Deanonymizing Web Browsing Data With Social Networks

7cef2fe99fbf
Stanford University

98549815bd
Stanford University

86f08df318
Stanford University

9f514851f6
Princeton University

Deanonymizing Web Browsing Data With Social Networks

Jessica Su
Stanford University

Ansh Shukla
Stanford University

Sharad Goel
Stanford University

Arvind Narayanan
Princeton University

Introduction

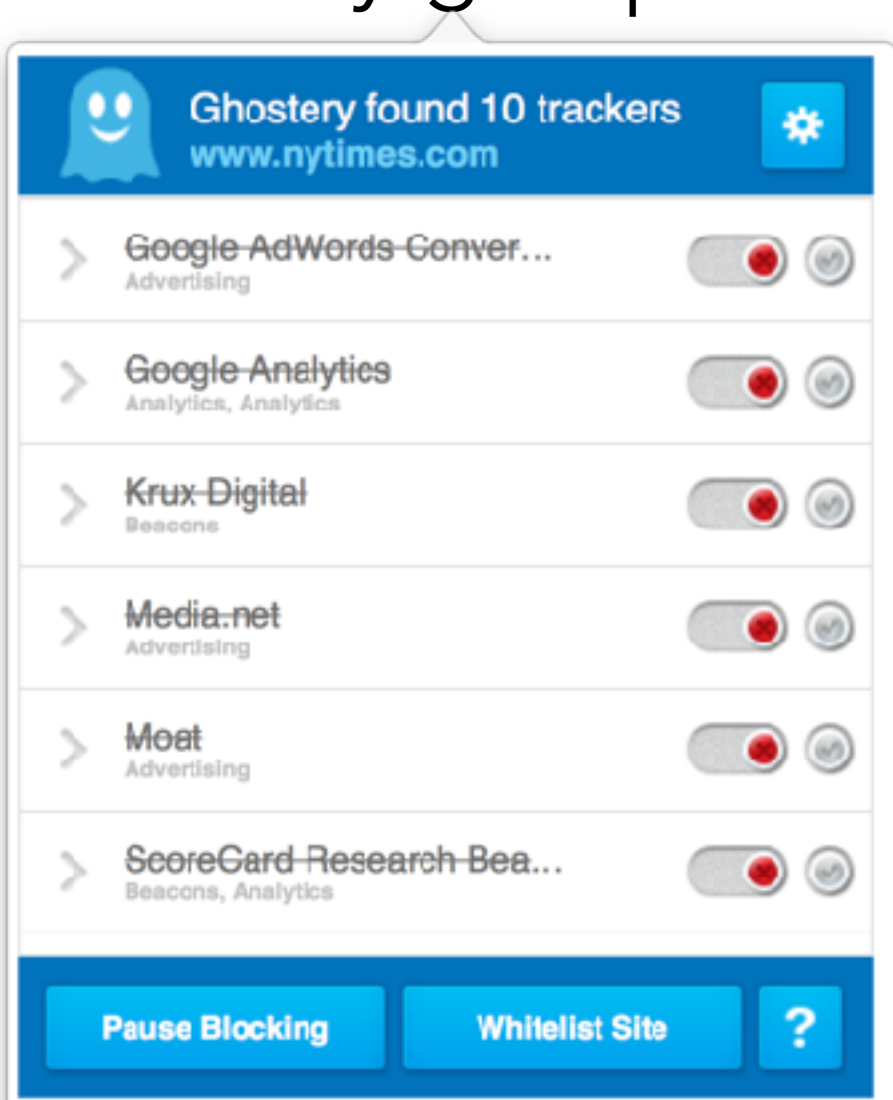
Introduction

Privacy is legally and fundamentally important.

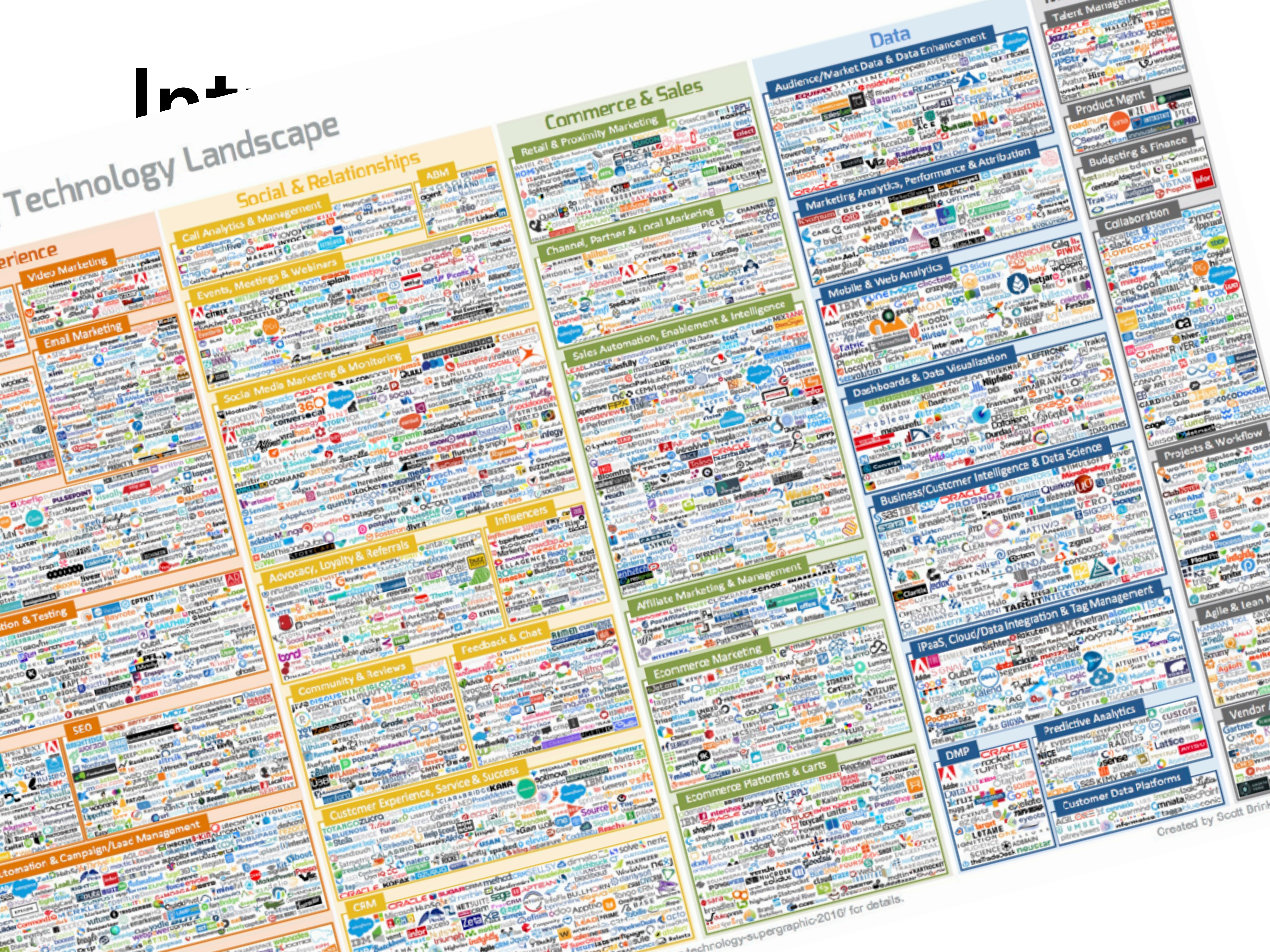
Introduction

Privacy is legally and fundamentally important.

Many groups collect private web browsing data.



Int


Technology Landscape
Video Marketing
Email Marketing
SEO
Social & Relationships
Call Analytics & Management
ABM
Events, Meetings & Webinars
Social Media Marketing & Monitoring
Influencers
Advocacy, Loyalty & Referrals
Feedback & Chat
Community & Reviews
Customer Experience, Service & Success
CRM
Commerce & Sales
Retail & Proximity Marketing
Channel, Partner & Local Marketing
Sales Automation, Enablement & Intelligence
Affiliate Marketing & Management
Ecommerce Marketing
Ecommerce Platforms & Carts
Data
Audience/Market Data & Data Enhancement
Marketing Analytics, Performance & Attribution
Mobile & Web Analytics
Dashboards & Data Visualization
Business/Customer Intelligence & Data Science
iPaaS, Cloud/Data Integration & Tag Management
DMP
Predictive Analytics
Customer Data Platforms
Product Mgmt
Budgeting & Finance
Collaboration
Projects & Workflow
Created by Scott Brinker

Introduction

Privacy is legally and fundamentally important.

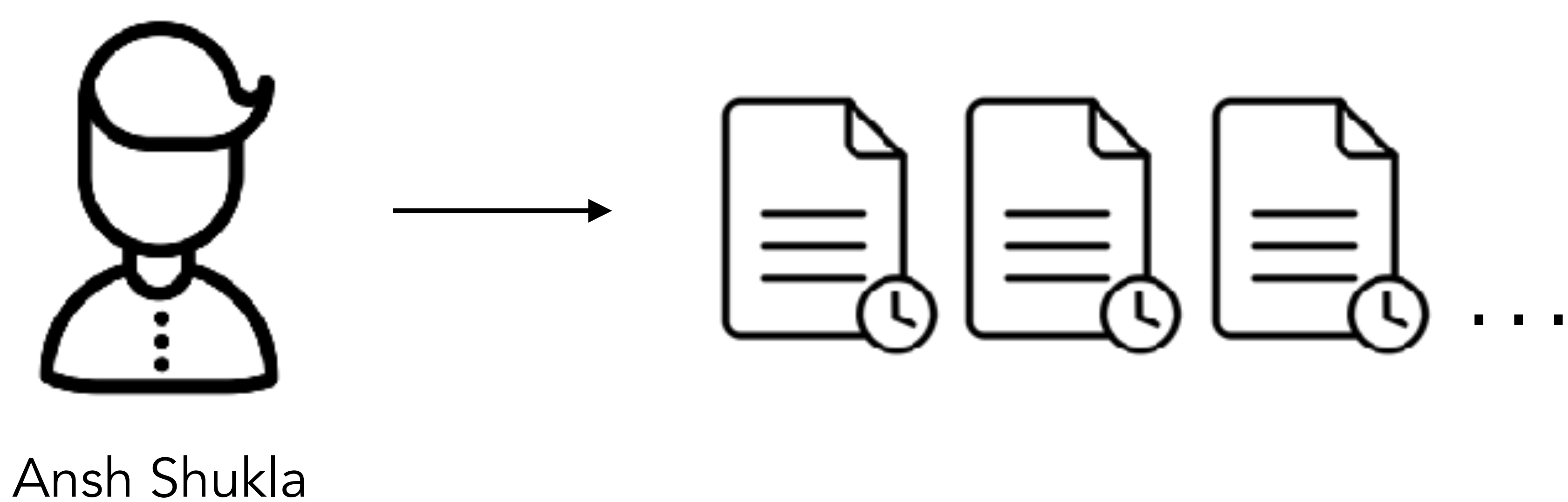
Many groups collect private web browsing data.

Introduction

Privacy is legally and fundamentally important.

Many groups collect private web browsing data.

Data collection is justified by scrubbing PII.

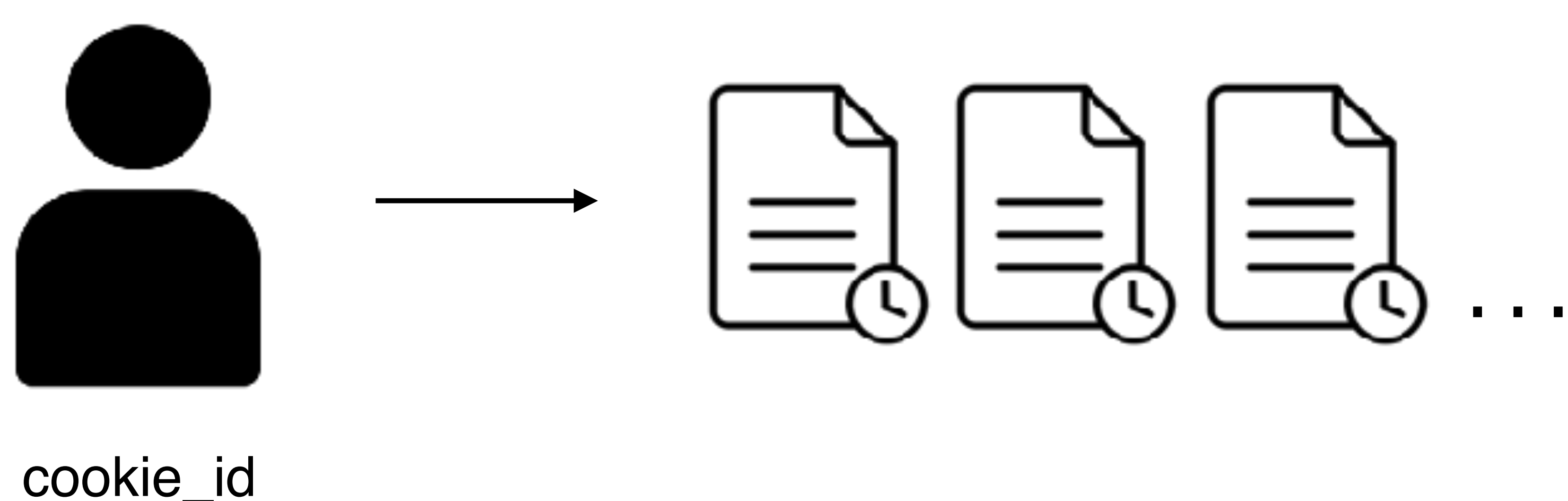


Introduction

Privacy is legally and fundamentally important.

Many groups collect private web browsing data.

Data collection is justified by scrubbing PII.



Introduction

Do “anonymized” web browsing histories protect privacy?

Introduction

Verify Your History

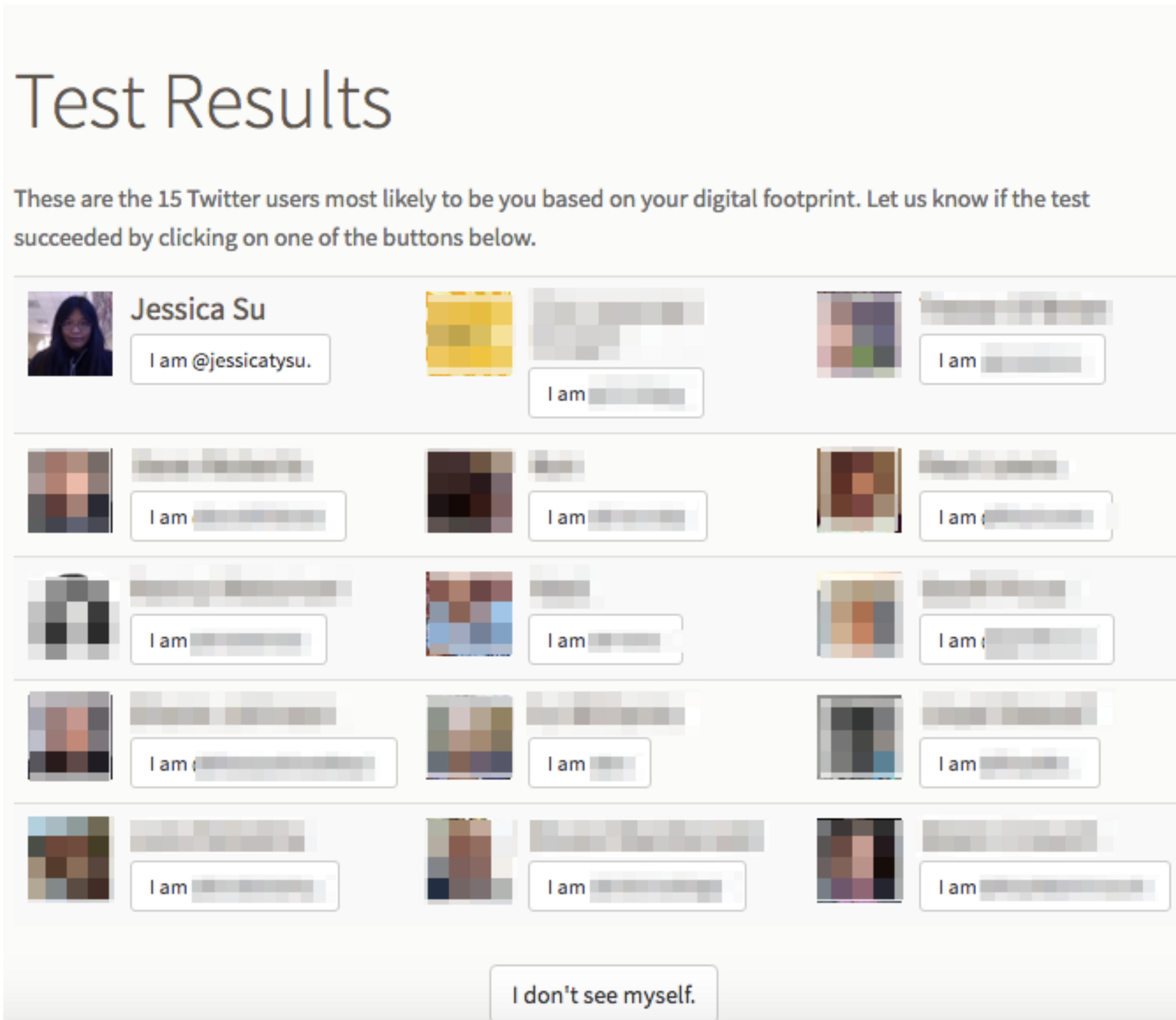
The following 16 links will be sent to our server and analyzed. Please confirm that you want to test this history by clicking the button below. If you do not want to test your history, click the "Don't send" button to uninstall the extension.

Link	***Expanded***
https://t.co/WBm8XdyVLY	*on.wsj.com/2c8O1ea*
https://t.co/iQbvXrFVen	*www.quora.com/What-are-the-economics-of-all-you-can-eat-buff...*
https://t.co/wDsnH2OxsD	*thecooperreview.com/6-tips-how-to-be-thought-leader/*
https://t.co/0EYHupFTrt	*dld.bz/eJm9B*
https://t.co/JNqFhFyylc	*www.quora.com/Did-ancient-people-perceive-less-colours-than-...*
https://t.co/0QI9lKTVxL	*waitbutwhy.com/2016/09/marriage-decision.html*
https://t.co/COTSo2FTJF	*www.washingtonexaminer.com/army-slide-lists-clinton-as-insid*

I confirm, let's continue.

Don't send these links.

Introduction



Introduction

72%

of people we tried to deanonymize
were correctly matched to their Twitter profile.

How does it work?

People tend to click on links that appear in their Twitter feed

Check if the browsing history contains a lot of obscure links from someone's feed

The set of people you follow
on Twitter is very distinctive,
and many links posted on Twitter
are shown to a small set of people

The Twitter links in your
browsing history are often
enough to uniquely identify you!

Observe that your privacy
can be violated, even if you
don't post anything

Problem definition

Twitter feed

Browsing history

https://facebook.com

http://cs246.stanford.edu

http://voxeu.org/article/...

Given an anonymous browsing history,
match it to the closest possible Twitter feed

What is the "best feed?"

Naive approach

Choose the Twitter feed that contains the most links from the browsing history.

https://facebook.com

http://cs246.stanford.edu

http://voxeu.org/article/...

Intersection size: 1

Naive approach

Choose the Twitter feed that contains the most links from the browsing history.

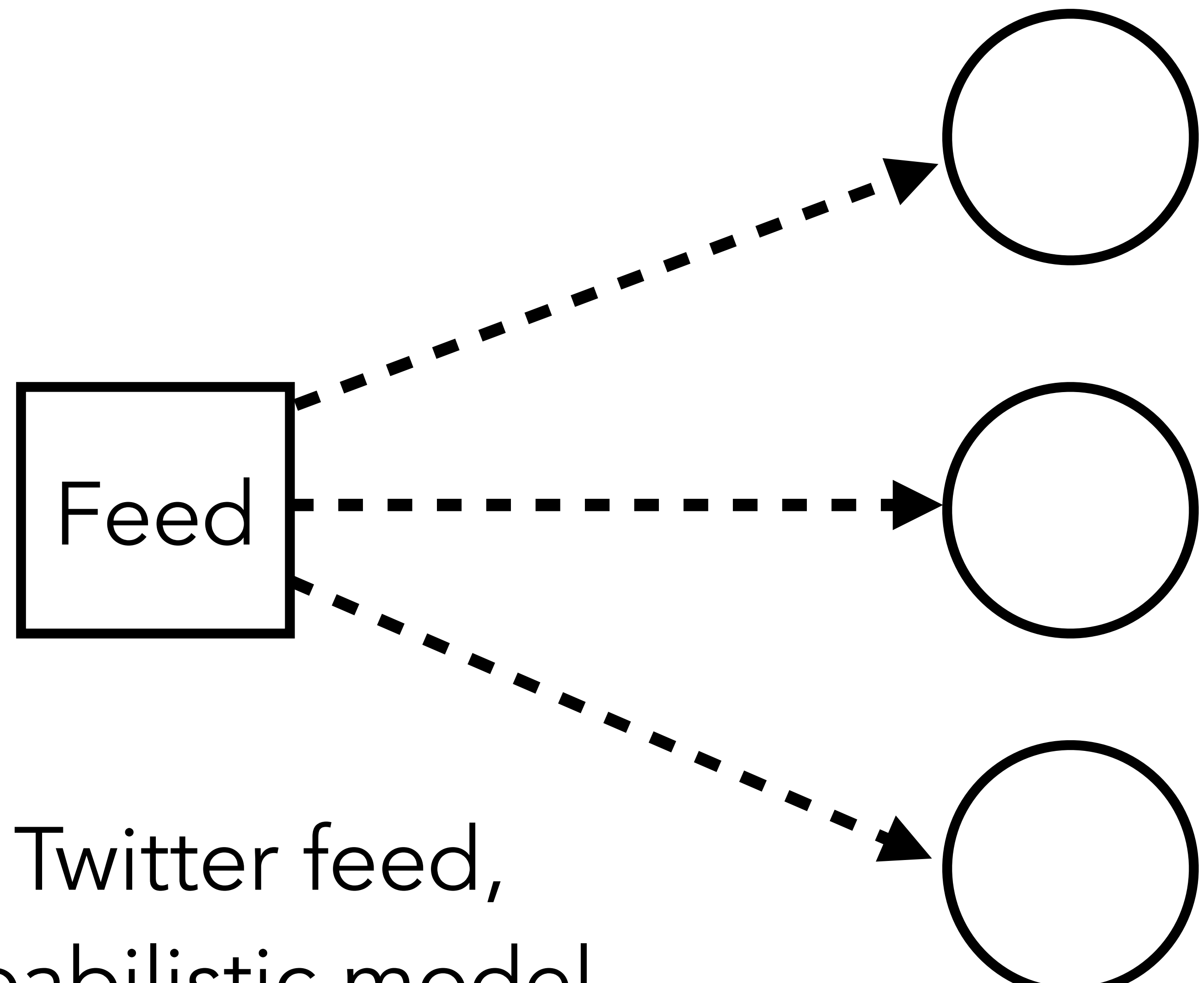
https://facebook.com
http://cs246.stanford.edu
http://voxeu.org/article/...

Intersection size: 1

Problem: Doesn't account for feed size.

Our approach

Step 1: Create a model of web navigation



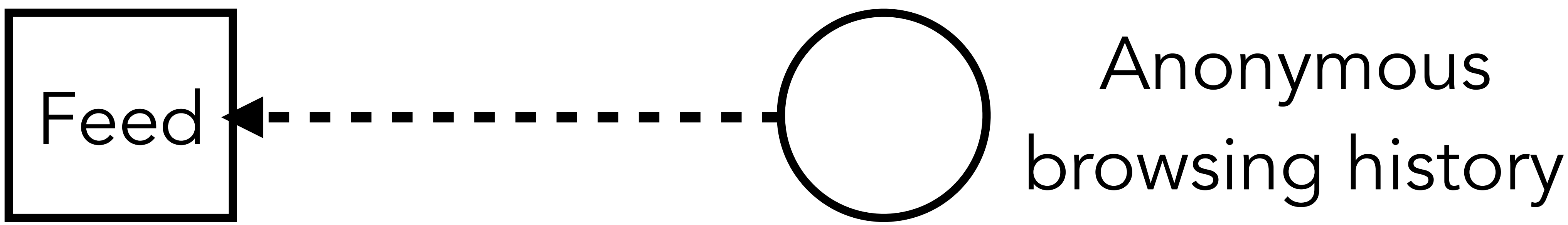
Potential browsing histories, generated based on the Twitter feed

Given a Twitter feed, use a probabilistic model to assign a probability to any sequence of web visits

The Twitter feed is a parameter of the model

Our approach

Step 2: Maximize the likelihood



Given an anonymous browsing history, find the model parameters that maximize the likelihood of the history

The model parameters correspond to the set of links in a person's Twitter feed, which tells you the identity of the user

Web navigation model

Probability of visiting a URL is proportional to

rp	if the URL is in your Twitter feed
p	otherwise

r is a parameter that depends on the user
p is the baseline popularity of the specific URL

Maximum likelihood estimation

Roughly equivalent to choosing the user whose feed maximizes

$$\texttt{intersection\_size} \cdot \log\left(\frac{\texttt{intersection\_size}}{\texttt{feed\_size}}\right)$$

Maximum likelihood estimation

Roughly equivalent to choosing the user whose feed maximizes

$$\texttt{intersection\_size} \cdot \log\left(\frac{\texttt{intersection\_size}}{\texttt{feed\_size}}\right)$$

This balances finding Twitter feeds that contain a lot of the links from the browsing history with finding Twitter feeds that don't contain too many links in general

How do we run this in real-time?

Implementation

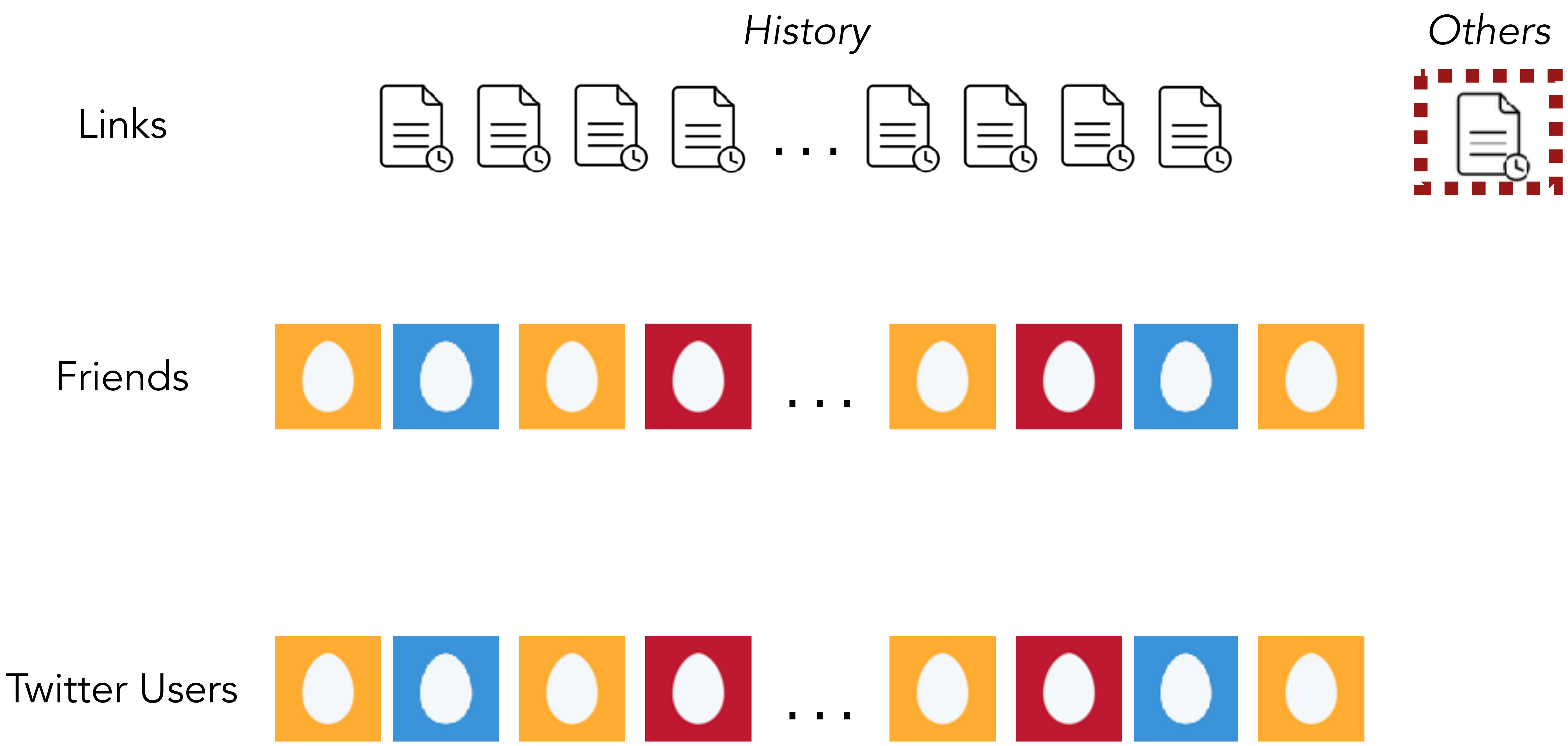
Need feed_size and intersection_size to calculate MLE score.

Need MLE score for *all* users in order to rank.

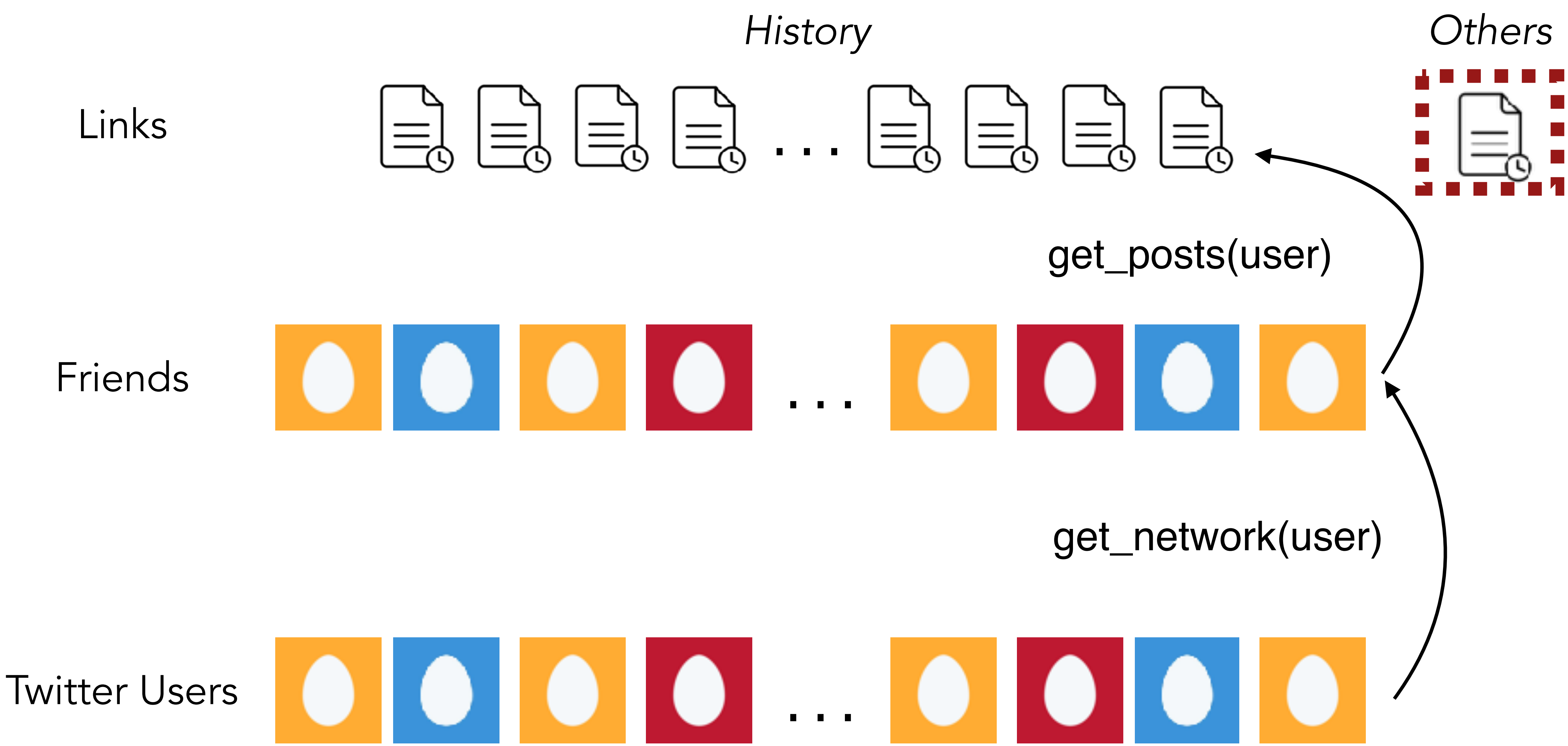
Given three actions:

get_network(user)
get_posts(user)
find_posters(link)

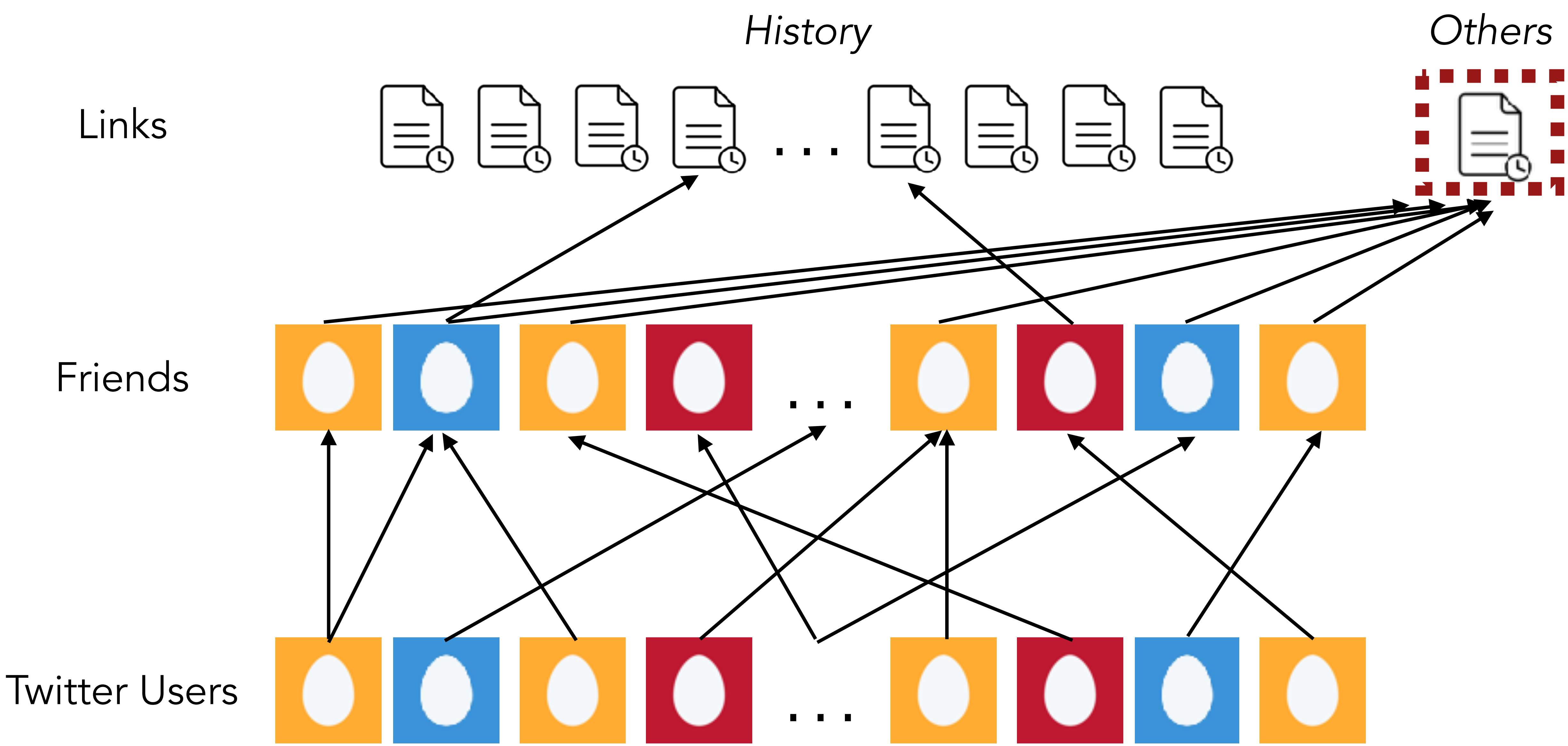
Implementation: Naive



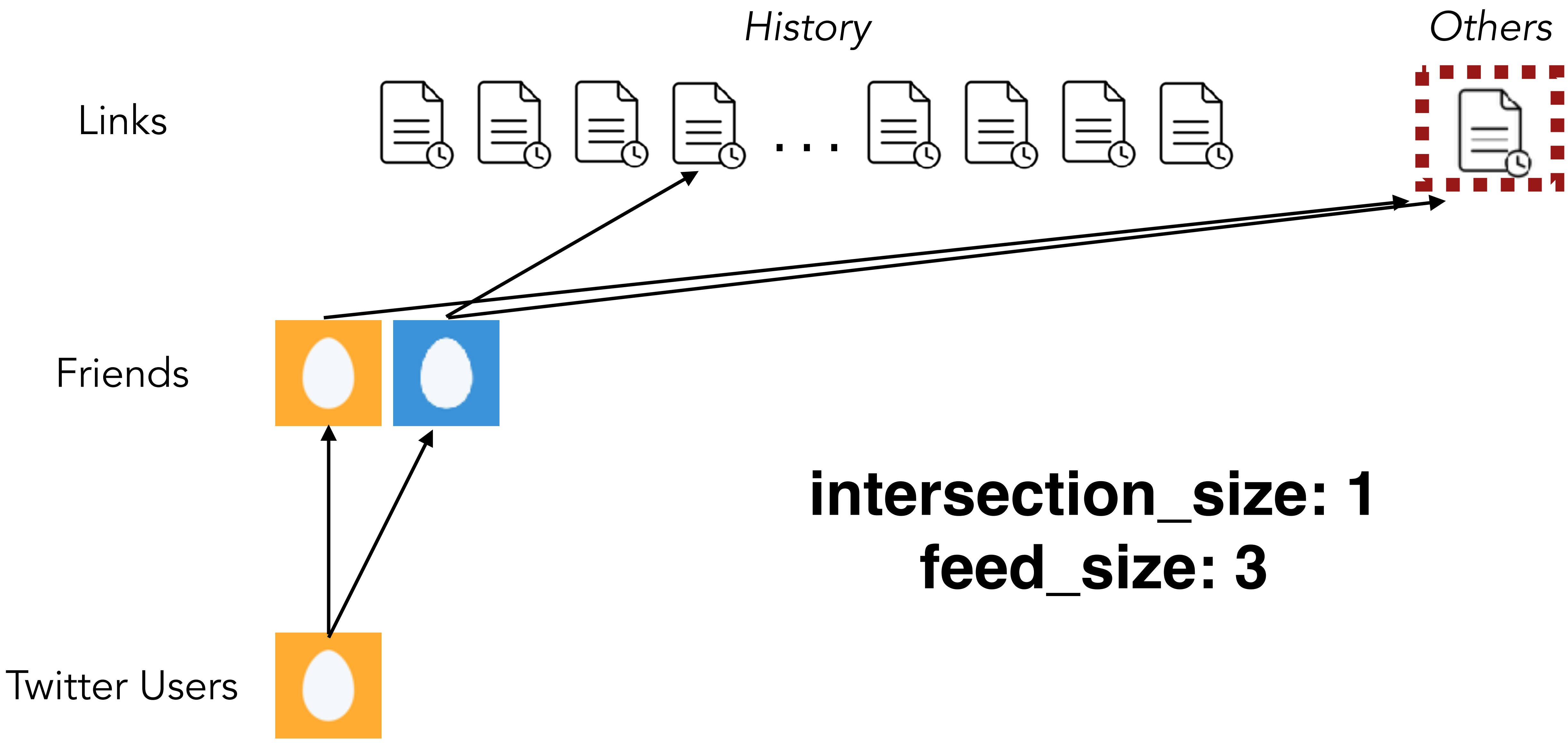
Implementation: Naive



Implementation: Naive



Implementation: Naive

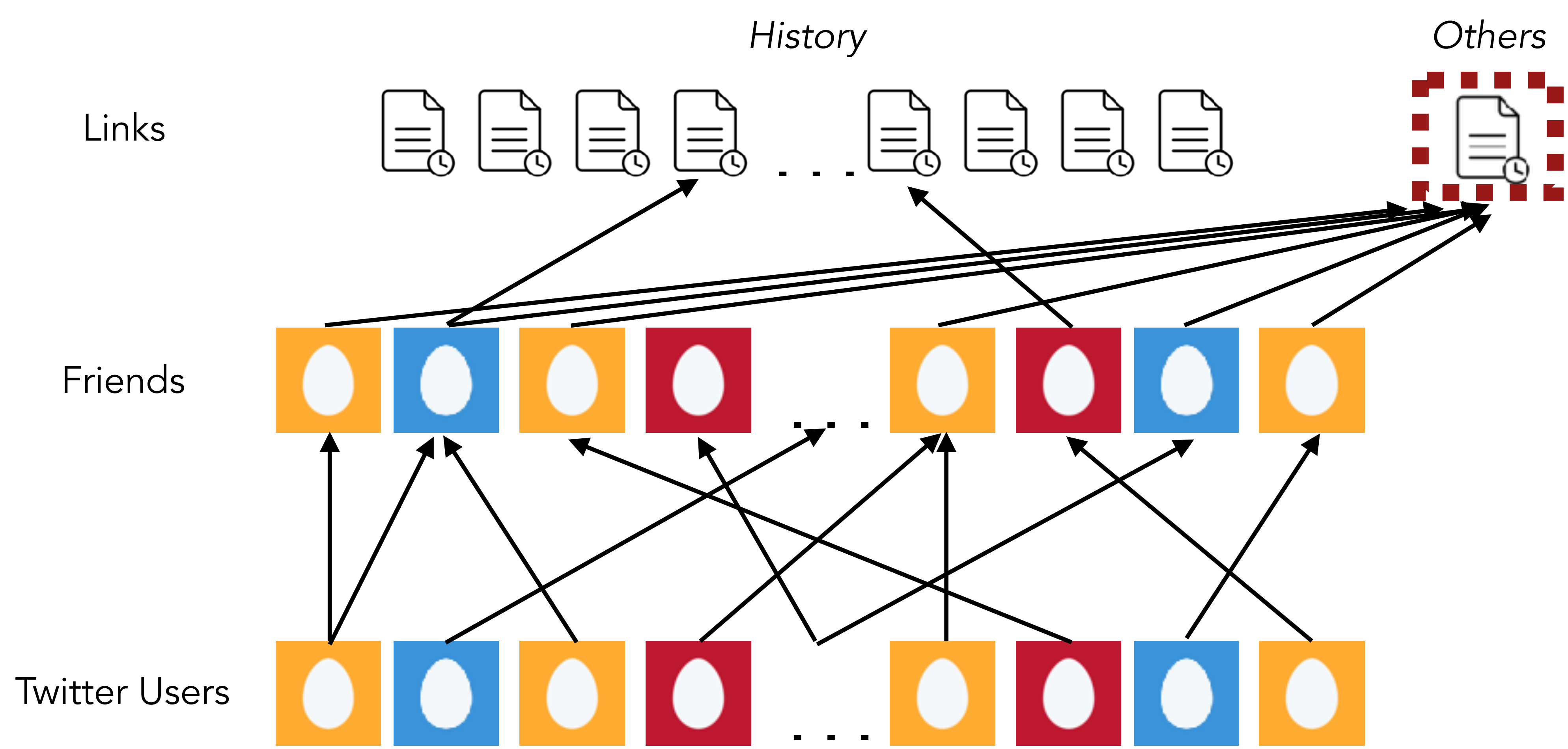


intersection_size: 1
feed_size: 3

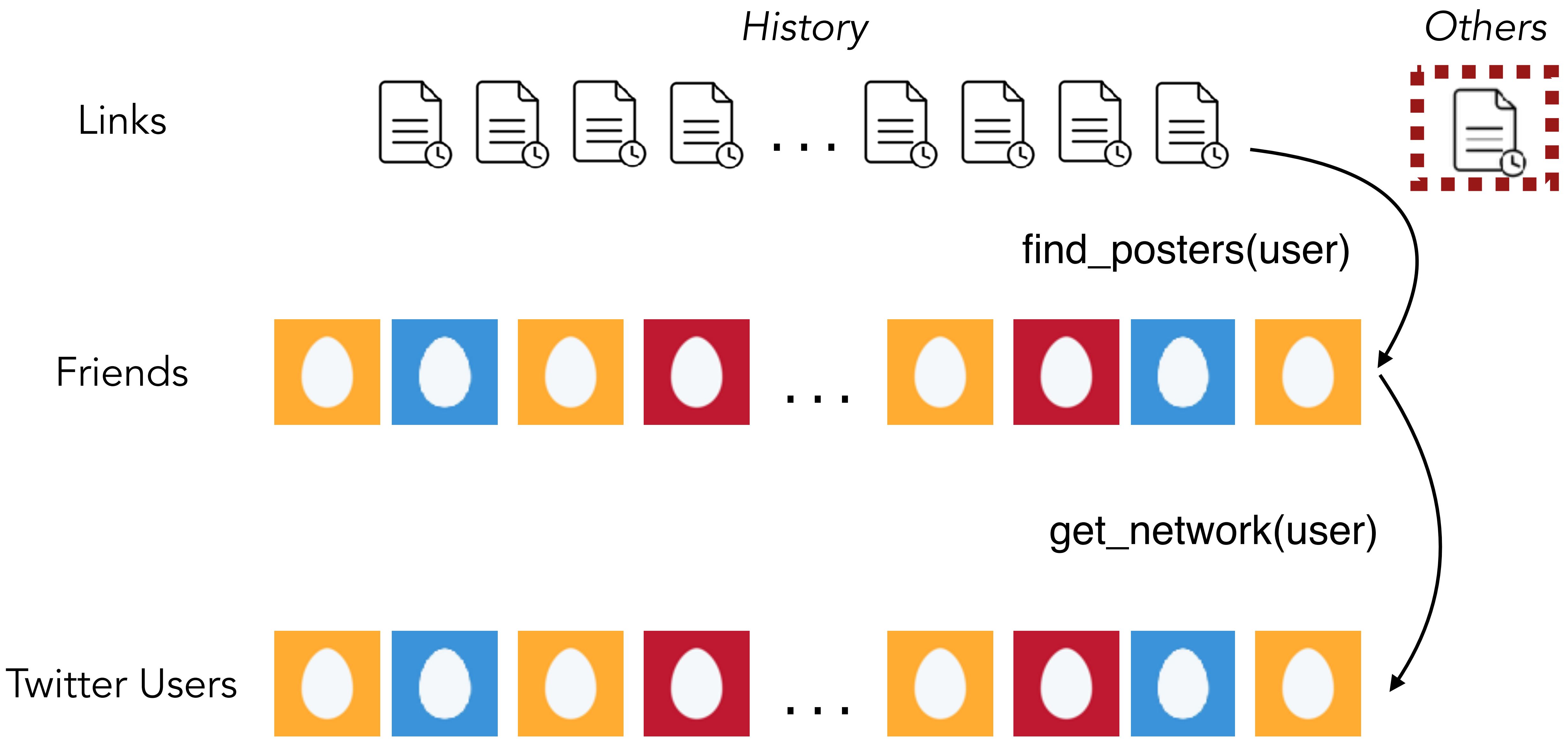
Implementation: Naive

Extremely inefficient because ~500m Twitter users

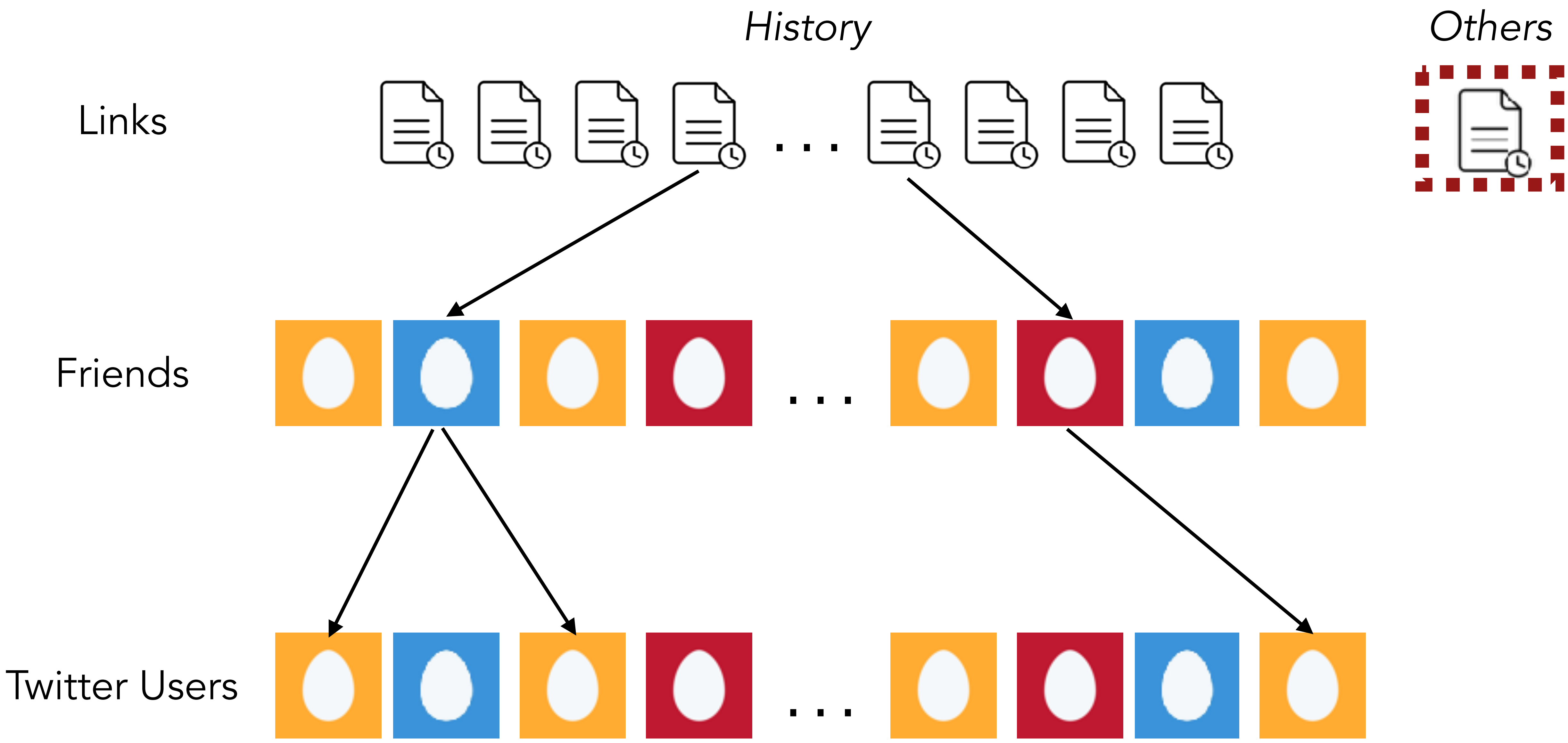
Most users have no intersection, MLE $-\infty$



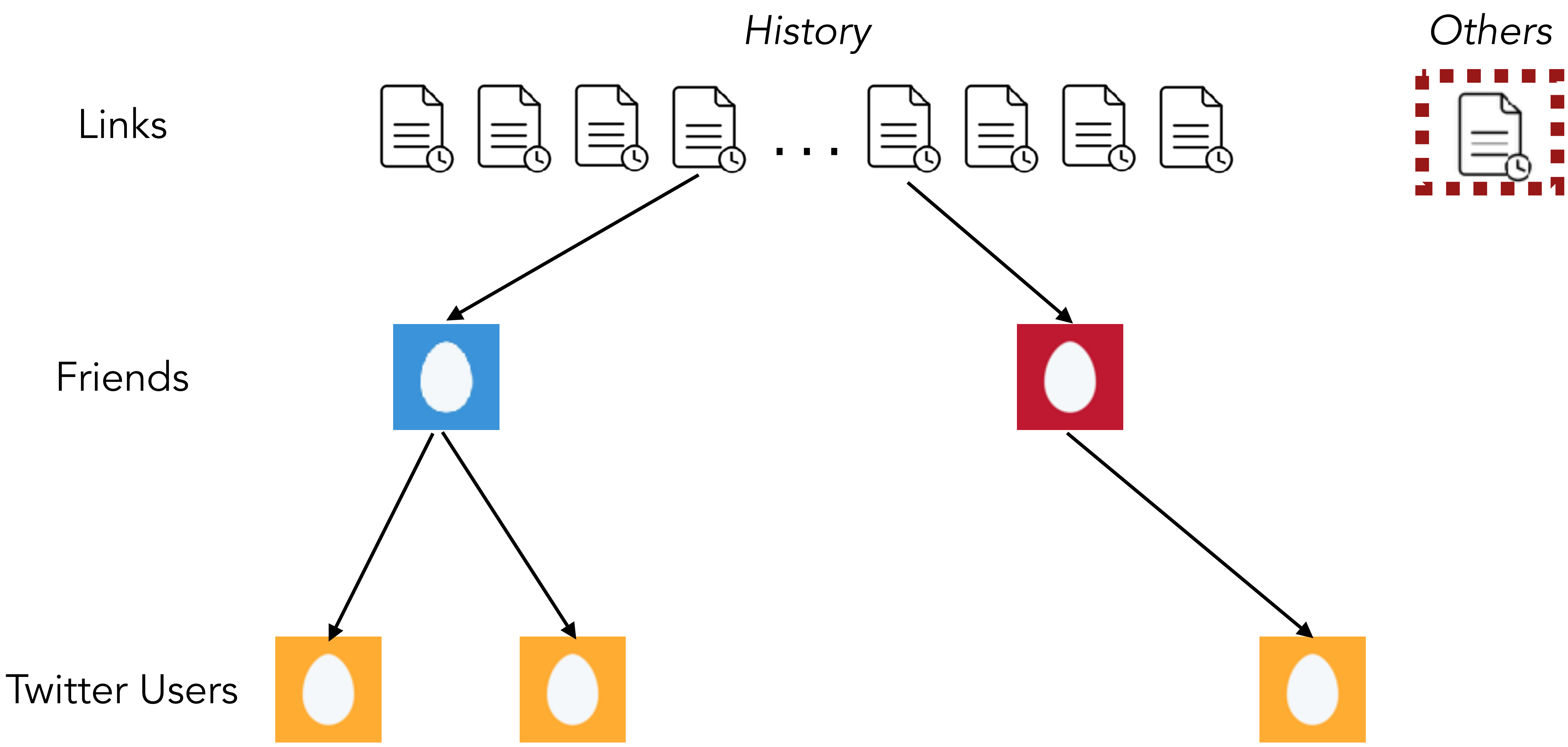
Implementation: Efficient



Implementation: Efficient

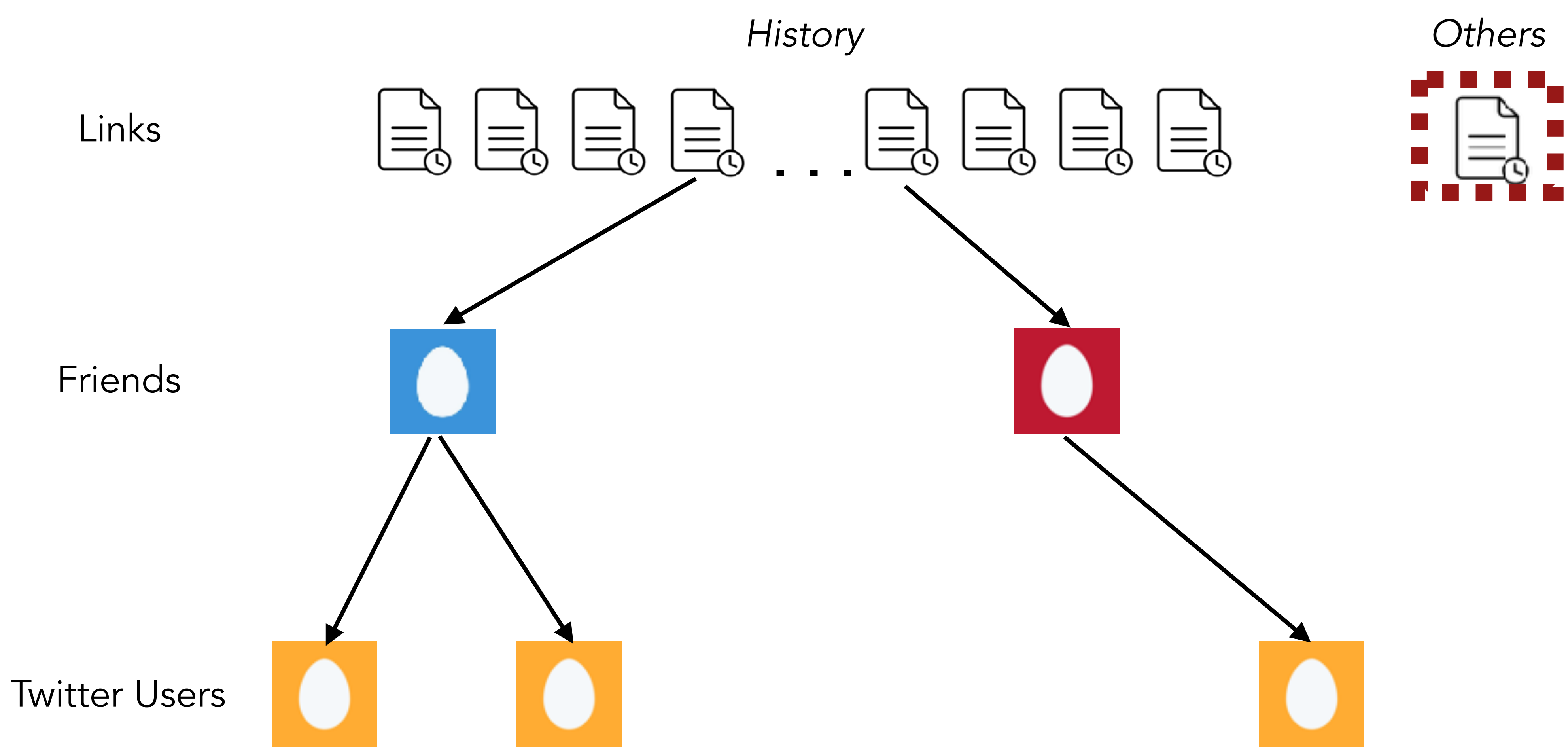


Implementation: Efficient



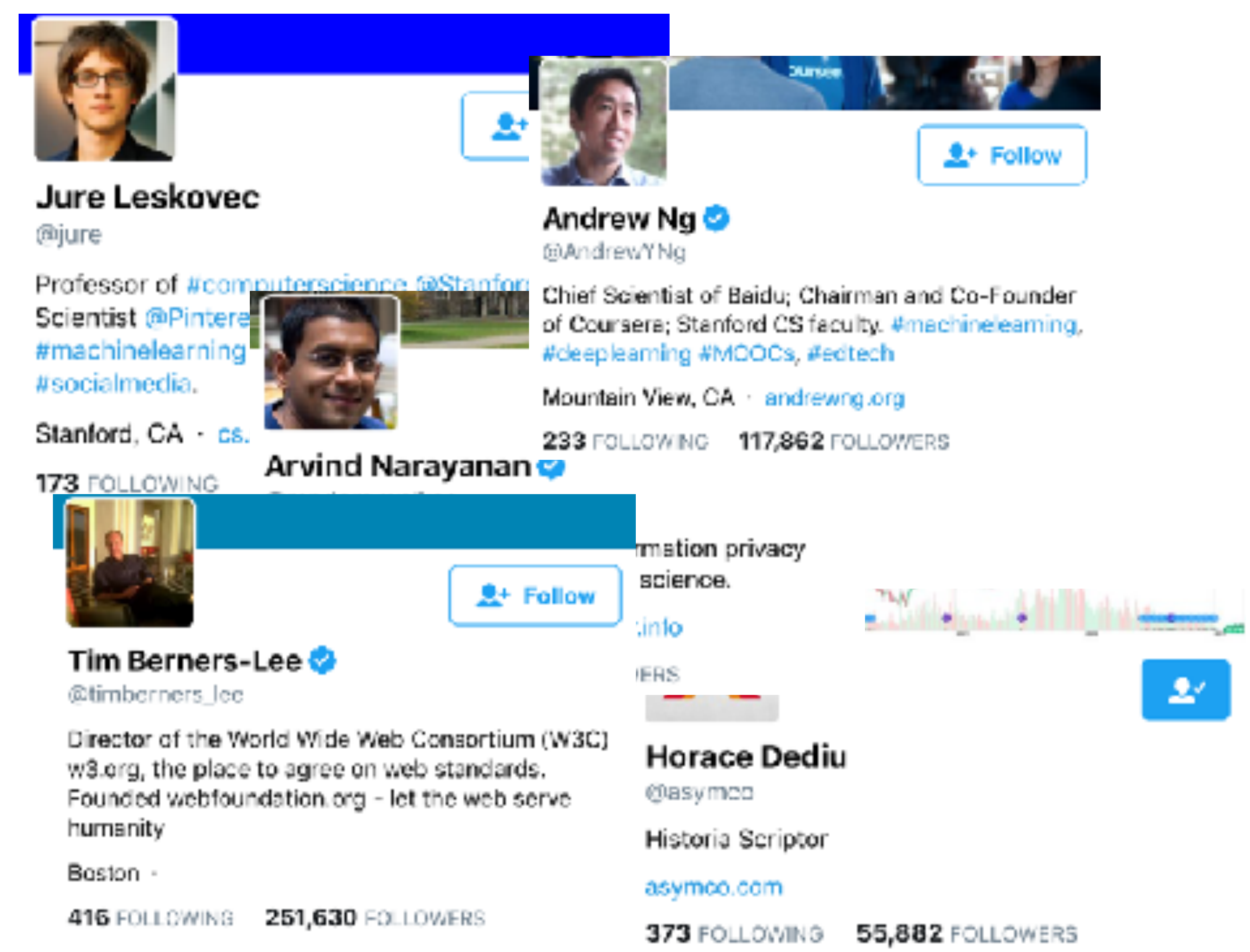
Implementation: Efficient

Lossless: only non-intersecting users are ignored.



Simplifications: get_network

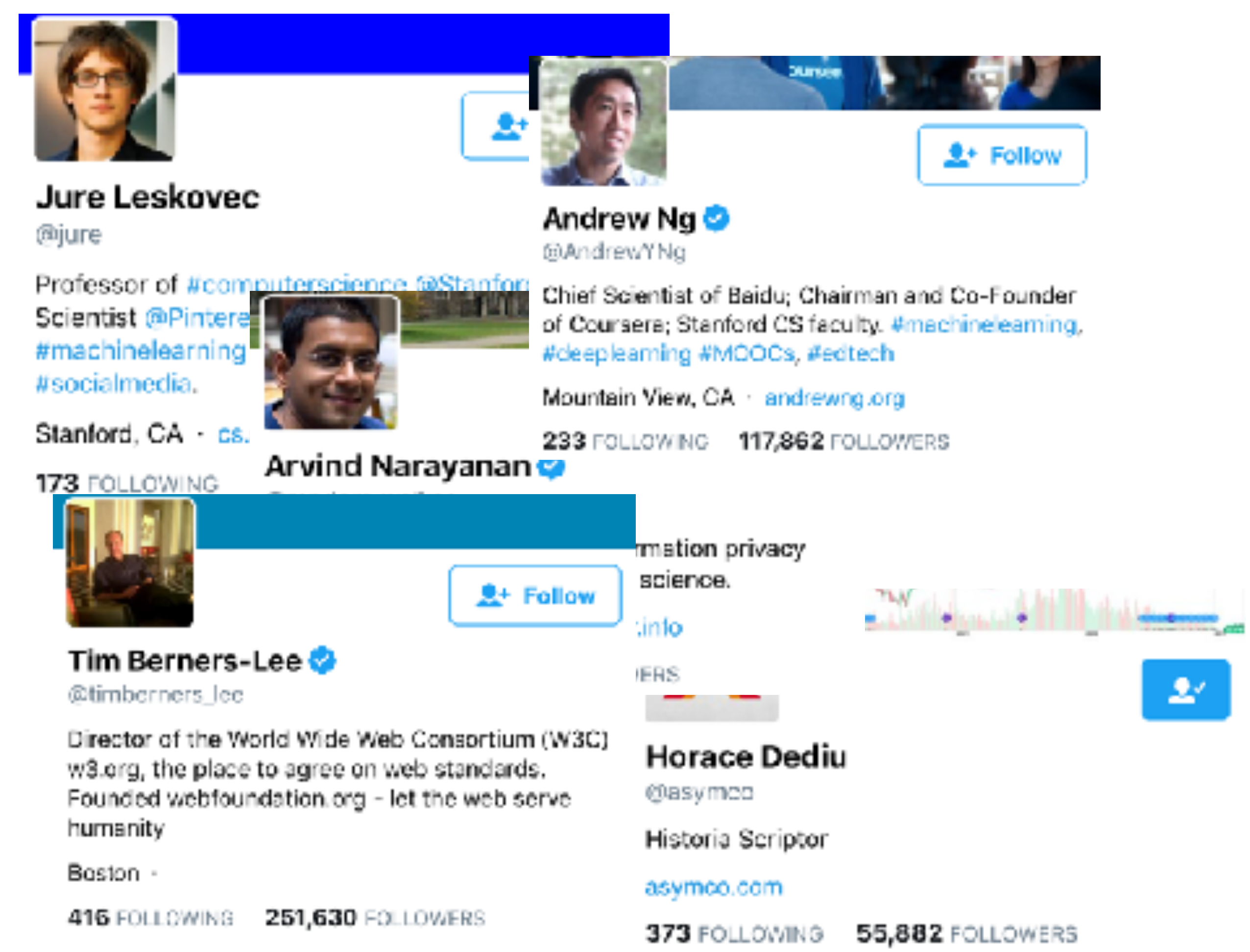
Expensive call if link seen by large network.



Simplifications: get_network

Expensive call if link seen by large network.

Ignore non-informative links.



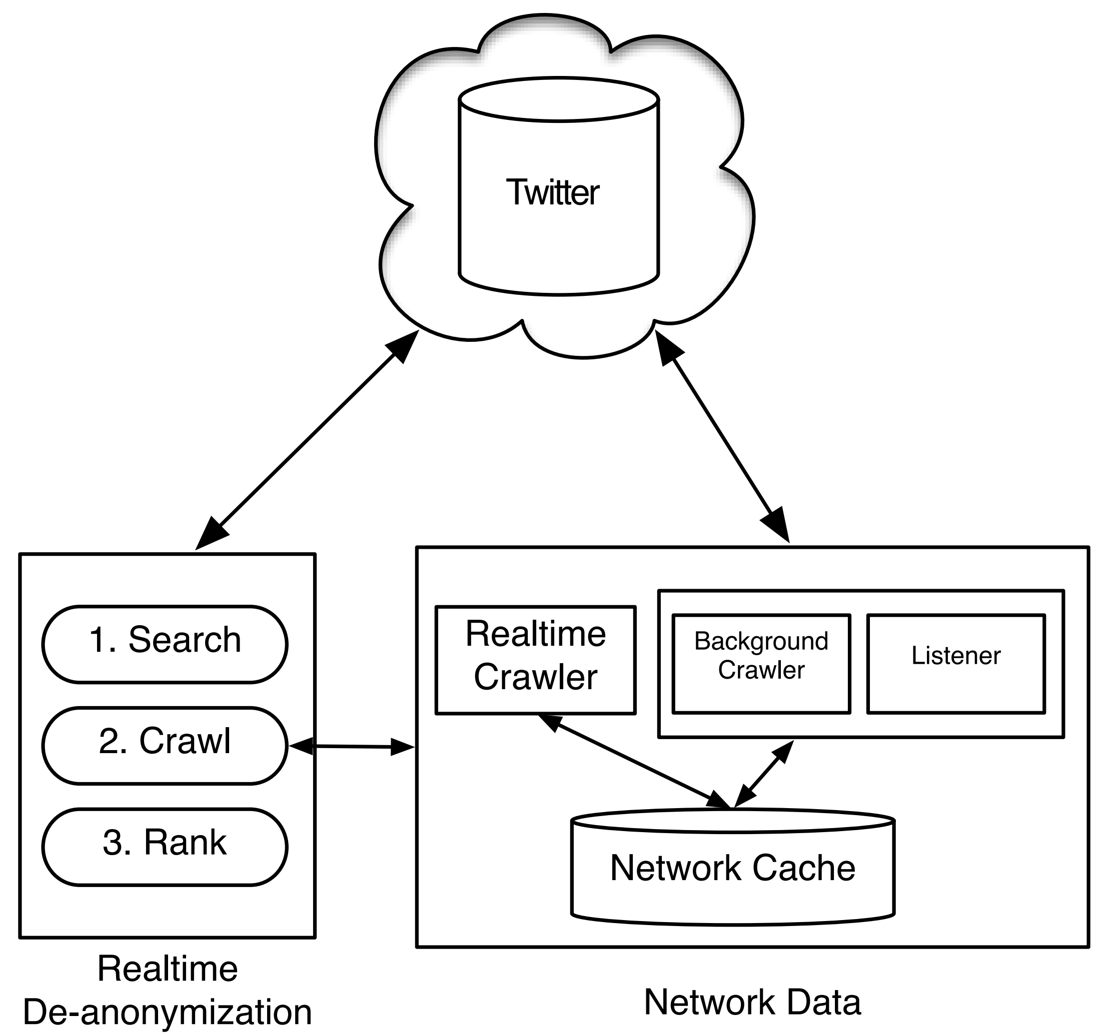
Simplifications: get_network

Still expensive for network size bigger than ~10,000.

Simplifications: get_network

Still expensive for network size bigger than ~10,000.

Background crawl for users with 10,000 - 500,000 followers.



Final Implementation

Ignores expensive, non-informative links.

Final Implementation

Ignores expensive, non-informative links and estimates feed size.

Uses offline crawl database of over 470,000 users.

Final Implementation

Ignores expensive, non-informative links and estimates feed size.

Uses offline crawl database of over 470,000 users.

Runs deanonymization operation in under 30 seconds.

How well does it work?

72%

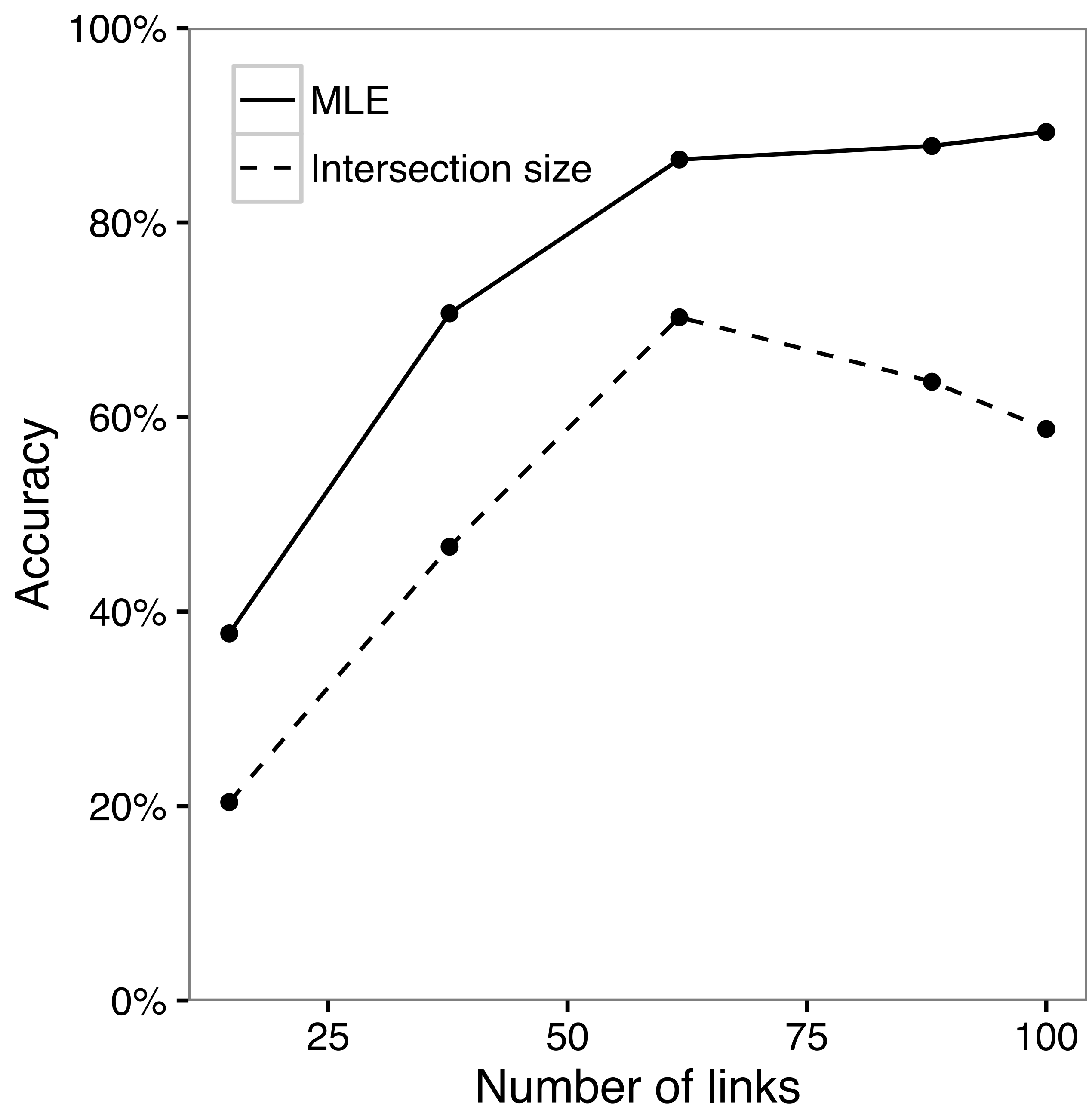
of the 374 users we tried to deanonymize were matched to the correct Twitter account.

81%

were in the Top 15.

IRB #34095

Main result

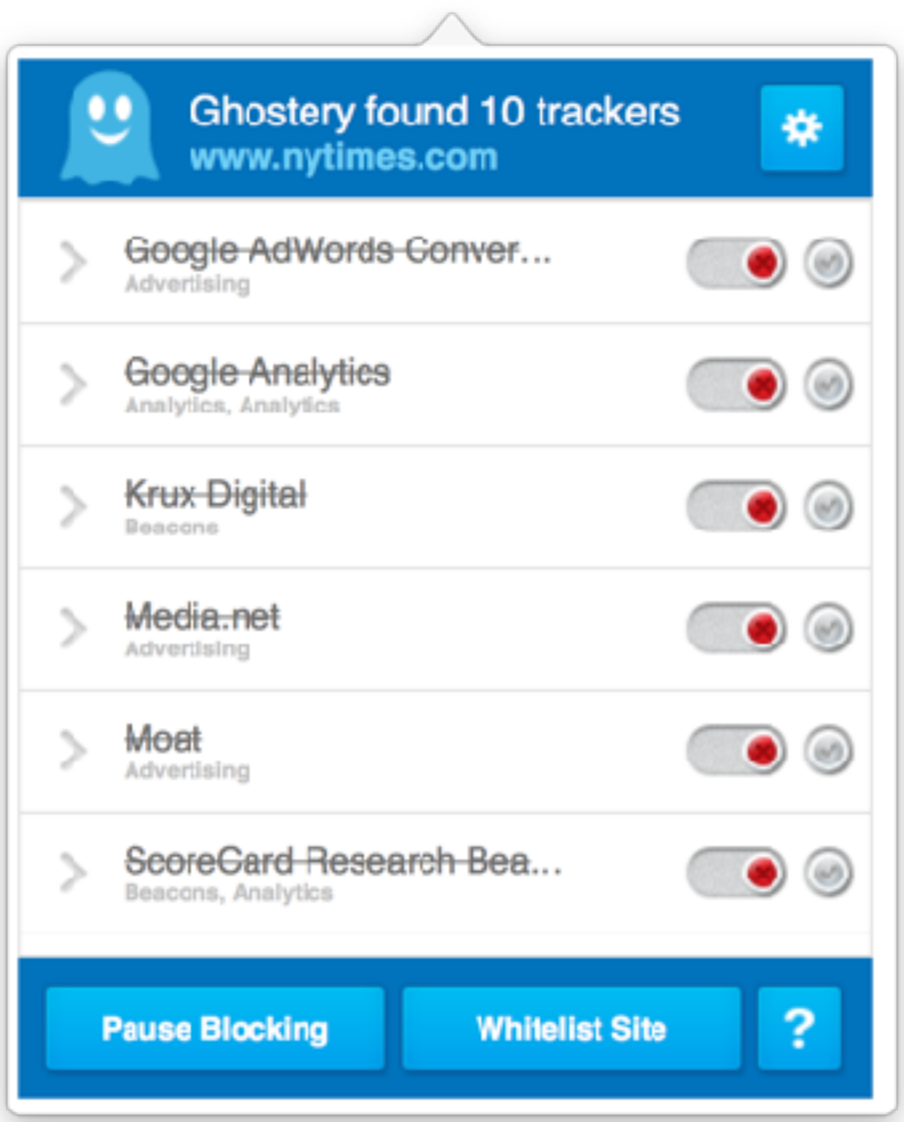


Accuracy increases when there are more URLs in the history

Our approach performs substantially better than baseline

How companies would use this

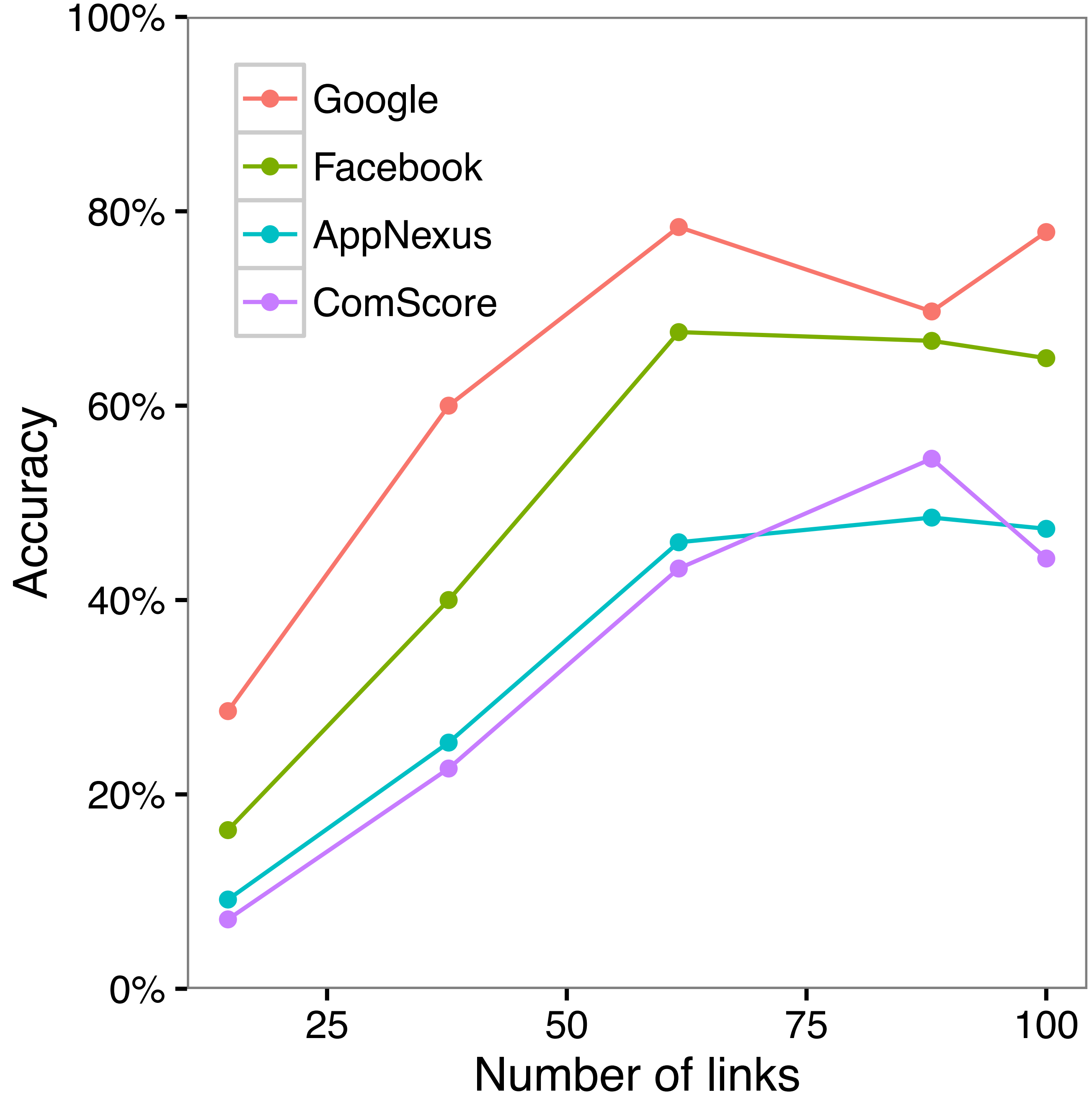
We had complete browsing history,
but companies do not



Companies only see URLs in your browsing history if they have trackers on that page

Retry the attack using only the part of the browsing history that a company has access to

Deanonymization accuracy for 3rd party trackers



Note that companies can collect this data even if you are logged out

Takeaways

Propose and test a successful model to deanonymize browsing data.

Mitigations are limited; attack exploits nature of the network.

Browsing data is sensitive regardless of anonymization.

Thanks for listening

We thank **Twitter** for access to the Gnip search API,
and **Henri Stern** for his help building the online experiment.

Icons created by Costa Galaktionov, Mihail Zhukov, and Sylwia Bartyzel of the Noun Project.

Full form of the MLE equation

The maximum likelihood estimator primarily depends on the size of the feed, and the number of URLs the feed and the browsing history have in common

$$\hat{R} = \underset{R \in C}{\operatorname{argmax}} \left[q_R \log \left(\frac{q_R}{p_R} \right) + (1 - q_R) \log \left(\frac{1 - q_R}{1 - p_R} \right) \right]$$

p_R: feed size

- sum(p_i) for all URLs i in the feed

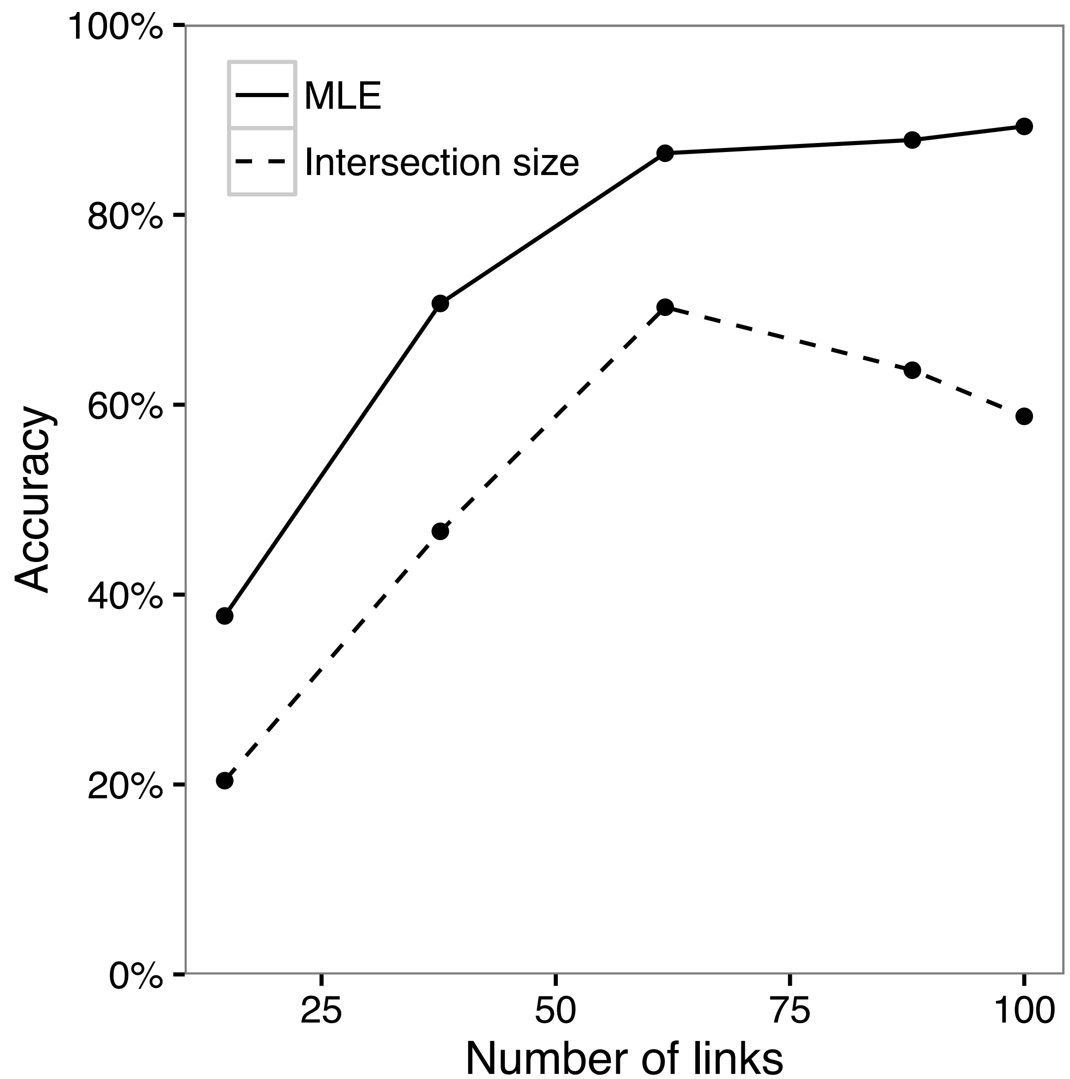
q_R: intersection size

- (fraction of history links that are in the feed)

C: set of candidates

R: the feed ("recommendation set")

FAQ: How well does this study generalize to other populations of Twitter users?



Effectiveness depends on the number of links.

Our main result (shown here) should generalize to the broader Twitter population.

The exact fraction of people who can be deanonymized depends on the history size distribution of the people in the sample. Our sample had a large number of active users.

FAQ: How accurate is the model?

Doesn't Twitter sort tweets by relevance, so that some tweets in your feed get higher priority than others?

Answer: The model doesn't have to be completely true to life, because most of the time, the obscure links give so much signal that crude modeling techniques are enough to reliably capture it.

We expect that a wide range of modeling decisions would have produced similar results.

(e.g. sorting by intersection size / log (number of friends))

FAQ: How did we decide who to reject from our study?

We required users to have visited at least **4** informative links

84% of users passed this filter

92% of users with at least **10** links in their browsing history passed this filter

97% of users with at least **20** links in their browsing history passed this filter

Link informativeness was based on how many people tweeted the link and how many people saw the link in their feed