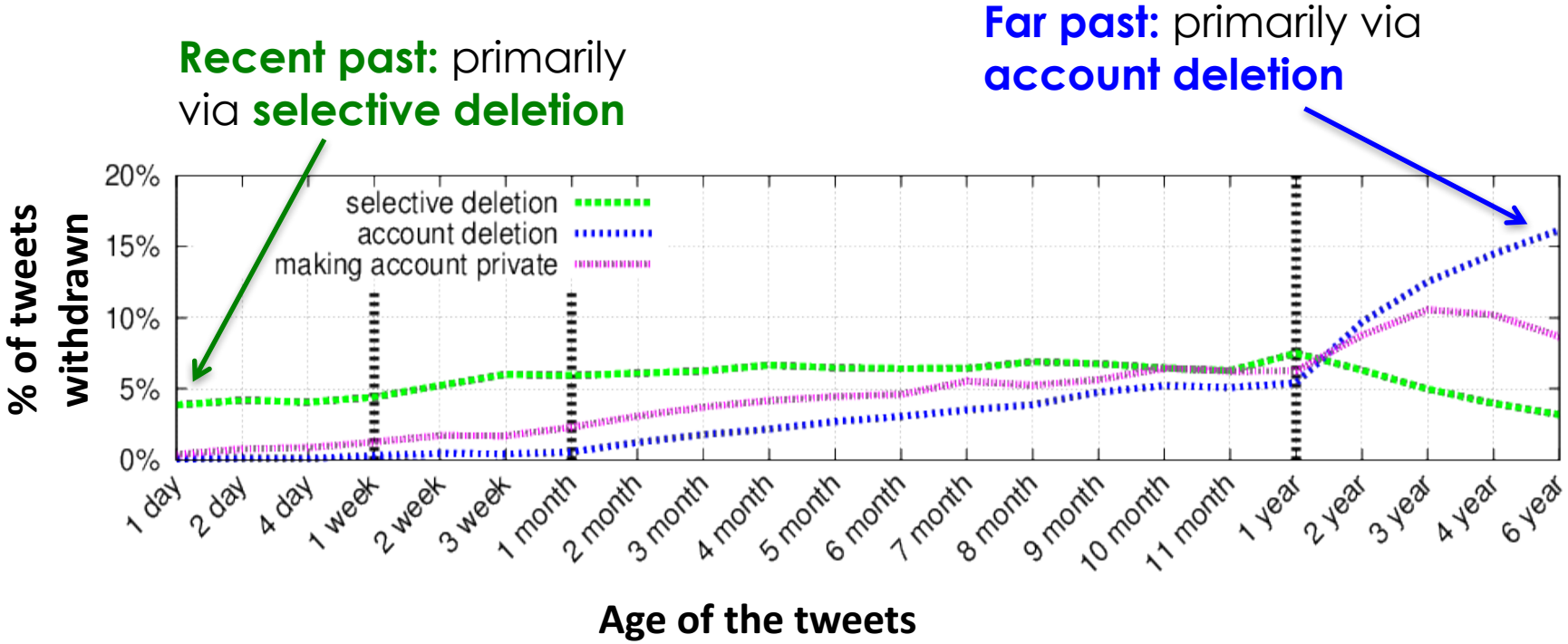
All content including residual activities are deleted after T

But they **do not facilitate interaction** with content

- Other users do not have enough time to comment or share
- **Content might not become popular!**

Can we improve age based withdrawal to facilitate interactions?

How do users change privacy preferences?



Very different mechanisms for changing privacy for content from far past compare to recent past

Improvement of longitudinal exposure control mechanisms

Emerging OSMs deal residual activities by **age based withdrawal**

- **Withdraw all content** after a preset time T (e.g. 24 hours)
- Snap chat, Cyber dust

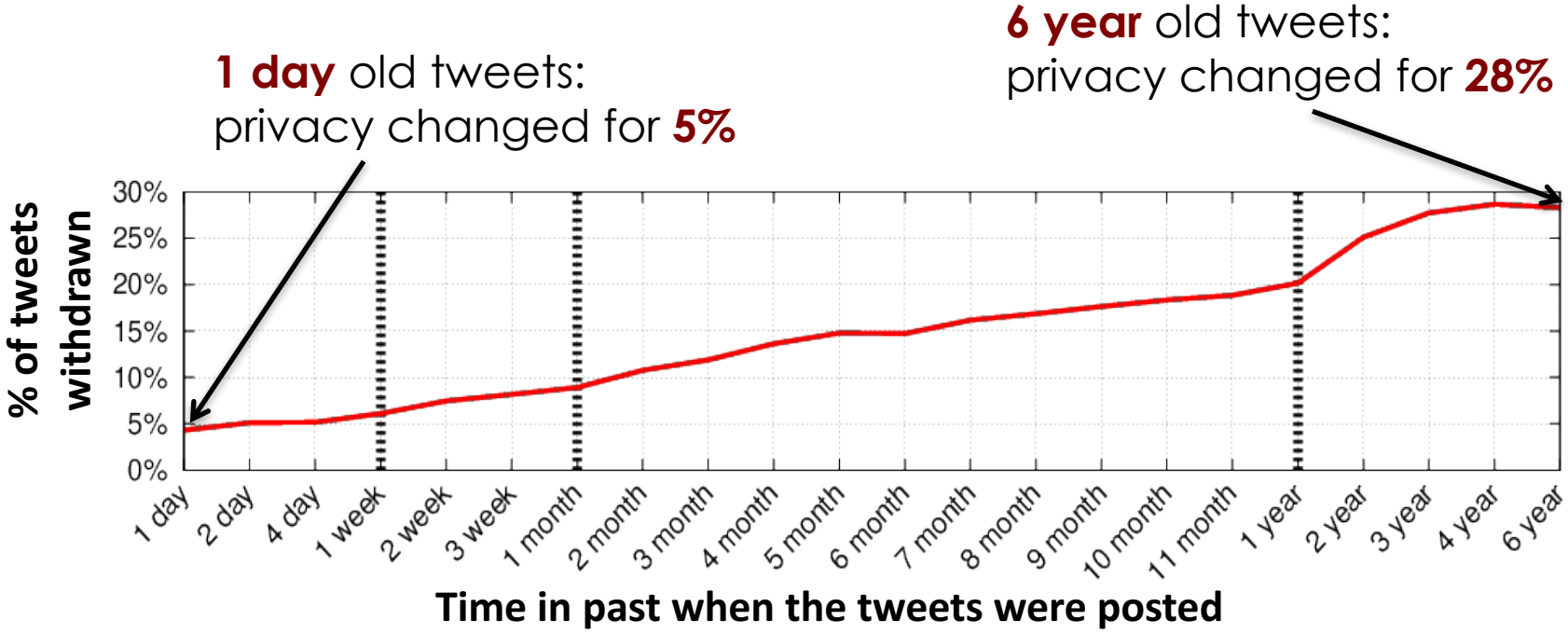
All content including residual activities are deleted after T

But they **do not facilitate interaction** with content

- Other users do not have enough time to comment or share
- **Content might not become popular!**

Can we improve age based withdrawal to facilitate interactions?

Do users change privacy preferences over time?



Users change privacy for increasing amount of old data with time

How do these users change privacy of this content?

Collecting data on users changing privacy preferences

In this study we focus on Twitter 

Simple privacy preferences

Either publicly visible to everyone

Or withdrawn from public domain (by deletion or making account private)

We archived tweets continuously, starting from 2009

All those tweets were **public when they were posted**

We try to **re-fetch** these archived tweets **today**

If **inaccessible**, then the privacy preferences changed over time

In other words users **controlled longitudinal exposure**

Information leakage via residual activities

Accounts have residual activities **in the form of user mentions**

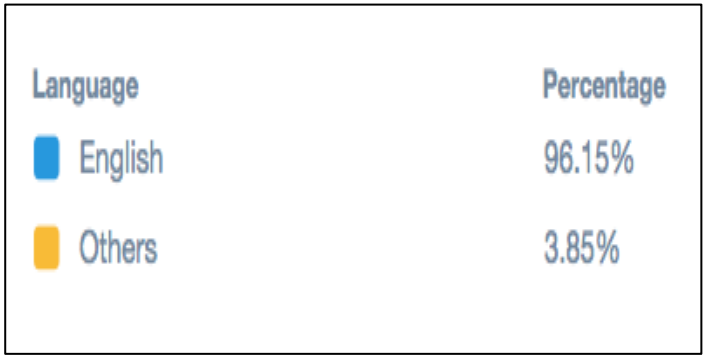
We developed an app for users to check residual activities

The app is at: **http://twitter-app.mpi-sws.org/footprint/**

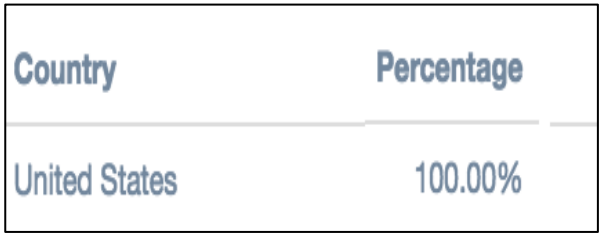
User demographics revealed by residual activities

Residual activities reveals likely **location or language** of user

Check it out at: **http://twitter-app.mpi-sws.org/footprint/**



Language	Percentage
English	96.15%
Others	3.85%



Country	Percentage
United States	100.00%

Tweets also have residual activities in the form of **replies**

Reveal information about **meaning of the tweet**

Check out our app and paper for details

Comparison of age and Inactivity based withdrawal

We simulated both the strategies for 30k tweets with retweets
Take retweets as proxy for interaction

	Interactions (Retweets) stopped
Withdraw tweets after **1 day of posting**	7,798
Withdraw tweets after **1 day of inactivity**	4,117

Inactivity based withdrawal **allows more interactions**

Especially for popular tweets -- Details in the paper

Inactivity based withdrawal stops information leak via residual activities and at the same time facilitates interaction in OSMs

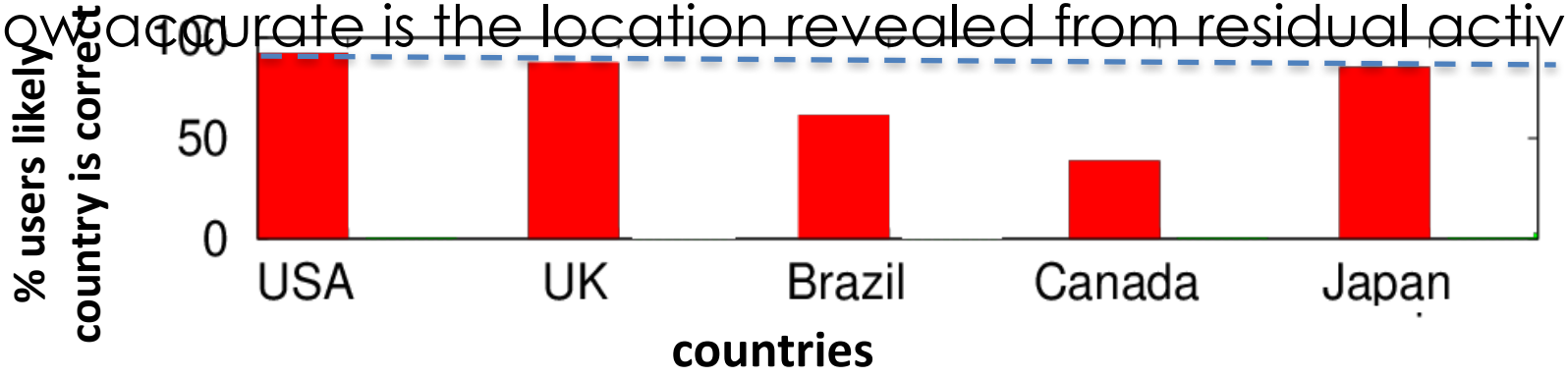
User demographics from residual activities

Residual activities reveals likely **location or language** of user

Likely location: country of users generating majority residual activities

Ground truth: Country reported in the deleted/private accounts

How accurate is the location revealed from residual activities?



Even for countries like **Japan** revealed location is **highly accurate**

Tweets also have residual activities in the form of **replies**

Reveal information about **meaning of the tweet**

Check out our app and paper for details

Demographics of users changing privacy

We investigated the demographics of our users from 2009

Inferred **Gender** and location form Twitter profile

User type	% female users
Random sample of Twitter users	50.3%
Users who **did not delete** any content	**44.5%**
Deleted tweet selectively	**55.7%**
Deleted account/Made **account private**	**61.5%**

Female users are more likely to **change privacy** of old content

User interests revealed by residual activities

Residual activities reveal **hashtags** which identify likely **user interests**