“I Knew They Clicked When I Saw Them With Their Friends”
Identifying your silent web visitors on social media

Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Columbia University

Oct 2, 2014
Motivation

- Primary use of social network is information sharing
- Typically on twitter, we are involved in a public social conversation
Motivation

- Primary use of social network is information sharing
- Typically on twitter, we are involved in a public social conversation
- Sometimes we want to remain anonymous:
  - Most of us are very careful when we post
  - Most of us are just lurkers
Motivation

- Primary use of social network is information sharing
- Typically on twitter, we are involved in a public social conversation
- Sometimes we want to remain anonymous:
  - Most of us are very careful when we post
  - Most of us are just lurkers
- Are we completely protected?
Related Work

Public conversation + Private browsing behavior

- Lot of work in deanonymizing users based on behavior

---

Related Work

Public conversation + Private browsing behavior

- Lot of work in deanonymizing users based on behavior
  - Inferring users from the (private) Netflix ratings and (public) imdb ratings\(^1\)\(^2\)
  - Inferring user attributes from some (private) known users and the social network graph\(^3\)

---


\(^3\) A. Mislove, B. Viswanath, K. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. *WSDM 2010.*
Related Work

Public conversation + Private browsing behavior

- Lot of work in deanonymizing users based on behavior
  - Inferring users from the (private) Netflix ratings and (public) imdb ratings \(^1\) \(^2\)
  - Inferring user attributes from some (private) known users and the social network graph \(^3\)
- However, many people have no behavior on the network

---


\(^3\) A. Mislove, B. Viswanath, K. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. *WSDM 2010.*
Related Work

Public conversation + Private browsing behavior

- Lot of work in deanonymizing users based on behavior
  - Inferring users from the (private) Netflix ratings and (public) imdb ratings \(^1\) \(^2\)
  - Inferring user attributes from some (private) known users and the social network graph \(^3\)
- However, many people have no behavior on the network
  - e.g.: 77% of users receiving tweets about the nytimes.com tweets had at most 1 post

---


\(^3\) A. Mislove, B. Viswanath, K. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. *WSDM 2010.*
Related Work

Public conversation + Private browsing behavior

- Lot of work in deanonymizing users based on behavior
  - Inferring users from the (private) Netflix ratings and (public) imdb ratings
  - Inferring user attributes from some (private) known users and the social network graph
- However, many people have no behavior on the network
  - e.g.: 77% of users receiving tweets about the nytimes.com tweets had at most 1 post
- Can we de-anonymize these people?

---

Our Work

Public conversation + Private browsing behavior

Differences from other work:
Public conversation + Private browsing behavior

Differences from other work:

- In our framework, users are silent
Our Work

Public conversation + Private browsing behavior

Differences from other work:

- In our framework, users are silent
- We would like a generic method that applies broadly to any domain
  - no assumptions in addition to user browsing
Our Work

Public conversation + Private browsing behavior

Differences from other work:

- In our framework, users are silent
- We would like a generic method that applies broadly to any domain
  - no assumptions in addition to user browsing
- We don’t assume that the user has an explicit account with a 3rd party
Problem Definition

- Assumptions:
Problem Definition

- Assumptions:
  - There is a single local persistent id maintained across a domain
    (e.g.: fingerprinting /cookies)

Introduction

How?
Correctness
Who is useful?
Conclusion
Faster Inference
Problem Definition

- Assumptions:
  - There is a single local persistent id maintained across a domain (e.g.: fingerprinting 4/cookies)
  - The public social conversation can be monitored
Problem Definition

Assumptions:

- There is a single local persistent id maintained across a domain (e.g.: fingerprinting /cookies)
- The public social conversation can be monitored 🤖
- Access to information in the http request referrer field

Problem Definition

- Assumptions:
  - There is a single local persistent id maintained across a domain
    (e.g.: fingerprinting /cookies)
  - The public social conversation can be monitored 🦅
  - Access to information in the http request referrer field
    - e.g.: NYT treats users differently based on their origin
Problem Definition

- Assumptions:
  - There is a single local persistent id maintained across a domain
    (e.g.: fingerprinting /cookies)
  - The public social conversation can be monitored
  - Access to information in the http request referrer field
    - e.g.: NYT treats users differently based on their origin

Non-assumptions:
- Not all users participate
- Attacker is controlling only a single domain
Problem Definition

- Assumptions:
  - There is a single persistent id maintained (e.g.: fingerprinting \(^5\)/cookies)
  - The public social conversation is monitored
  - Access to information in the http request referrer field
  - e.g.: NYT treats users differently based on their origin

Non-assumptions:
- Not all users participate
- Attacker is controlling only a single domain

---

Questions we asked

- Given knowledge of anonymous users who clicked links and the social network of those users, how can we re-identify the anonymous users?
Questions we asked

- Given knowledge of anonymous users who clicked links and the social network of those users, how can we re-identify the anonymous users?
- What are the factors affecting inference (such as click rates)?
Questions we asked

- Given knowledge of anonymous users who clicked links and the social network of those users, how can we re-identify the anonymous users?
- What are the factors affecting inference (such as click rates)?
- Can we make identification more robust to these factors?
Questions we asked

- Given knowledge of anonymous users who clicked links and the social network of those users, how can we re-identify the anonymous users?
- What are the factors affecting inference (such as click rates)?
- Can we make identification more robust to these factors?
- Who is useful in re-identification?
Dataset

- Datasets should be more than the gardenhose of tweets

---


Dataset

- Datasets should be more than the gardenhose of tweets
- Users posting domain links

---

6 A. May, A. Chaintreau, N. Korula, and S. Lattanzi. Filter & Follow: How Social Media Foster Content Curation. *SIGMETRICS 2014*

Dataset

- Datasets should be more than the gardenhose of tweets
  - Users posting domain links
  - AND their followers

---

6 A. May, A. Chaintreau, N. Korula, and S. Lattanzi. Filter & Follow: How Social Media Foster Content Curation. *SIGMETRICS 2014*

Identifying silent visitors
Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction

How?
Correctness
Who is useful?

Conclusion

Faster Inference

Dataset

Datasets should be more than the gardenhose of tweets

Users posting domain links

NYT: Twitter posts containing URLs to nytimes.com
AND their followers
for 1 week in Dec 2012

346k unique users
70k unique URLs


Dataset

- Datasets should be more than the gardenhose of tweets
  - Users posting domain links
  - AND their followers
- NYT: Twitter posts containing URLs to nytimes.com for 1 week in Dec 2012
  - 346k unique users
  - 70k unique URLs
- DIGG: Twitter posts containing URLs to digg.com over July 2009.
  - 44k unique users
  - 216k unique URLs

---

6 A. May, A. Chaintreau, N. Korula, and S. Lattanzi. Filter & Follow: How Social Media Foster Content Curation. *SIGMETRICS 2014*

Dataset

- Datasets should be more than the gardenhose of tweets
  - Users posting domain links
  - AND their followers
- NYT: Twitter posts containing URLs to nytimes.com for 1 week in Dec 2012
  - 346k unique users
  - 70k unique URLs
- DIGG: Twitter posts containing URLs to digg.com over July 2009.
  - 44k unique users
  - 216k unique URLs
- KAIST: Twitter posts over July 2009
  - 8m unique users
  - 37m unique URLs

---

6 A. May, A. Chaintreau, N. Korula, and S. Lattanzi. Filter & Follow: How Social Media Foster Content Curation. *SIGMETRICS 2014*

Method

nytimes.com

nyti.ms/1pJ06oL
Top Colleges That Enroll Rich, Middle Class and Poor

Social network
Method

Identifying silent visitors

Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction

How?
Correctness
Who is useful?
Conclusion
Faster Inference

nytimes.com

nyti.ms/1pJ06oL
Top Colleges That Enroll Rich, Middle Class and Poor

Social network
Identifying silent visitors

Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction

How?

Correctness

Who is useful?

Conclusion

Faster Inference

**Method**

nytimes.com

nyti.ms/1pJ060L

Top Colleges That Enroll Rich, Middle Class and Poor

Social network
Method

Identifying silent visitors

Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction

How?

Correctness

Who is useful?

Conclusion

Faster Inference
Method

Identifying silent visitors

Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction

How?

Correctness

Who is useful?

Conclusion

Faster Inference

nytimes.com

nyti.ms/1pJ06oL
Top Colleges That Enroll Rich, Middle Class and Poor
Method

Identifying silent visitors

Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction

How?

Correctness

Who is useful?

Conclusion

Faster Inference

nytimes.com

nyti.ms/1pJ06oL
Top Colleges That Enroll Rich, Middle Class and Poor

nyti.ms/1qMfdIn
Steve Jobs Was a Low-Tech Parent
Method

Identifying silent visitors
Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction
How?
Correctness
Who is useful?
Conclusion
Faster Inference

nytimes.com
nyti.ms/1pJ06oL
Top Colleges That Enroll Rich, Middle Class and Poor

nyti.ms/1qMfdIN
Steve Jobs Was a Low-Tech Parent
Method

Identifying silent visitors

Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction
How?
Correctness
Who is useful?
Conclusion
Faster Inference

nytimes.com

nyti.ms/1pJ06oL
Top Colleges That Enroll Rich, Middle Class and Poor

nyti.ms/1qMfdln
Steve Jobs Was a Low-Tech Parent

Social network
Method

Identifying silent visitors

Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction

How?

Correctness

Who is useful?

Conclusion

Faster Inference

nytimes.com

nyti.ms/1pJ06oL
Top Colleges That Enroll Rich, Middle Class and Poor

nyti.ms/1qMfdln
Steve Jobs Was a Low-Tech Parent
Method

Identifying silent visitors
Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction

How?
Correctness
Who is useful?
Conclusion
Faster Inference

nytimes.com
nyti.ms/1pJ06oL
Top Colleges That Enroll Rich, Middle Class and Poor

nyti.ms/1qMfdIn
Steve Jobs Was a Low-Tech Parent

Social network
Method

Identifying silent visitors
Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction
How?
Correctness
Who is useful?
Conclusion
Faster Inference

nytimes.com
nyti.ms/1pJ06oL
Top Colleges That Enroll Rich, Middle Class and Poor

nyti.ms/1qMfdIn
Steve Jobs Was a Low-Tech Parent
Method

Identifying silent visitors
Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction
How? Correctness
Who is useful?
Conclusion Faster Inference

nytimes.com
nyti.ms/1pJ06oL
Top Colleges That Enroll Rich, Middle Class and Poor

nytimes.com
nyti.ms/1qMfdln
Steve Jobs Was a Low-Tech Parent

nytimes.com
nyti.ms/1oj85cU
Fashion Week Now
Method

Indentifying silent visitors
Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction

Who is useful?

How?

Correctness

Conclusion

Faster Inference

nytimes.com

nyti.ms/1pJ06oL
Top Colleges That Enroll Rich, Middle Class and Poor

nyti.ms/1qMfdln
Steve Jobs Was a Low-Tech Parent

nyti.ms/1oj85cU
Fashion Week Now

Social network
Method

Identifying silent visitors
Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction
How?
Correctness
Who is useful?
Conclusion
Faster Inference

nytimes.com
nyti.ms/1pJ06oL
Top Colleges That Enroll Rich, Middle Class and Poor

nyti.ms/1qMfdIn
Steve Jobs Was a Low-Tech Parent

nyti.ms/1oj85cU
Fashion Week Now
Method

Identifying silent visitors
Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction
How?
Correctness
Who is useful?
Conclusion
Faster Inference

nytimes.com

nyti.ms/1pJ06OL
Top Colleges That Enroll Rich, Middle Class and Poor

nyti.ms/1qMfdln
Steve Jobs Was a Low-Tech Parent

nyti.ms/1oj85cU
Fashion Week Now

Social network
Identifying silent visitors

Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction

How?

Correctness

Who is useful?

Conclusion

Faster Inference

Method

nytimes.com

nyti.ms/1pJ06oL
Top Colleges That Enroll Rich, Middle Class and Poor

nyti.ms/1qMfdIn
Steve Jobs Was a Low-Tech Parent

nyti.ms/1oj85cU
Fashion Week Now
Potential reasons why the method won’t work

- Most URLs are posted by a lot of people
- URLs responsible for a lot of traffic are seen by a lot of people
- Some people receive a lot of URLs
- Is there sufficient diversity of links?
## Results: Re-identification using URLs

Does our method work?

<table>
<thead>
<tr>
<th>% visitors</th>
<th>CTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>53.7%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td>5%</td>
</tr>
<tr>
<td></td>
<td>1%</td>
</tr>
</tbody>
</table>
## Results: Re-identification using URLs

- How much traffic can we identify?

<table>
<thead>
<tr>
<th></th>
<th>CTR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>% visitors</td>
<td>53.7%</td>
</tr>
<tr>
<td>% traffic</td>
<td>91.0%</td>
</tr>
</tbody>
</table>
Results: Re-identification using URLs

How many users can we uniquely identify?

![Graph showing CCDF and fraction of people uniquely identifiable against number of URLs received. The graph indicates that on average, 10-20 URLs are needed to identify half the users.]
Results: Re-identification using URLs

- How many users can we uniquely identify?

- On average 10-20 URLs to identify half the users
Results: Re-identification using URLs

What factors affect identification?

<table>
<thead>
<tr>
<th>% visitors</th>
<th>CTR 100%</th>
<th>CTR 30%</th>
<th>CTR 5%</th>
<th>CTR 1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>53.7%</td>
<td>91.0%</td>
<td>79.2%</td>
<td>43.8%</td>
<td>11.8%</td>
</tr>
<tr>
<td>30.5%</td>
<td>79.2%</td>
<td>43.8%</td>
<td>11.8%</td>
<td></td>
</tr>
<tr>
<td>7.9%</td>
<td>43.8%</td>
<td>11.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8%</td>
<td>11.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Results: Re-identification using URLs

- Can we improve results when we know who exactly you received a URL from (attribution)?

<table>
<thead>
<tr>
<th></th>
<th>No Attribution</th>
<th>Attribution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>% visitors</strong></td>
<td>53.7%</td>
<td>69.1%</td>
</tr>
<tr>
<td><strong>% traffic</strong></td>
<td>91.0%</td>
<td>97.0%</td>
</tr>
<tr>
<td><strong>CTR</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100%</td>
<td>30.5%</td>
<td>49.3%</td>
</tr>
<tr>
<td>30%</td>
<td>79.2%</td>
<td>91.9%</td>
</tr>
<tr>
<td>5%</td>
<td>43.8%</td>
<td>70.3%</td>
</tr>
<tr>
<td>1%</td>
<td>11.8%</td>
<td>31.7%</td>
</tr>
</tbody>
</table>

**SUMMARY:** We need 10-15 clicks to identify a user
Who are useful in re-identification?

- Not much difference between the individuals used for identification and those who were not helpful for identification.

![Bar chart showing the number of followers of users used to identify you, with two types: useful and useless.](chart.png)
Conclusion

- Public social conversation identifies us very broadly
  - Even when we aren’t active on the social network
- The only way to really prevent this is to take drastic measures
- What mechanisms can address this type of privacy attack?
Conclusion

- Public social conversation identifies us very broadly
  - Even when we aren’t active on the social network
- The only way to really prevent this is to take drastic measures
- What mechanisms can address this type of privacy attack?

Any Questions?
Re-identification using URLs

Identifying silent visitors

Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction

How?

Correctness

Who is useful?

Conclusion

Faster Inference

CCDF

Dataset

NYT
Re-identification using URLs

Dataset

NYT

Fraction of people uniquely identifiable

Number of URLs Received

CCDF

Number of URLs Received

Dataset

NYT
Re-identification using URLs

Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau

Introduction

How?

Correctness

Who is useful?

Conclusion

Faster Inference

![CCDF Graph]

- CCDF
- Dataset (DIGG, NYT)
- Number of URLs Received
- Fraction of people uniquely identifiable

0.00 0.25 0.50 0.75 1.00

1 2 3 5 10 20 50 100 500 1000 5000

Number of URLs Received
Re-identification using URLs

CCDF

Fraction of people uniquely identifiable

Number of URLs Received

Dataset
DIGG
KAIST
NYT

Introduction
How?
Correctness
Who is useful?
Conclusion
Faster Inference
Can we do better?

- Comparing two methods
  - With URLs
  - With Attributed URLs
Can we do better?

- Comparing two methods
  - With URLs
  - With Attributed URLs

```
<table>
<thead>
<tr>
<th>Number of urls received</th>
<th>Fraction of people uniquely identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>CCOF</td>
</tr>
<tr>
<td>1</td>
<td>CTR 100%</td>
</tr>
</tbody>
</table>
```

Method for identification
- Attributed URLs
- URLs
Can we do better?

- Comparing two methods
  - With URLs
  - With Attributed URLs
Can we do better?

- Comparing two methods
  - With URLs
  - With Attributed URLs

![Graph showing comparison between URLs and Attributed URLs](image)

**CCDF**

<table>
<thead>
<tr>
<th>Number of urls received</th>
<th>Fraction of people uniquely identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>0.50</td>
</tr>
<tr>
<td>5</td>
<td>0.75</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**CTR**

- 100%
- 5%
- 1%

**Method for identification**

- Attributed URLs
- URLs