Identifying Personal Information in Internet Traffic

Yabing Liu[†] Han Hee Song[‡] Ignacio Bermudez[§]
Alan Mislove[†] Mario Baldi[‡] Alok Tongaonkar[§]

[†]Northeastern University
[‡]Cisco Systems
[§]Symantec Corporation

November 2, 2015, COSN'15

Web-based services

Most popular Internet-based services

- Web sites, smartphone apps
- Traditional PCs, tablets, and smartphones
- Facebook (1.44 B) WhatApp (800 M)

Users share significant data explicitly

- Name, gender, email, locations…
- Photos, videos, blogs, news, statuses…

Applications collect user data implicitly

- Monetizing personal information (third parties)

Web-based services

Users don’t have control

- Cannot keep content secret from provider
- Little visibility into what apps do with PI

Organizations concerned about their user privacy

- Companies, universities, …
- Alert users about potential leak

Goal: Important to understand PI transmitted

- Develop system which can automatically detect it

Personal Information

Definition of PI

- Anything the web site or app can receive about the user

Users today have many types of PI

- Name, birthday, income, interests, user ID, …
- Photos, videos, statuses, …

Focus: certain types of text-based PI

Motivating Experiment

Controlled Lab traffic in Aug. 2014

- Set up web/HTTPS-MITM proxy
- Configured iPhone to use the proxy
- Downloaded and ran top 35 free apps from the App Store
- Examined network traces (only HTTP/HTTPS)

PI in App Traffic

What is the fraction of HTTP VS. HTTPS flows?

- 62% HTTP VS. 38% HTTPS

What applications are collecting user PI?

- All of them!
- Examples: Email, Name, UserID, Location, Gender, …

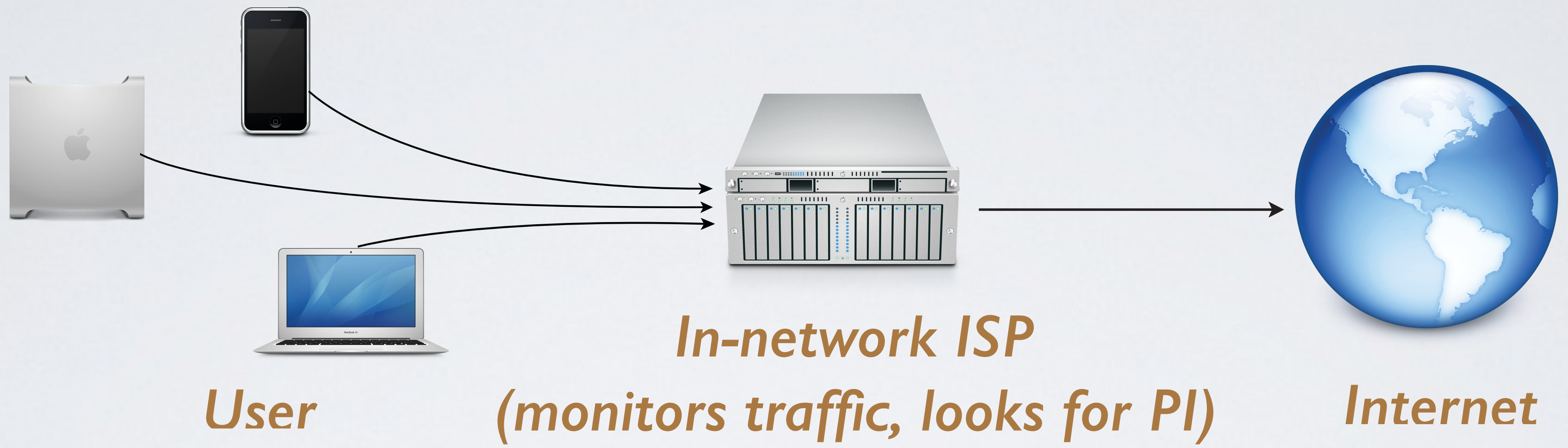
What fraction of flows have PI?

- 3%

Upshot: Lots of PI, but needle in a haystack

Goal

Automatically detect when web sites or smartphone apps collect PI



Explore in-network measurement and analysis

- Large organizations who control the network
- Not end-host-based approach (e.g., devices, browsers)
- Only HTTP transactions (44% of ground truth PI from Lab traffic)

Reasons

- Significantly lower barriers to deployment
- Higher coverage than end-host-based approach

Outline

- ~~Motivation~~
- Dataset
- Methodology
- Evaluation

Dataset

Real ISP operational traffic

- 24 hour PCAP data [Aug. 2011, one European City]
- 13K users without ground truth
- To test methodologies at scale

Dataset	HTTP flows
ISP traffic	40,775,119

Locate the flows with PI

Domain-Keys

Deconstruct fields from HTTP traffic trace

- Key — HTTP GET request, Referrer header, Cookie
- Domain — Host header
- <Domain, Key> (DK) - Value pairs

Observed HTTP transaction

GET /foo.html?user_firstname=Alice HTTP/1.1
Host: imagevenue.com
Cookie: a=293&g=00s9229daa&age=39&id=27
ETag: 2039-2dc90ea2-12
Referer: http://www.facebook.com/?user_id=89
Accept-Encoding: deflate,gzip

HTTP/1.1 200 OK
Date: Mon, 23, May 2013 22:38:34 GMT

Domain-Keys

Deconstruct fields from HTTP traffic trace

- Key — HTTP GET request, Referrer header, Cookie
- Domain — Host header
- <Domain, Key> (DK) - Value pairs

Observed HTTP transaction

GET /foo.html?user_firstname=Alice HTTP/1.1
Host: imagevenue.com
Cookie: a=293&g=00s9229daa&age=39&id=27
ETag: 2039-2dc90ea2-12
Referer: http://www.facebook.com/?user_id=89
Accept-Encoding: deflate,gzip

HTTP/1.1 200 OK
Date: Mon, 23, May 2013 22:38:34 GMT

Derived domain-keys and values

Domain	Key	Field	Value
imagevenue.com	user_firstname	GET	*Alice*
imagevenue.com	a	Cookie	*293*
imagevenue.com	g	Cookie	*00s9229da*
imagevenue.com	age	Cookie	*39*
imagevenue.com	id	Cookie	*27*
imagevenue.com	user_id	Referer	*89*

Domain-Keys

Deconstruct fields from HTTP traffic trace

- Key — HTTP GET request, Referrer header, Cookie
- Domain — Host header
- <Domain, Key> (DK) - Value pairs

Tuples	Domain-keys
51,368,712	3,113,696

Observed HTTP transaction

GET /foo.html?user_firstname=Alice HTTP/1.1
Host: imagevenue.com
Cookie: a=293&g=00s9229daa&age=39&id=27
ETag: 2039-2dc90ea2-12
Referer: http://www.facebook.com/?user_id=89
Accept-Encoding: deflate,gzip

HTTP/1.1 200 OK
Date: Mon, 23, May 2013 22:38:34 GMT

Derived domain-keys and values

Domain	Key	Field	Value
imagevenue.com	user_firstname	GET	*Alice*
imagevenue.com	a	Cookie	*293*
imagevenue.com	g	Cookie	*00s9229da*
imagevenue.com	age	Cookie	*39*
imagevenue.com	id	Cookie	*27*
imagevenue.com	user_id	Referer	*89*

Seeded Approach

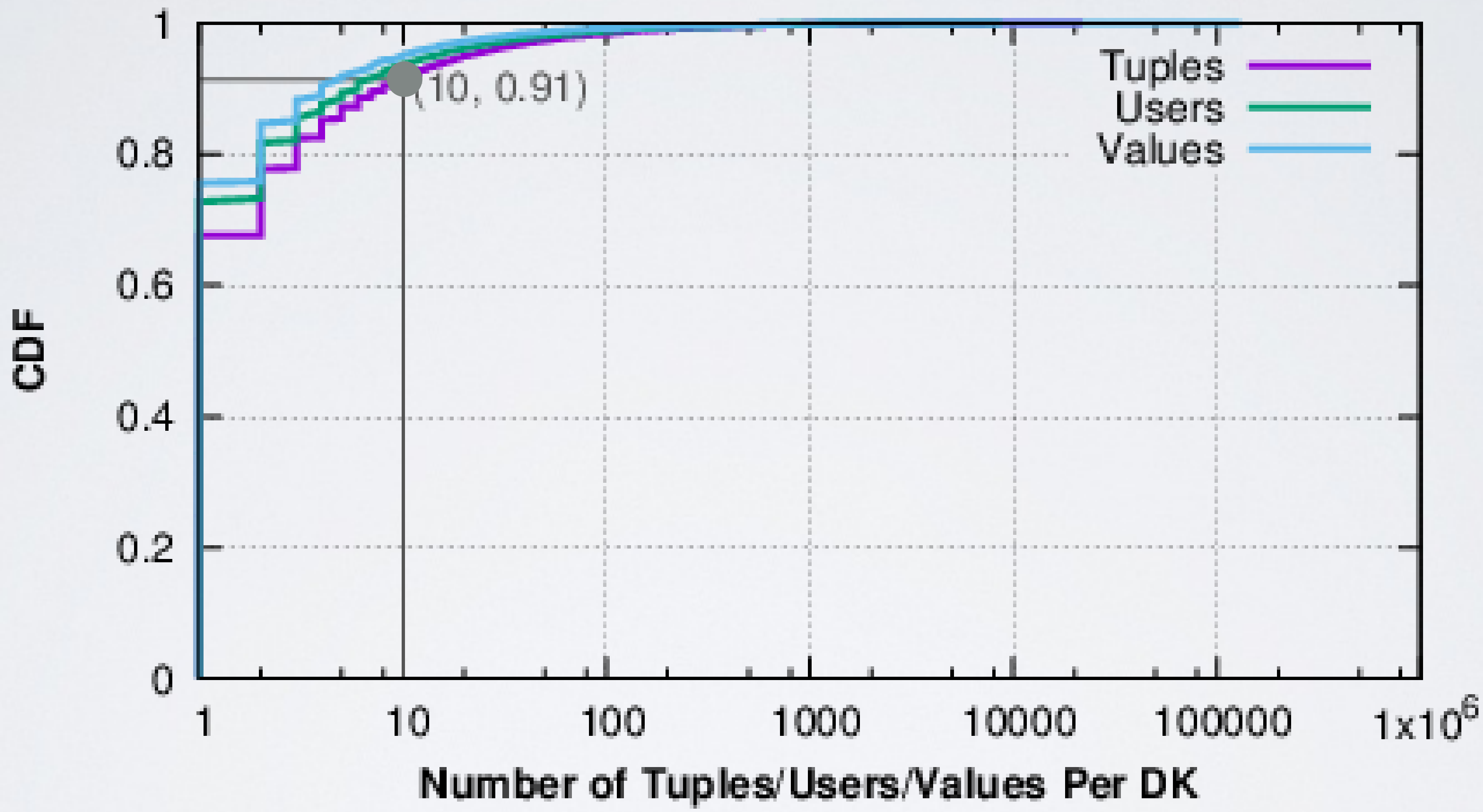
Look for domain-keys with many values that "look like" PI

But many challenges in analyzing data

1. Do every domain-keys have enough number of values?
2. What kinds of value are PI we look for?
3. How to filter out keys with many mismatched values?
4. How to discover missing values?

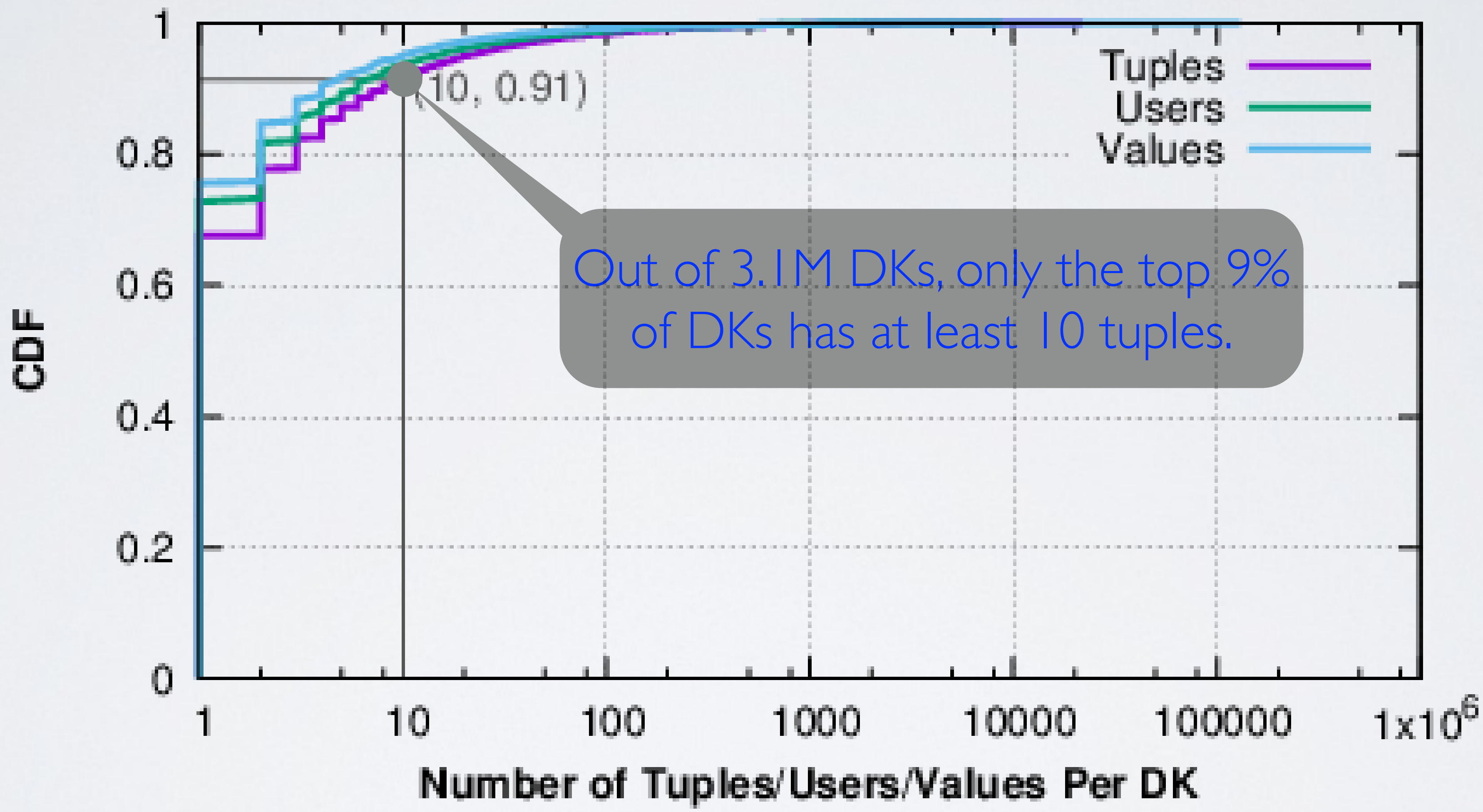
Step1: Pre-processing

1 Does every DK have enough number of values?



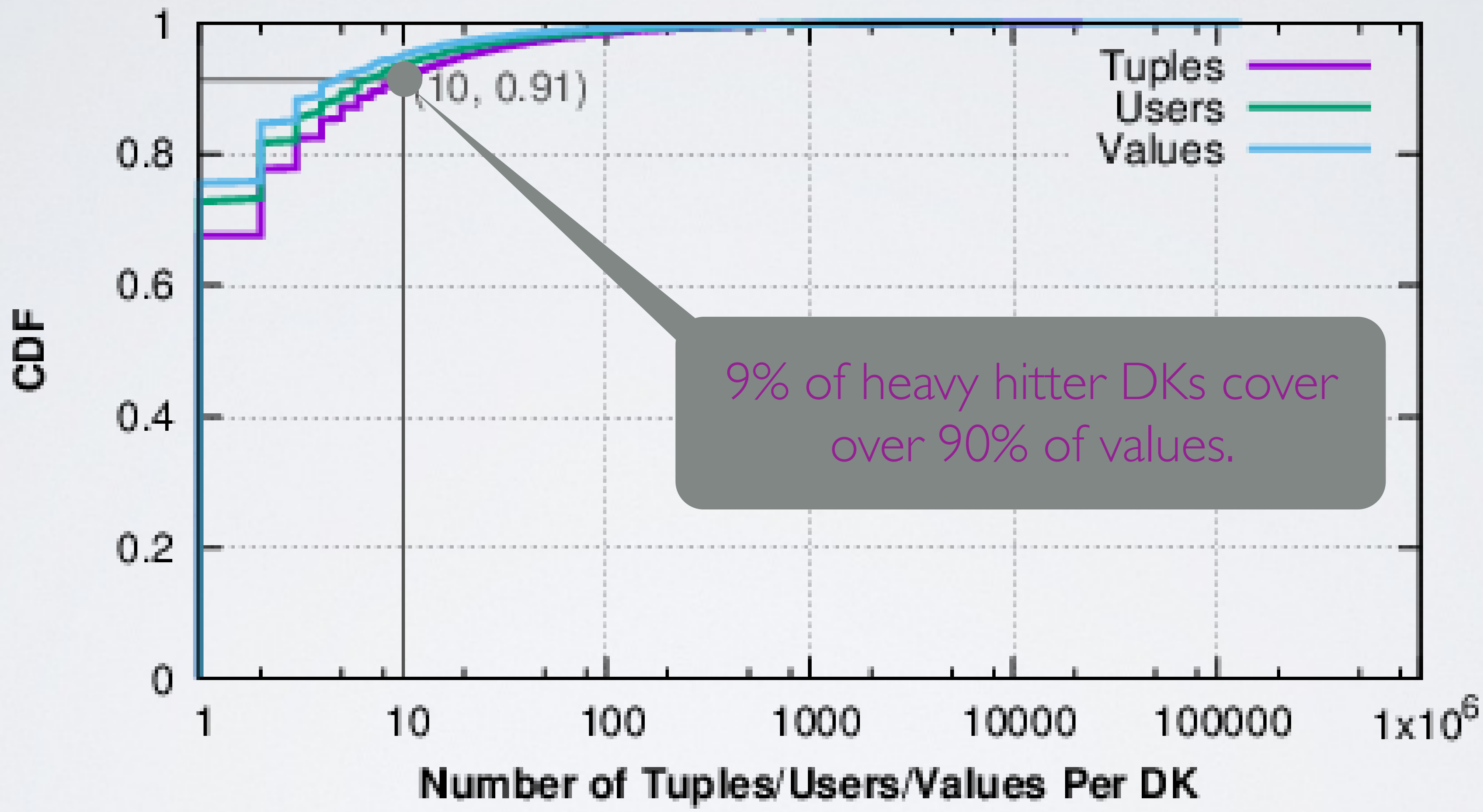
Step1: Pre-processing

1 Does every DK have enough number of values?



Step1: Pre-processing

1 Does every DK have enough number of values?



Step2: Seed rules

2 What kinds of value are PI we look for?

- Regular expressions with constraints and dictionaries

<table>
<tr><th>PI Type</th><th>Seed Rules</th></tr>
<tr><td>AgeRange</td><td>/^[0-9]{1,3}-[0-9]{1,3}$/ (where the second number is larger than the first)</td></tr>
<tr><td>City</td><td>Dictionary of cities, such as {“boston”, “new york”, “chicago”, …}</td></tr>
<tr><td>Email</td><td>/^(\w|\-|_|\.)+\@((\w|\-|_)+\.)+[a-zA-Z]{2,}$/</td></tr>
<tr><td>Geo</td><td>/^[\+\-]{0,1}\d+\.\d{4}\d+$/ (where the value is within the range of the country)</td></tr>
<tr><td>Gender</td><td>/^[mf]$/ or /^(fe)?male$/ or the corresponding words for the male/female in local language</td></tr>
<tr><td>Name</td><td>Dictionary of boy and girl names, such as {“alice”, “christian”, …}</td></tr>
<tr><td>Phone</td><td>/^([+]code?((38[{8,9}|0])|(34[{7-9}|0])|(36[6|6|0])|(33[{3-9}|0])|(32[{3-9}|0])|(32[{8,9}]))([\d]{7})$/</td></tr>
</table>

Step2: Seed rules

2 What kinds of value are PI we look for?

- Regular expressions with constraints and dictionaries

<table>
<tr><th>PI Type</th><th>Seed Rules</th></tr>
<tr><td>AgeRange</td><td>/^[0-9]{1,3}-[0-9]{1,3}$/ (where the second number is larger than the first)</td></tr>
<tr><td>City</td><td>Dictionary of cities, such as {“boston”, “new york”, “chicago”, …}</td></tr>
<tr><td>Email</td><td>/^(\w|\-|_|\.)+\@((\w|\-|_)+\.)+[a-zA-Z]{2,}$/</td></tr>
<tr><td>Geo</td><td>/^[\+\-]{0,1}\d+\.\d{4}\d+$/ (where the value is within the range of the country)</td></tr>
<tr><td>Gender</td><td>/^[mf]$/ or /^(fe)?male$/ or the corresponding words for the male/female in local language</td></tr>
<tr><td>Name</td><td>Dictionary of boy and girl names, such as {“alice”, “christian”, …}</td></tr>
<tr><td>Phone</td><td>/^([+]code?((38[{8,9}|0])|(34[{7-9}|0])|(36[6|6|0])|(33[{3-9}|0])|(32[{3-9}|0])|(32[{8,9}]))([\d]{7})$/</td></tr>
</table>

Step2: Seed rules

2 What kinds of value are PI we look for?

- Regular expressions with constraints and dictionaries

<table>
<tr><th>PI Type</th><th>Seed Rules</th></tr>
<tr><td>AgeRange</td><td>/^[0-9]{1,3}-[0-9]{1,3}$/ (where the second number is larger than the first)</td></tr>
<tr><td>City</td><td>Dictionary of cities, such as {“boston”, “new york”, “chicago”, …}</td></tr>
<tr><td>Email</td><td>/^(\w|\-|_|\.)+\@((\w|\-|_)+\.)+[a-zA-Z]{2,}$/</td></tr>
<tr><td>Geo</td><td>/^[\+\-]{0,1}\d+\.\d{4}\d+$/ (where the value is within the range of the country)</td></tr>
<tr><td>Gender</td><td>/^[mf]$/ or /^(fe)?male$/ or the corresponding words for the male/female in local language</td></tr>
<tr><td>Name</td><td>Dictionary of boy and girl names, such as {“alice”, “christian”, …}</td></tr>
<tr><td>Phone</td><td>/^([+]code?((38[{8,9}|0])|(34[{7-9}|0])|(36[6|6|0])|(33[{3-9}|0])|(32[{3-9}|0])|(32[{8,9}]))([\d]{7})$/</td></tr>
</table>

Step3: Filtering domain-keys

3 How to filter out DKs with many mismatched values?

- For each DK, plot ratio of matched values

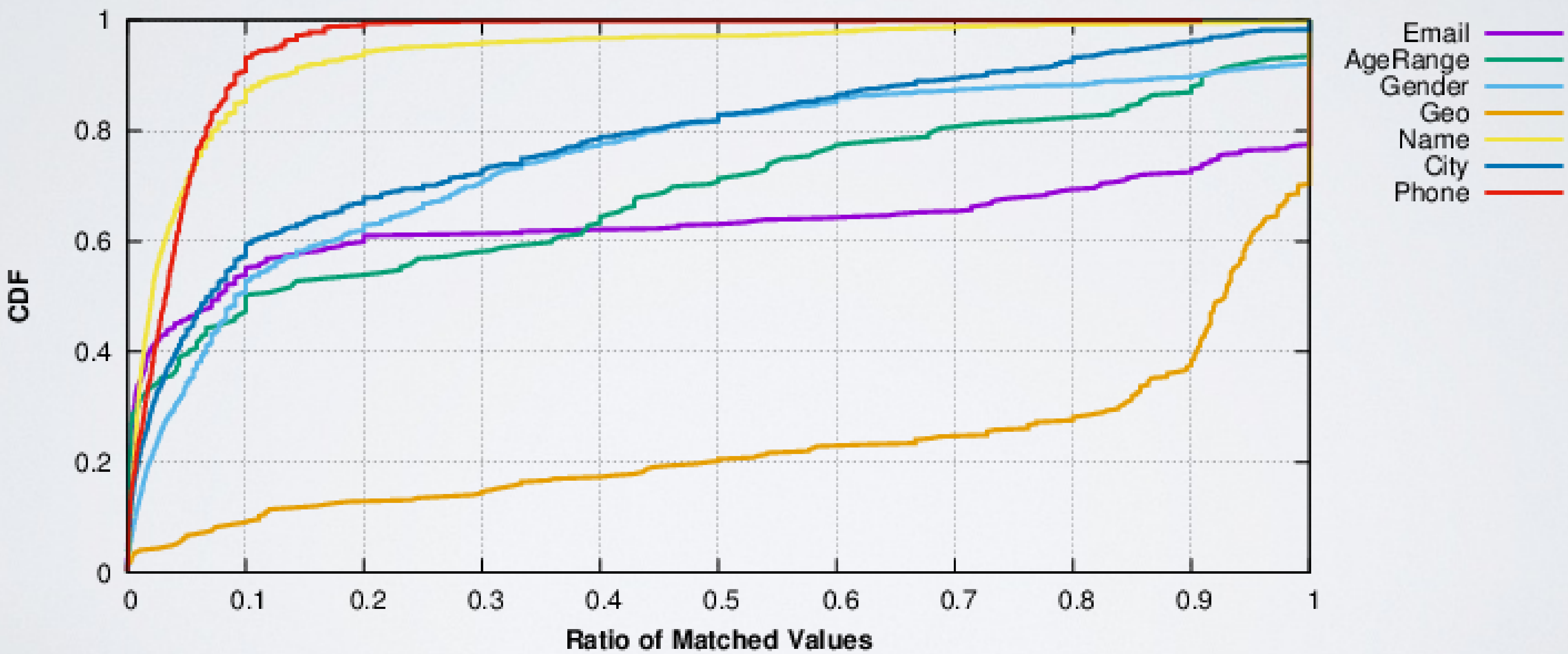
$$Ratio = \frac{NumofMatchedValues}{TotalValues}$$

Step3: Filtering domain-keys

3 How to filter out DKs with many mismatched values?

- For each DK, plot ratio of matched values

$$Ratio = \frac{NumofMatchedValues}{TotalValues}$$

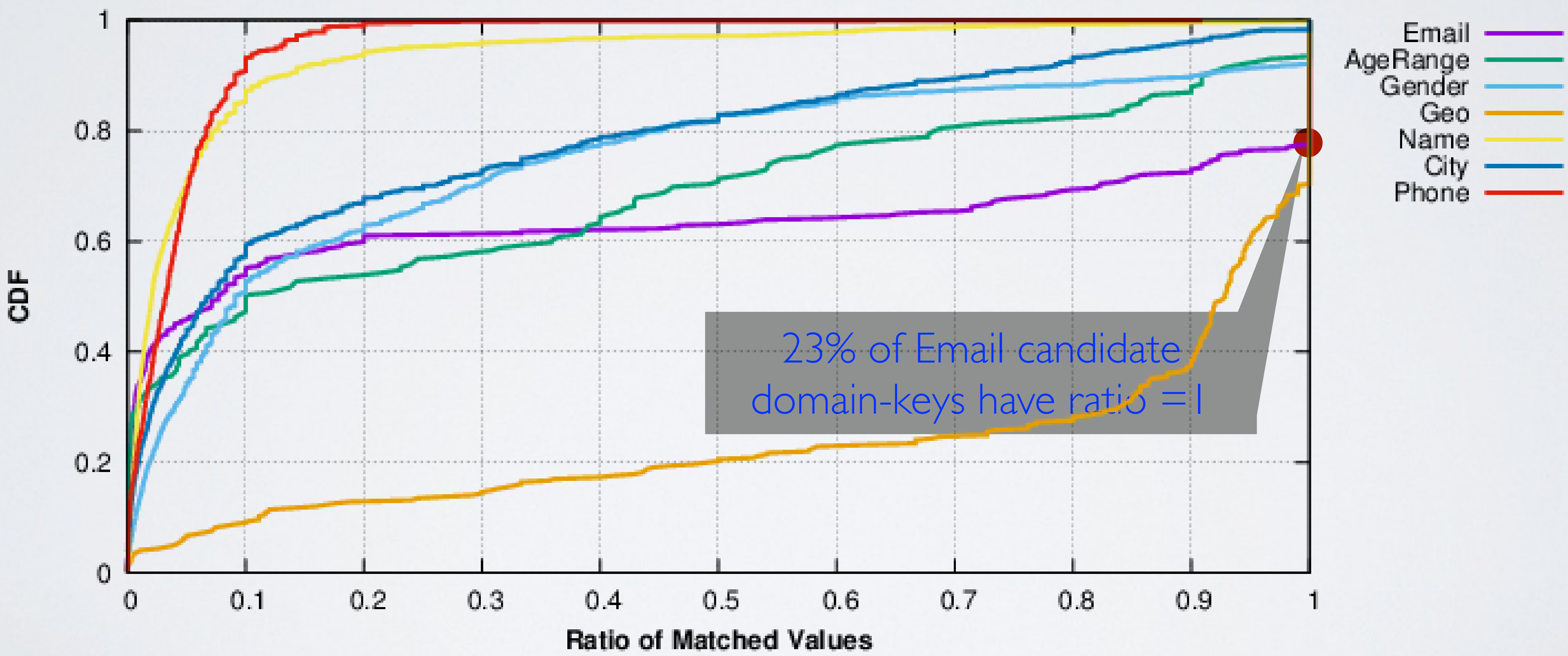


Step3: Filtering domain-keys

3 How to filter out DKs with many mismatched values?

- For each DK, plot ratio of matched values

$$Ratio = \frac{NumofMatchedValues}{TotalValues}$$

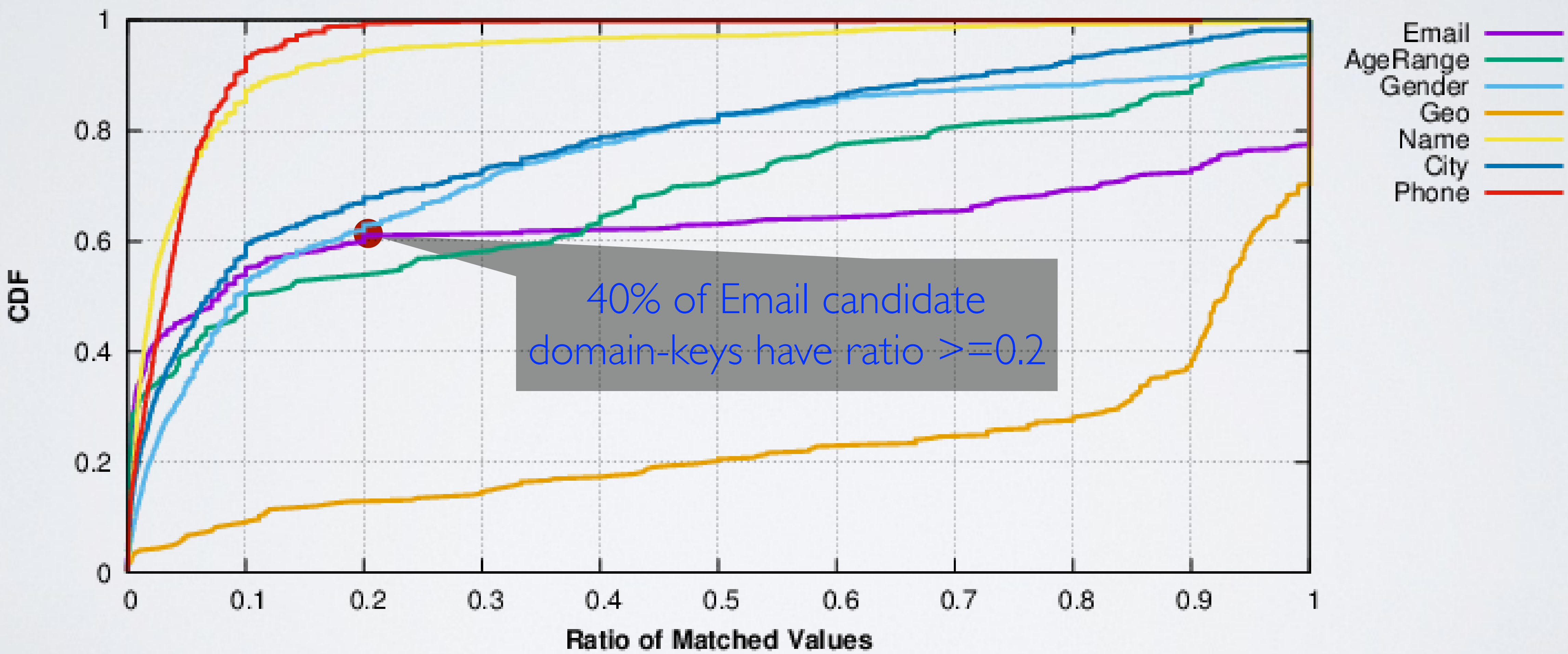


Step3: Filtering domain-keys

3 How to filter out DKs with many mismatched values?

- For each DK, plot ratio of matched values

$$Ratio = \frac{NumofMatchedValues}{TotalValues}$$



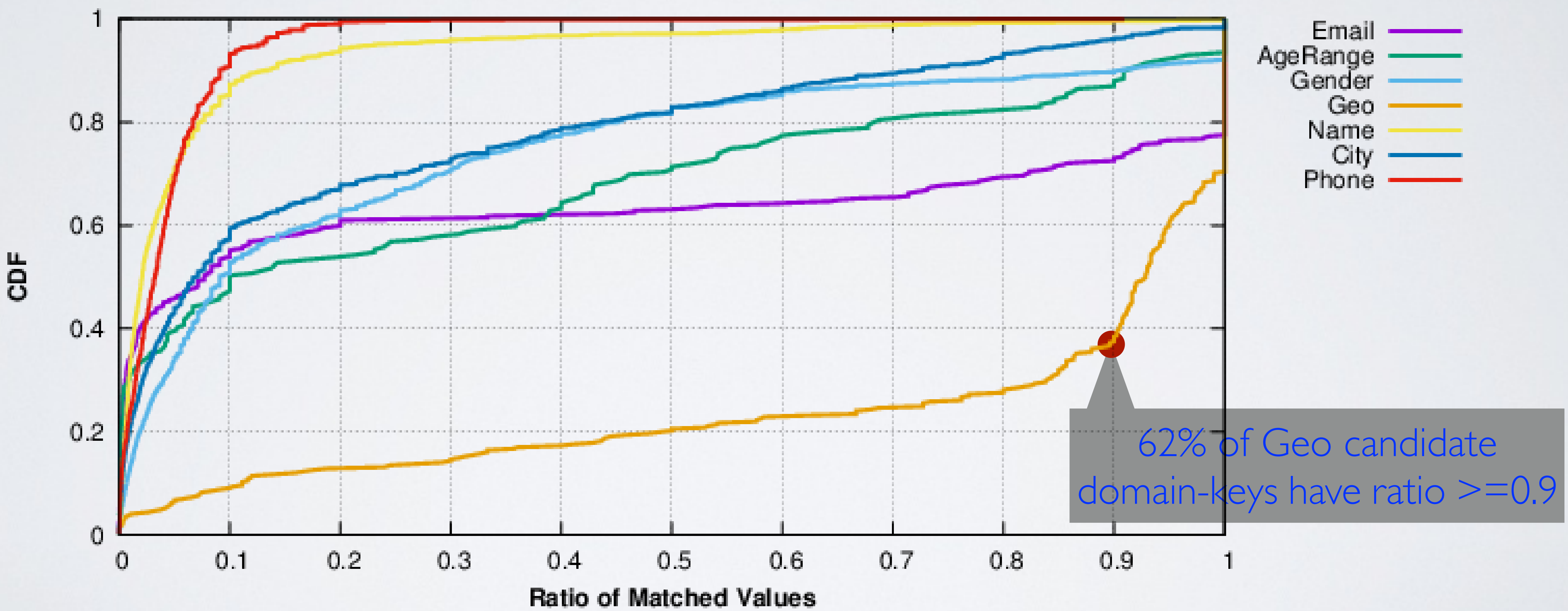
Pick knee points to select threshold

Step3: Filtering domain-keys

3 How to filter out DKs with many mismatched values?

- For each DK, plot ratio of matched values

$$Ratio = \frac{NumofMatchedValues}{TotalValues}$$



Pick knee points to select threshold

Step4: Expansion

4 How to expand the missing values?

- Seed rules do not cover all possible cases

User-Index	Domain	Key	Value
1	google-analytics.com	email	johnDoe@gmail.com
2	google-analytics.com	email	janeDoe@hotmail.com
1	google-analytics.com	email	johnDoe
2	google-analytics.com	email	janeDoe
3	facebook.com	gender	female
4	facebook.com	gender	m
5	facebook.com	gender	f
6	facebook.com	gender	1
7	facebook.com	gender	f-f
8	facebook.com	gender	f-m

Take all values of DKs with enough matches

Step4: Expansion

4 How to expand the missing values?

- Seed rules do not cover all possible cases

User-Index	Domain	Key	Value
1	google-analytics.com	email	*johnDoe@gmail.com*
2	google-analytics.com	email	*janeDoe@hotmail.com*
1	google-analytics.com	email	*johnDoe*
2	google-analytics.com	email	*janeDoe*
3	facebook.com	gender	*female*
4	facebook.com	gender	*m*
5	facebook.com	gender	*f*
6	facebook.com	gender	*l*
7	facebook.com	gender	*f-f*
8	facebook.com	gender	*f-m*

Take all values of DKs with enough matches

Step4: Expansion

4 How to expand the missing values?

- Seed rules do not cover all possible cases

User-Index	Domain	Key	Value
1	google-analytics.com	email	johnDoe@gmail.com
2	google-analytics.com	email	janeDoe@hotmail.com
1	google-analytics.com	email	johnDoe
2	google-analytics.com	email	janeDoe
3	facebook.com	gender	female
4	facebook.com	gender	m
5	facebook.com	gender	f
6	facebook.com	gender	1
7	facebook.com	gender	f-f
8	facebook.com	gender	f-m

Take all values of DKs with enough matches

Outline

- ~~Motivation~~
- ~~Dataset~~
- ~~Methodology~~
- Evaluation

Baseline approach

Key-semantic based approach

- Can we rely on semantics of Keys?

PI Type	Keywords
AgeRange	age
City	city, area, state, region, …
Email	email, account, login, logon, …
Geo	lat, lon, lng, geo
Gender	gen, gnd, gdr, ycg, sex, …
Name	name, nome, pers, author
Phone	phone, pid, …

Observed HTTP transaction

GET /foo.html?user_firstname=Alice HTTP/1.1
Host: imagevenue.com
Cookie: a=293&email=1&message=39&id=27
ETag: 2039-2dc90ea2-12
Referer: http://www.facebook.com/?user_id=89
Accept-Encoding: deflate,gzip

HTTP/1.1 200 OK
Date: Mon, 23, May 2013 22:38:34 GMT

Evaluation

Methodology

- Six human raters on sampling of results (domain-key + list of 10 values)
- Label as either positive, negative, or neutral

Evaluation

Methodology

- Six human raters on sampling of results (domain-key + list of 10 values)
- Label as either positive, negative, or neutral

PI Type	Seeded #DKs	False Positive	Baseline #DKs	False Positive
AgeRange	17	0.0%	3,729	88.0%
City	465	8.8%	3,191	76.0%
Email	154	3.9%	3,253	76.0%
Geo	147	10.0%	1,358	100.0%
Gender	214	0.0%	1,986	88.0%
Name	100	52.5%	2,142	92.0%
Phone	11	90.9%	3,864	100.0%
Total	1,108	13.6%	19,523	89.5%

Evaluation

Methodology

- Six human raters on sampling of results (domain-key + list of 10 values)
- Label as either positive, negative, or neutral

PI Type	Seeded #DKs	False Positive	Baseline #DKs	False Positive
AgeRange	17	0.0%	3,729	88.0%
City	465	8.8%	3,191	76.0%
Email	154	3.9%	3,253	76.0%
Geo	147	10.0%	1,358	100.0%
Gender	214	0.0%	1,986	88.0%
Name	100	52.5%	2,142	92.0%
Phone	11	90.9%	3,864	100.0%
Total	1,108	13.6%	19,523	89.5%

Evaluation

Methodology

- Six human raters on sampling of results (domain-key + list of 10 values)
- Label as either positive, negative, or neutral

PI Type	Seeded #DKs	False Positive	Baseline #DKs	False Positive
AgeRange	17	0.0%	3,729	88.0%
City	465	8.8%	3,191	76.0%
Email	154	3.9%	3,253	76.0%
Geo	147	10.0%	1,358	100.0%
Gender	214	0.0%	1,986	88.0%
Name	100	52.5%	2,142	92.0%
Phone	11	90.9%	3,864	100.0%
Total	1,108	13.6%	19,523	89.5%

- False-positive: 703 flagged domain-keys from 1,108 Seeded (13.6%)
- False-positive: 200 flagged domain-keys from 19,523 Baseline (89.5%)

Evaluation

Methodology

- Six human raters on sampling of results (domain-key + list of 10 values)
- Label as either positive, negative, or neutral

PI Type	Seeded #DKs	False Positive	Baseline #DKs	False Positive
AgeRange	17	0.0%	3,729	88.0%
City	465	8.8%	3,191	76.0%
Email	154	3.9%	3,253	76.0%
Geo	147	10.0%	1,358	100.0%
Gender	214	0.0%	1,986	88.0%
Name	100	52.5%	2,142	92.0%
Phone	11	90.9%	3,864	100.0%
Total	1,108	13.6%	19,523	89.5%

- False-negative: 1000 flagged domain-keys from the rest (2.7%)

Conclusion

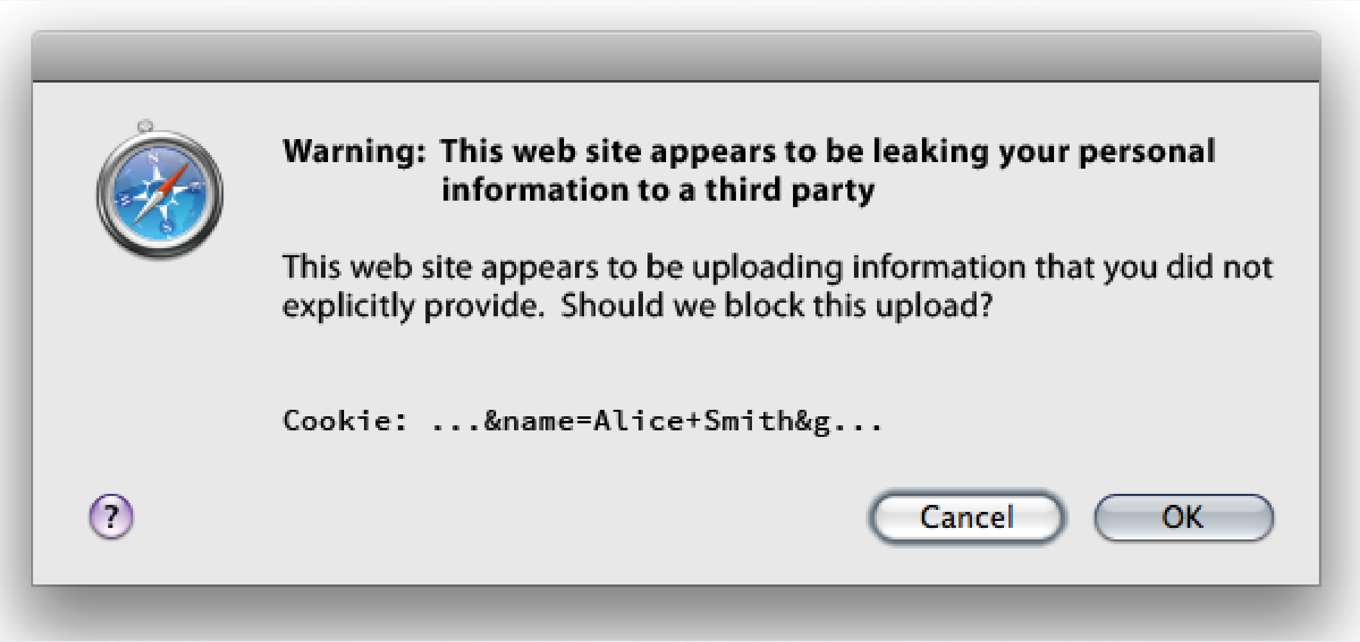
Proposed seeded approach

- Automatically locates rare PI embedded in network traffic
- Low false negative (2.7%) and false positive (13.6%)

Future work

- Select thresholds automatically (state space exploration)
- Differentiate between PI the user has intentionally shared and doesn't

Eventually: Inform user of what is being leaked automatically



Questions?