Anonymization of location data does not work: a large-scale measurement study

Hui Zang, Sprint
Jean Bolot, Technicolor

September 21, 2011

© 2005 Sprint Nextel. All rights reserved.

CDRs are useful, but …

- CDRs can be used for various purposes
  - >Marketing
  - >Business
  - >Security
  - >Location based applications and services
  - >Mobility modeling

- Privacy might be breached if such data is not anonymized and handled properly

© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Outline

- CDR
- k-anonymity
- Dataset
- Factors impacting size of anonymity sets
  - different location granularity levels
  - Distance between locations
  - geographical regions
  - extra side knowledge
- Solutions
  - Time domain
  - Spatial domain

© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

CDR Example

Attribute	Value
Mobile ID	123-456-7890
Time of call	2010 02 02 12 33 02
Call duration	300 seconds
Start Cell ID	153
Start Sector ID	2
End Cell ID	157
End Sector ID	1
Call direction	incoming
Caller ID	123-456-0987

© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Simple anonymization

Attribute	**Value**
Mobile ID	**00000001**
Time of call	2010 02 02 12 33 02
Call duration	300 seconds
Start Cell ID	153
Start Sector ID	2
End Cell ID	157
End Sector ID	1
Call direction	incoming
Caller ID	**00000002**

© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Simple anonymization for location record

Attribute	Value
Mobile ID	**00000001**
Time of call	2010 02 02 12 33 02
Call duration	300 seconds
Start Cell ID	153
Start Sector ID	2
End Cell ID	157
End Sector ID	1

who is 00000001 ?

Re-identification Attacks

© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Privacy in data publishing

- Re-identification attacks
- Majority of US population can be uniquely identified by (gender, zipcode, birth-date)

Quasi-identifier

- Anonymity set: individuals with the same (gender, zipcode, birth-date)
- Re-identifiable if ||Anonymity Set|| = 1

© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

K-anonymity

- K-anonymity constraint
  - At least k individuals have the same quasi-identifier
  - ||Anonymity set|| >= k
  - E.g. using first 4 digits of zipcode, k = 2

- Our contribution: k-anonymity in location data from cellular networks

© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Dataset

- Nation-wide CDR
- Feb – April 2010
- 25 M subscribers (subset)
- 30 B records
- >100k locations

© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Quasi-Identifiers

- Top N locations
- N = 1, 2, 3
  - User x's trace: 1-13, 1-23, 1-13, 1-23, 2-151
  - User x's top locations: 1-13:3, 1-23:2, 2-151:1
  - Anonymity sets:
    - Top 1: 1-13
    - Top 2: 1-13, 1-23
    - Top 3: 1-13, 1-23, 2-151
- Six granularity levels:
  - Sector, cell, zip-code, city, county, state
- For example,
  - Top 1 location at sector level:1-13-1
  - Top 3 locations at cell level: 1-13, 1-23, 2-151
  - Top 2 locations at state level: CA, CA

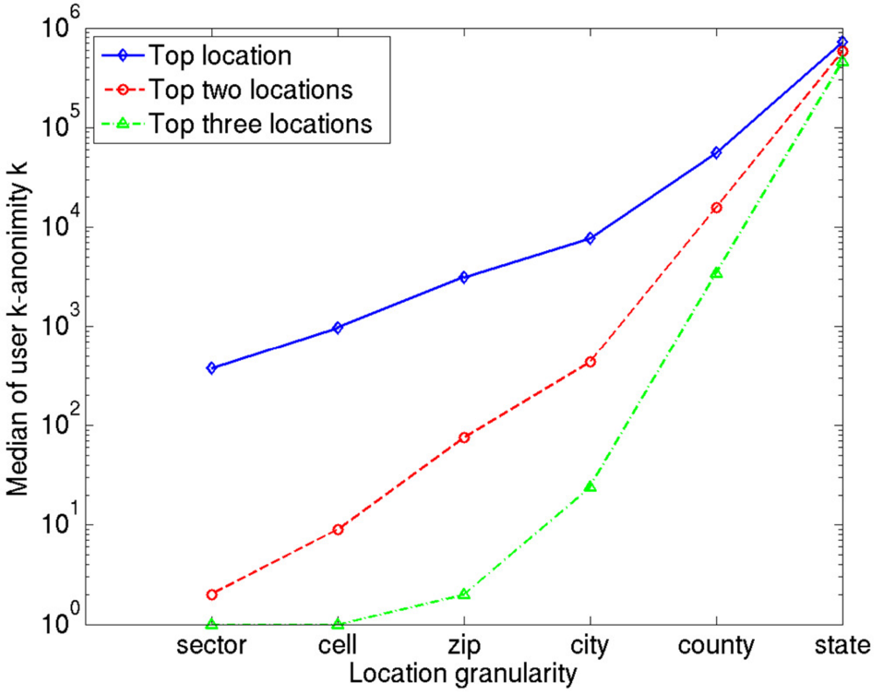
© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Factors affecting anonymity

- N
- Location granularity
- Distance between top N locations
- Geographical regions
- Other information

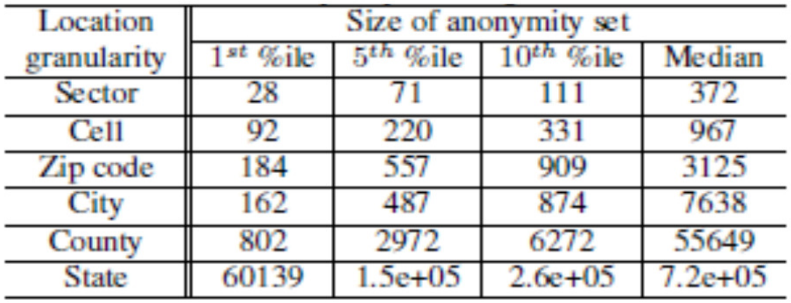
© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

N & granularity



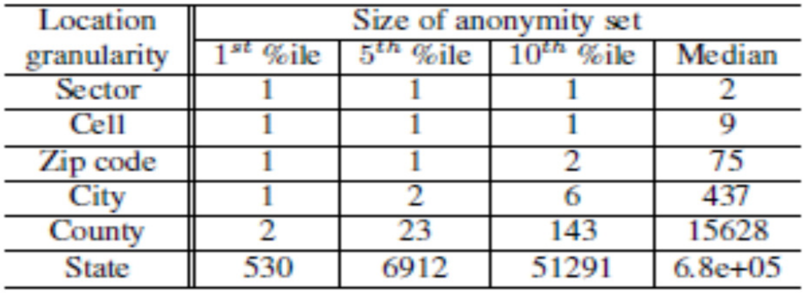
Median of users' k-anonymity at various granularity levels

Top 1 location



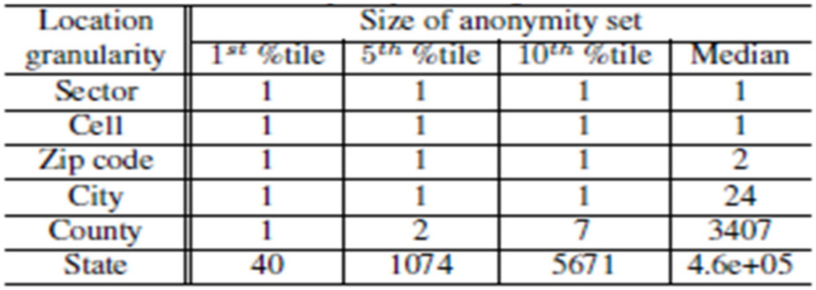
Location granularity	Size of anonymity set			
	1^{st} %ile	5^{th} %ile	10^{th} %ile	Median
Sector	28	71	111	372
Cell	92	220	331	967
Zip code	184	557	909	3125
City	162	487	874	7638
County	802	2972	6272	55649
State	60139	1.5e+05	2.6e+05	7.2e+05

Top 2 locations



Location granularity	Size of anonymity set			
	1^{st} %ile	5^{th} %ile	10^{th} %ile	Median
Sector	1	1	1	2
Cell	1	1	1	9
Zip code	1	1	2	75
City	1	2	6	437
County	2	23	143	15628
State	530	6912	51291	6.8e+05

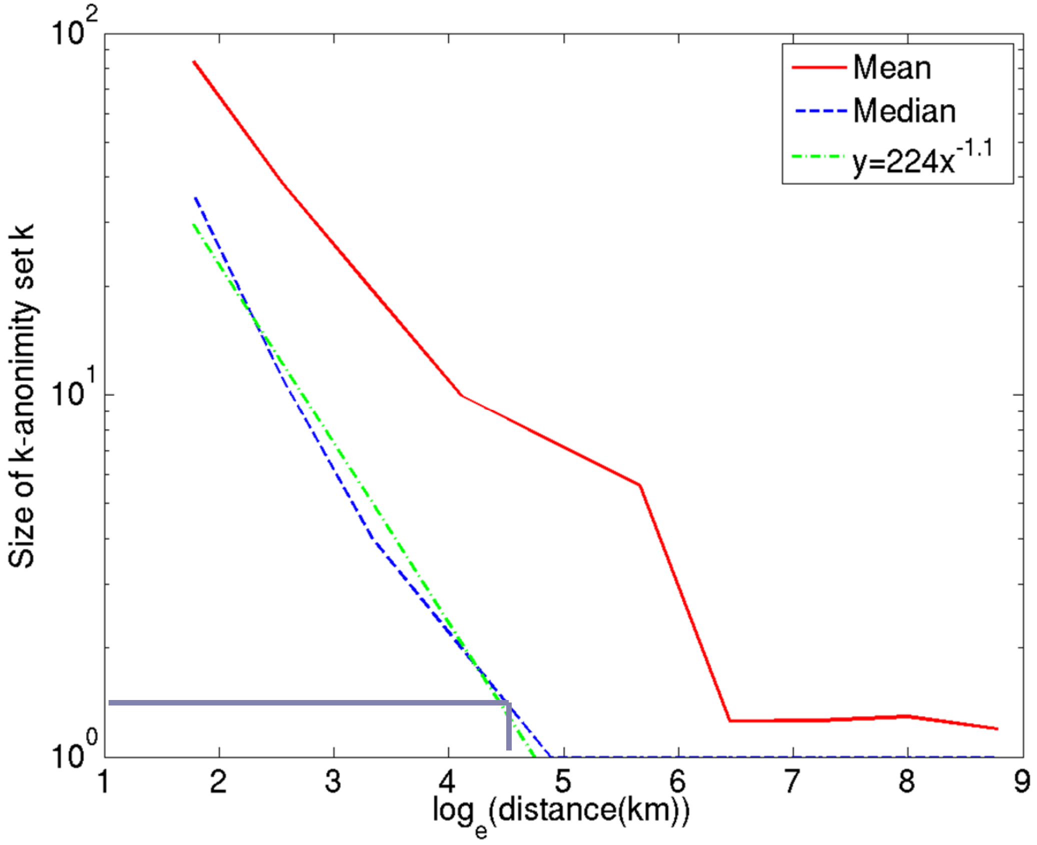
Top 3 locations



Location granularity	Size of anonymity set			
	1^{st} %tile	5^{th} %tile	10^{th} %tile	Median
Sector	1	1	1	1
Cell	1	1	1	1
Zip code	1	1	1	2
City	1	1	1	24
County	1	2	7	3407
State	40	1074	5671	4.6e+05

© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

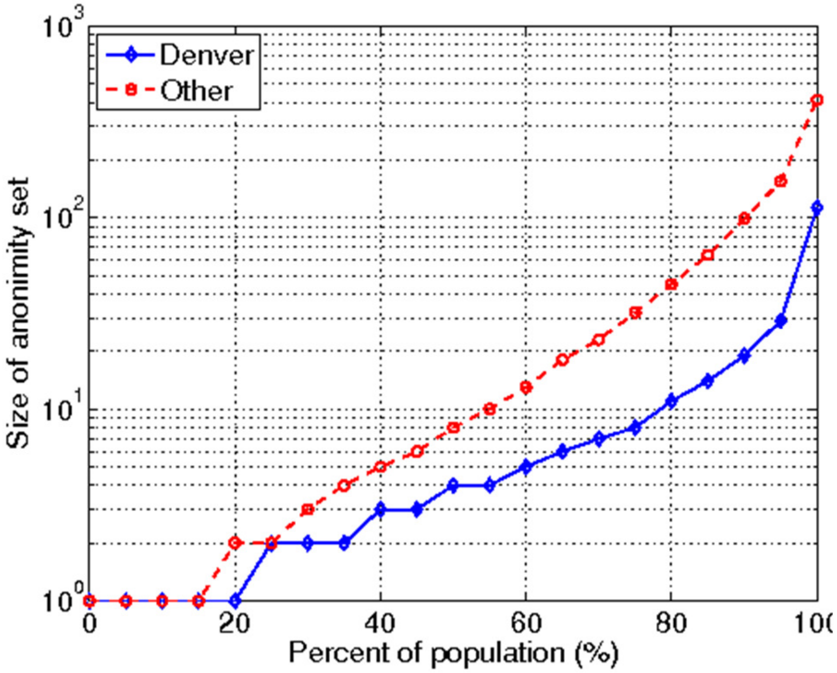
Distance between top 2 locations



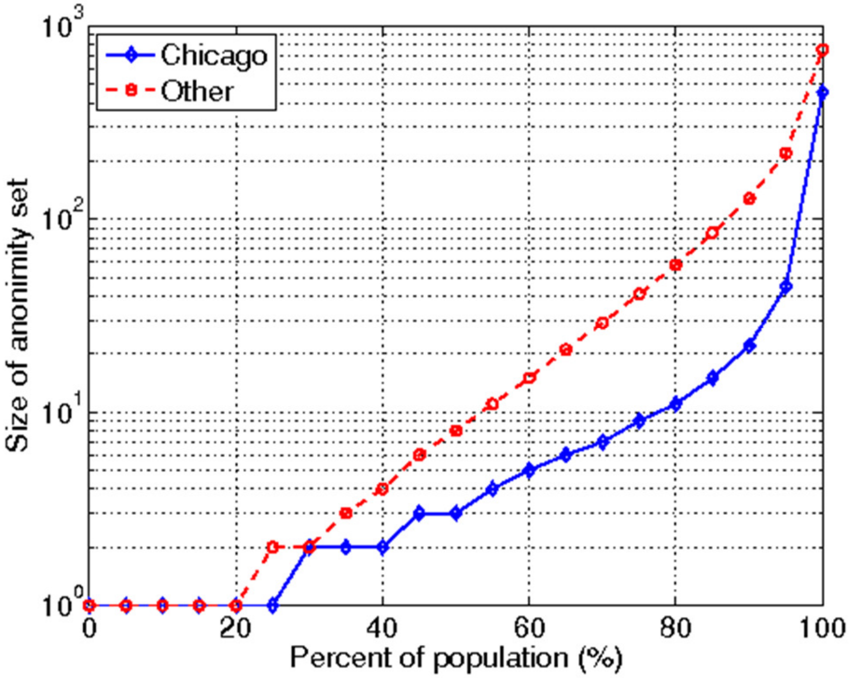
100 km

© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Geographical Regions – urban vs. rural



Colorado



Illinois

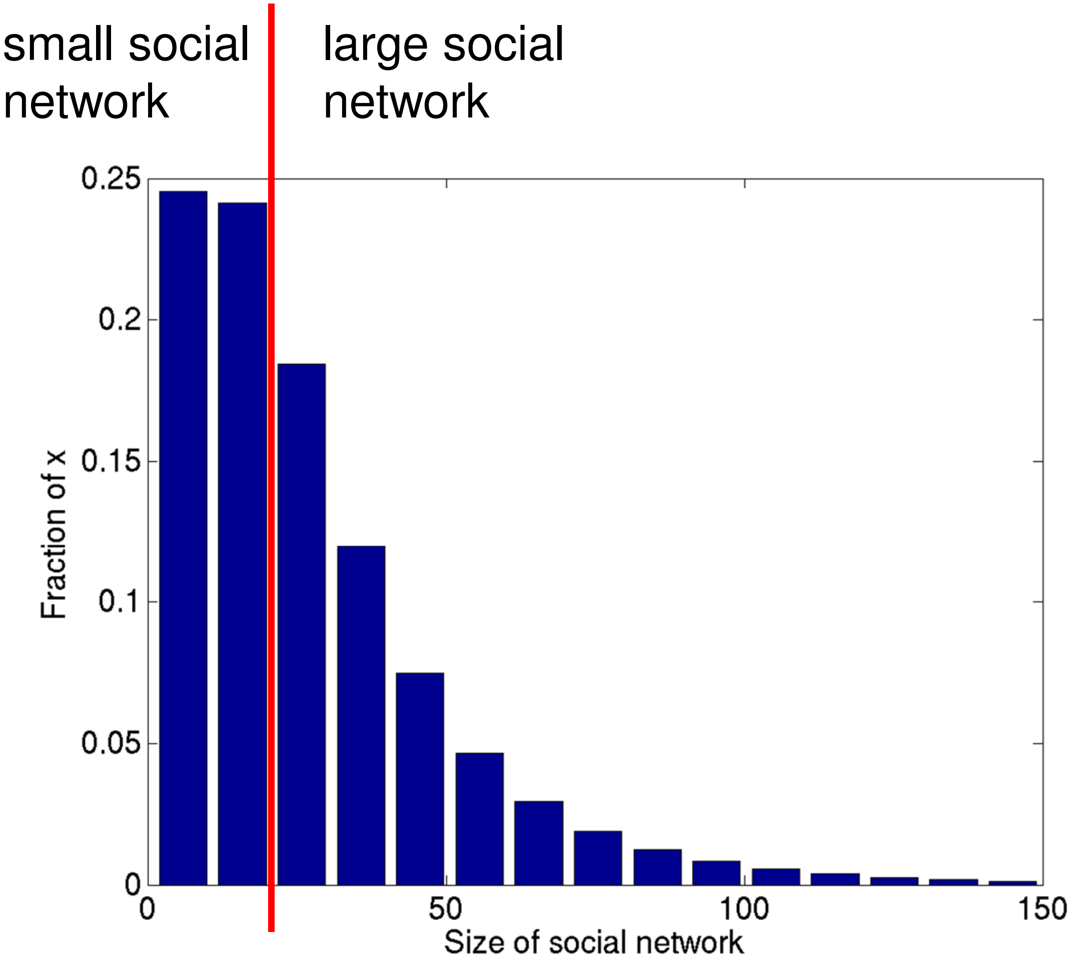
© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Extra side information

Attribute	**Value**
Mobile ID	**00000001**
Time of call	2010 02 02 12 33 02
Call duration	300 seconds
Start Cell ID	153
Start Sector ID	2
End Cell ID	157
End Sector ID	1
Call direction	incoming
Caller ID	***00000002***

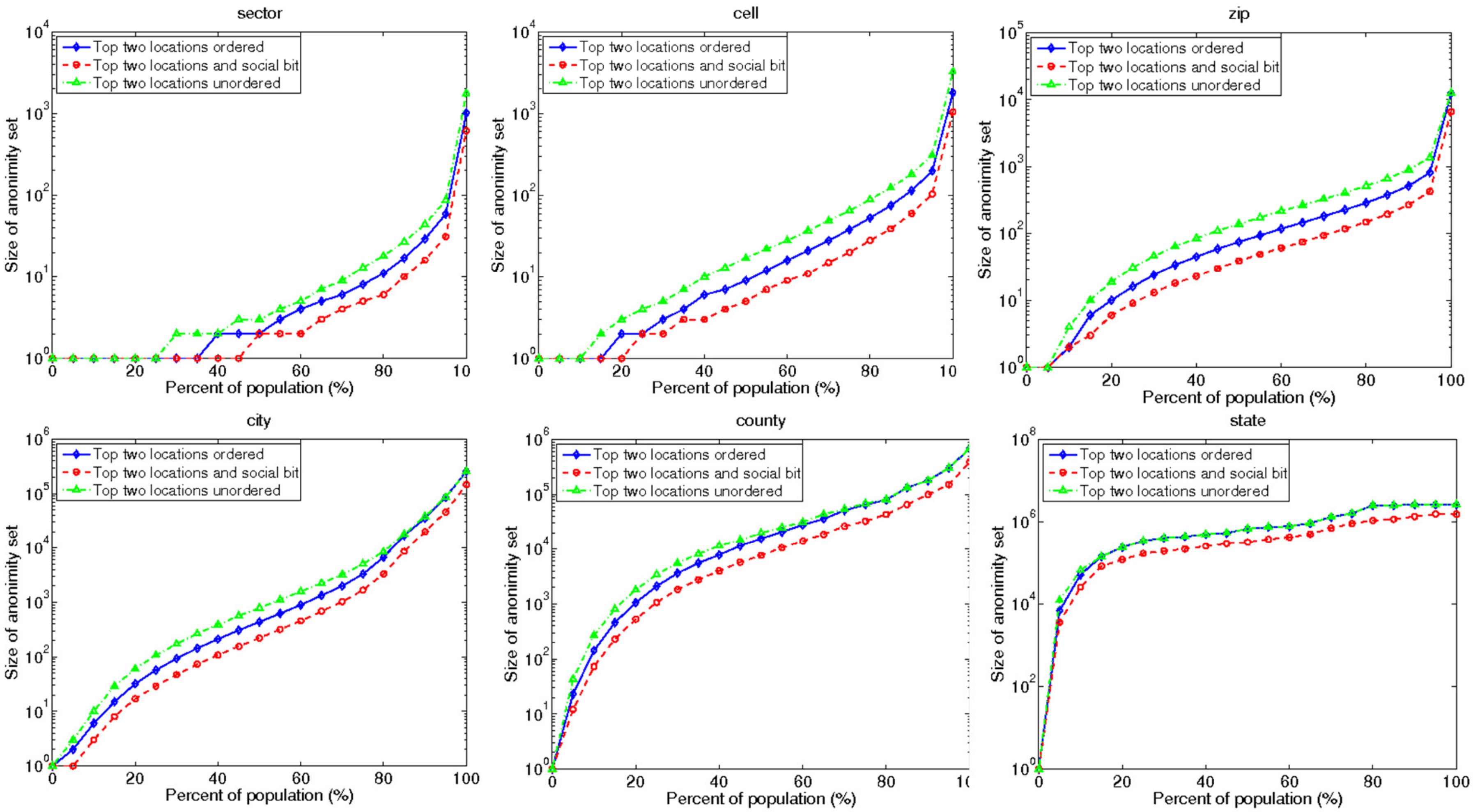
© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Extra side information



© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Extra side information



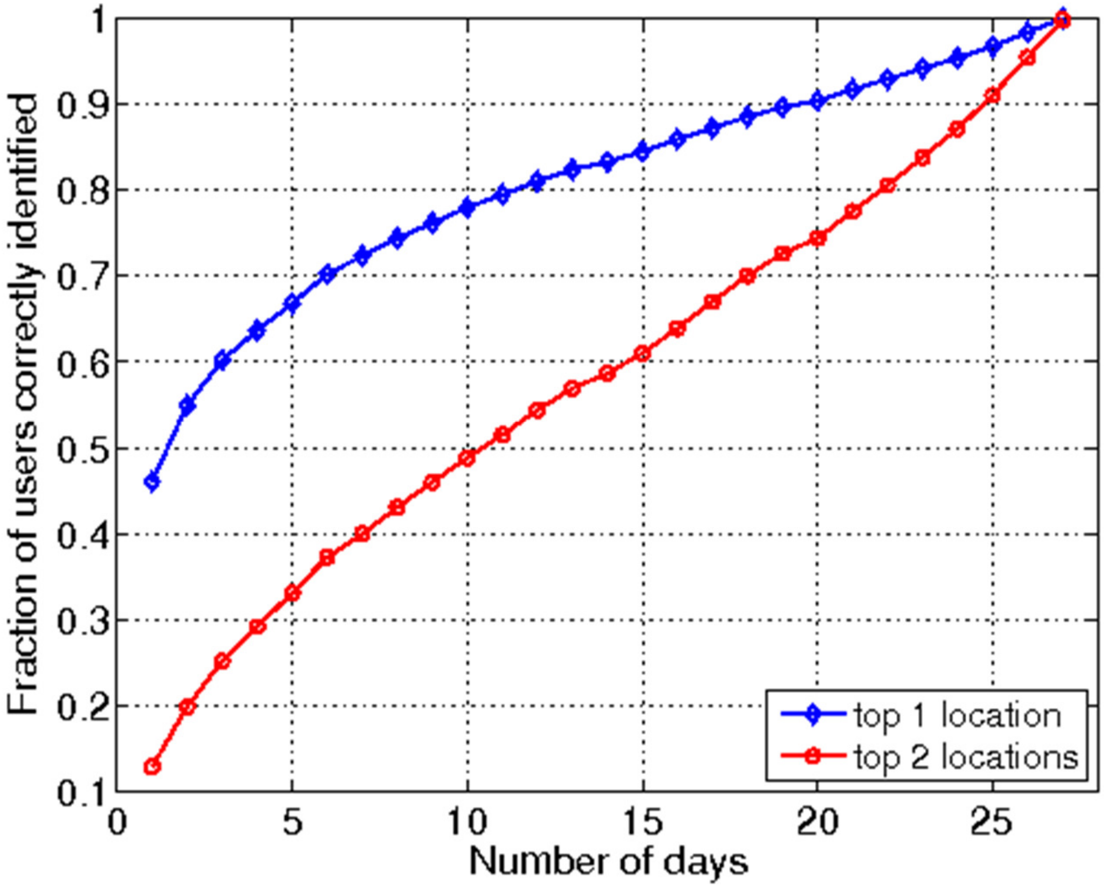
Red curves: size of anonymity sets reduces by half

Solutions

- Spatial and time domain solutions:
  - Publish traces at zip-code granularity or above
  - Publish short traces, such as a day

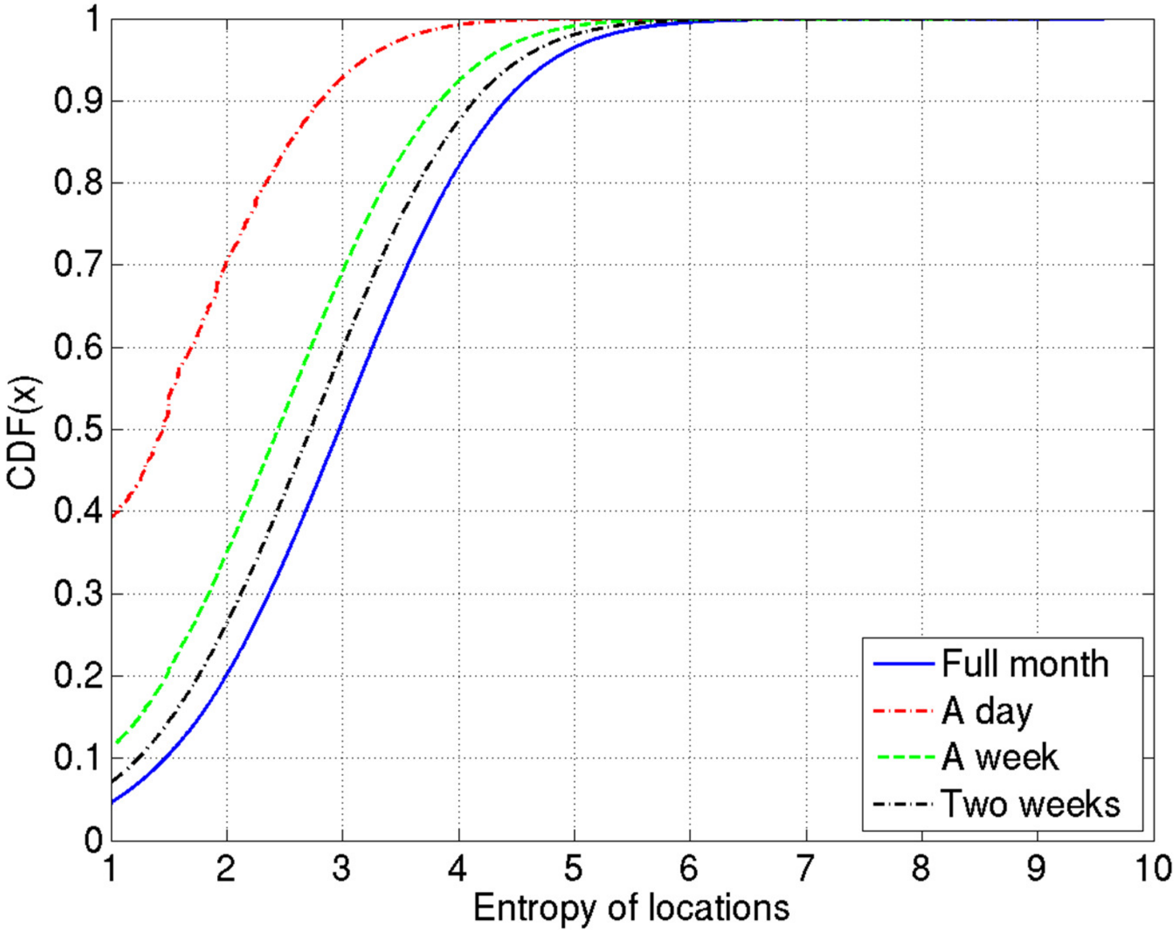
- Reduction of utility of published traces
  - Mobility modeling
  - Identifying preferred locations

Fraction of users whose top locations are correctly extracted



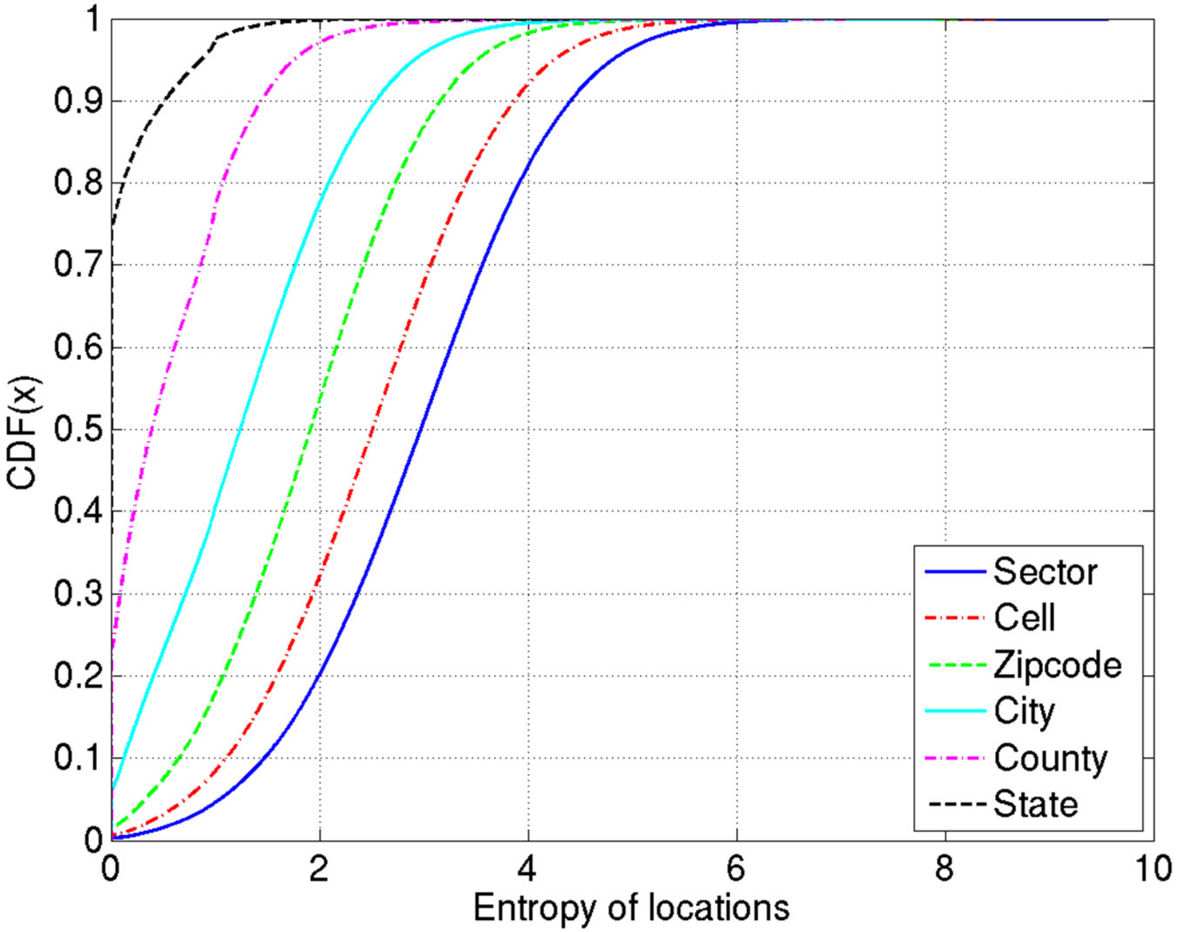
© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Entropy of traces of different durations



© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Entropy of traces of different location granularity



© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Conclusions

- Availability of large scale cell phone data has enabled and will continue to enable a wide range of new services and applications

- Cell phone data are economically valuable

- Subscribers' privacy is at risk if such data is not anonymized and handled properly
  - Anonymity depends on N, granularity, geographical regions, etc.
  - Time domain and spatial domain approaches are proposed to improve anonymization

© 2011 Sprint. This information is subject to Sprint policies regarding use and is the property of Sprint and/or its relevant affiliates and may contain restricted, confidential or privileged materials intended for the sole use of the intended recipient. Any review, use, distribution or disclosure is prohibited without authorization.

Thank you !

hui.zang@sprint.com