On Your Social Network De-anonymizablity: Quantification and Large Scale Evaluation with Seed Knowledge

Shouling Ji, Weiqing Li, and Raheem Beyah

Georgia Institute of Technology

Neil Zhenqiang Gong

University of California, Berkeley

Prateek Mittal

Princeton University

Georgia Institute of Technology

Communications Assurance & Performance Group

Outline

- Introduction and Motivation
- System Model
- De-anonymization Quantification
- Evaluation and Observations
- Conclusion and Future Works

De-anonymization Attack

- Structured/Graph Data

De-anonymization Attack

- Structured/Graph Data
  - Social Networks

De-anonymization Attack

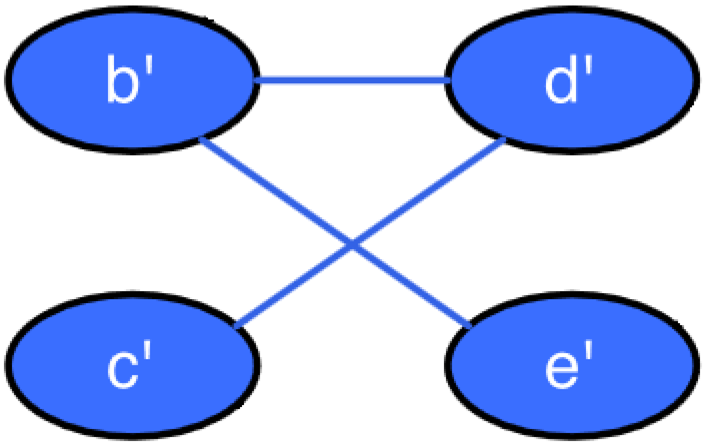
- Structured/Graph Data
  - Social Networks
  - Mobility Traces

De-anonymization Attack

- Structured/Graph Data
  – Social Networks
  – Mobility Traces
  – Email Networks

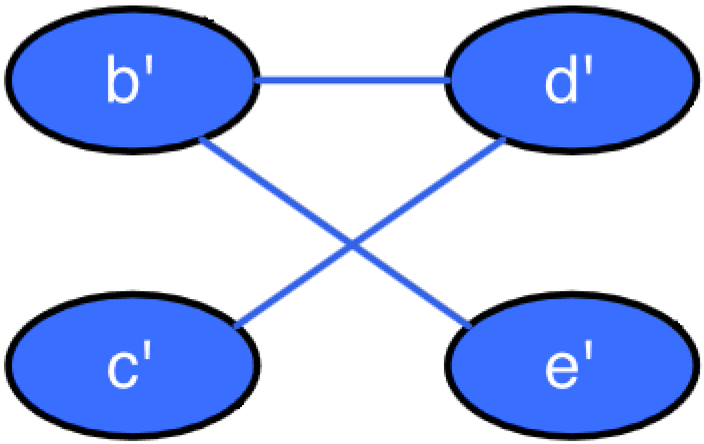
De-anonymization Attack

The Anonymized Graph
(e.g. Facebook)

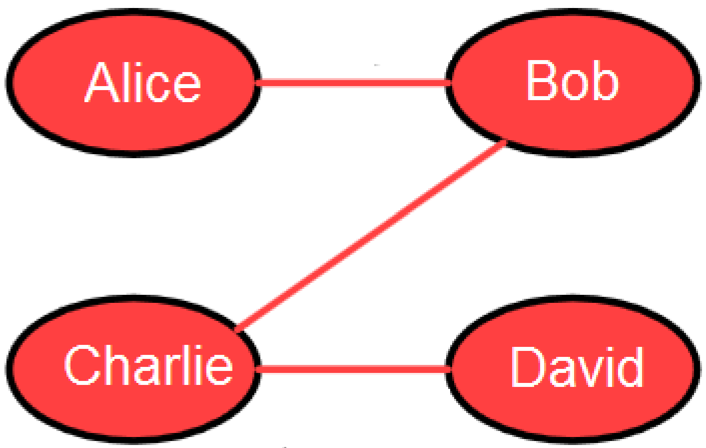


De-anonymization Attack

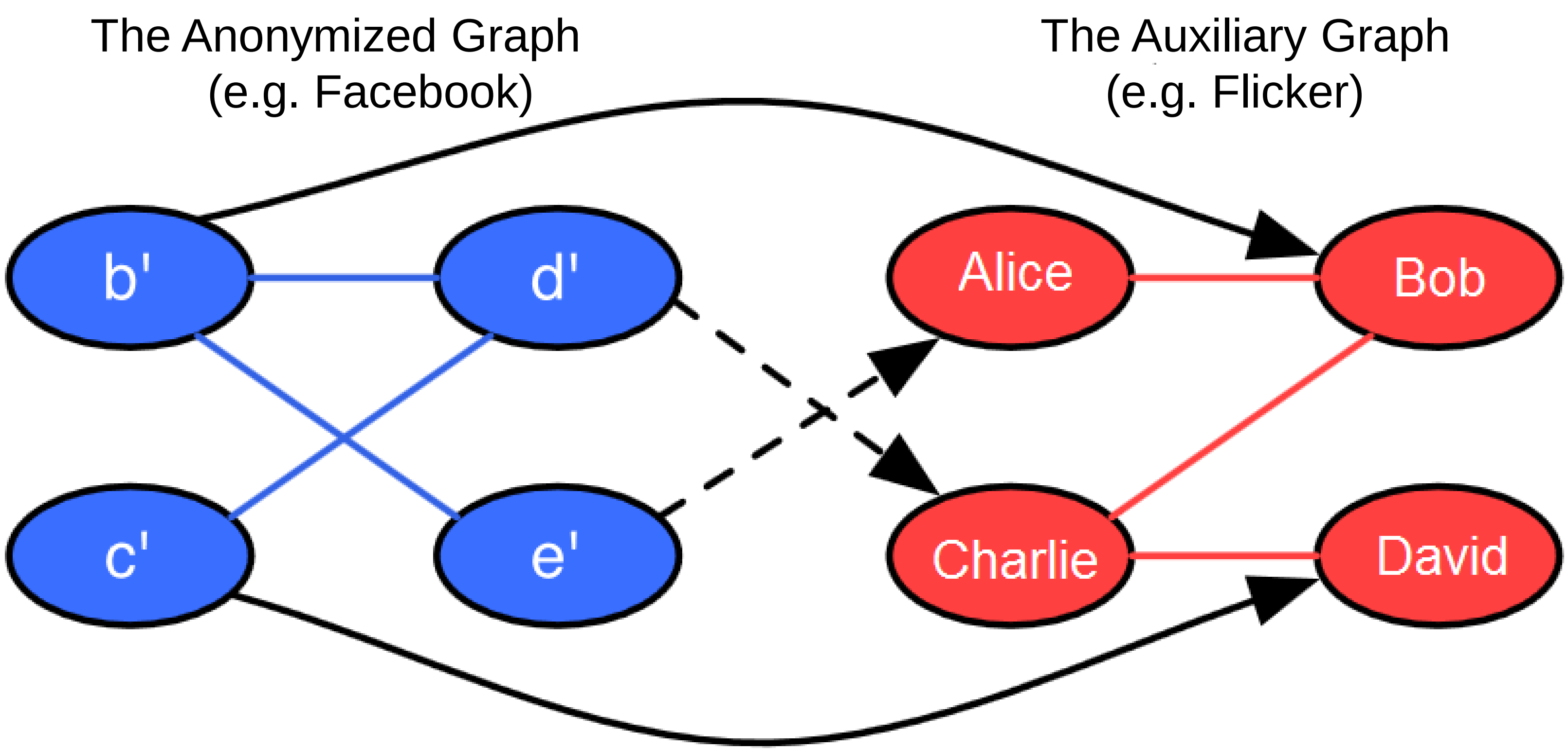
The Anonymized Graph
(e.g. Facebook)



The Auxiliary Graph
(e.g. Flicker)



De-anonymization Attack



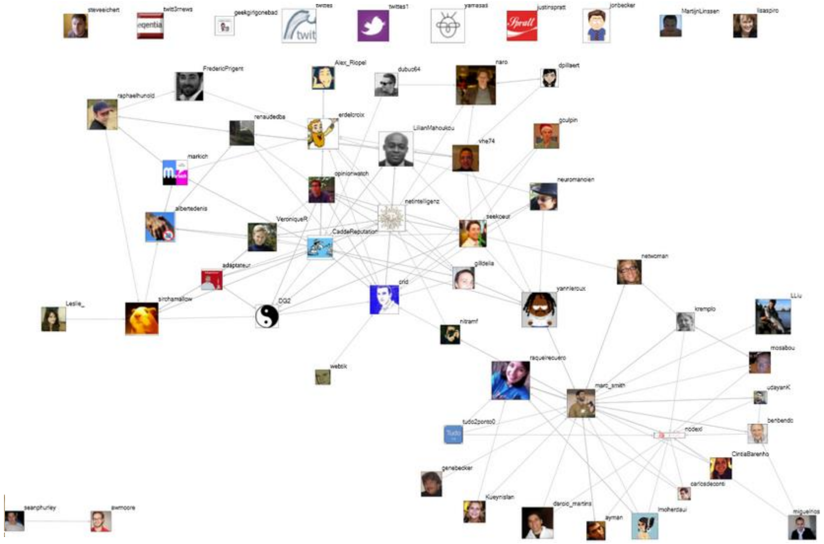
Narayanan-Shmatikov attack (IEEE S&P 2009)

- A. Narayanan and V. Shmatikov, *De-anonymizing Social Networks*, **IEEE S&P 2009**.

Narayanan-Shmatikov attack (IEEE S&P 2009)

- A. Narayanan and V. Shmatikov, *De-anonymizing Social Networks*, **IEEE S&P 2009**.
- Anonymized data: Twitter (crawled in late 2007)
  - A microblogging service
  - 224K users, 8.5M edges

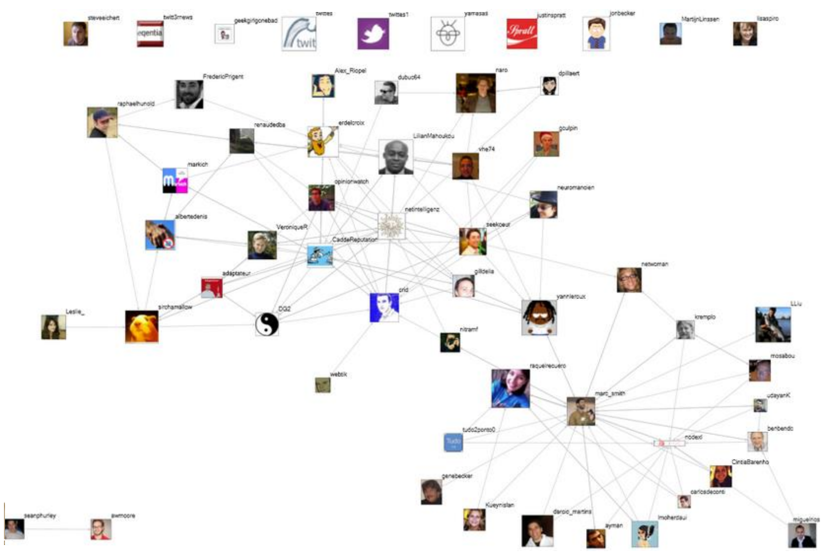
Twitter



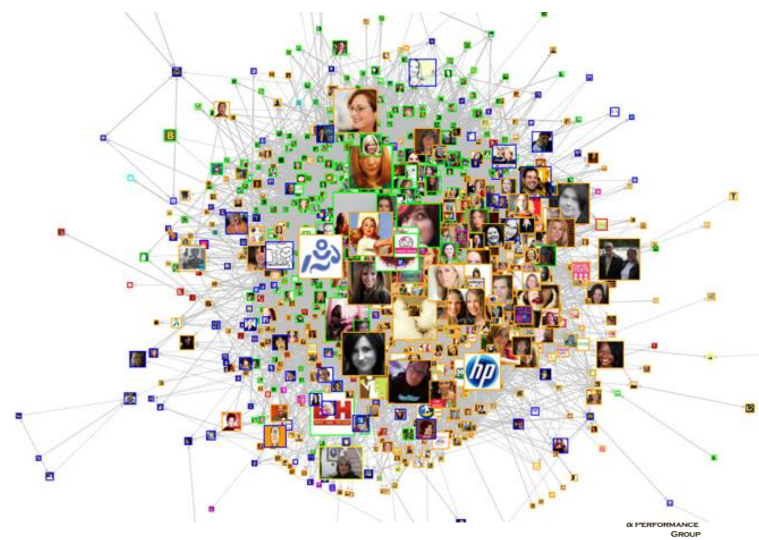
Narayanan-Shmatikov attack (IEEE S&P 2009)

- A. Narayanan and V. Shmatikov, *De-anonymizing Social Networks*, **IEEE S&P 2009**.
- Anonymized data: Twitter (crawled in late 2007)
  - A microblogging service
  - 224K users, 8.5M edges
- Auxiliary data: Flicker (crawled in late 2007/early 2008)
  - A photo-sharing service
  - 3.3M users, 53M edges

Twitter



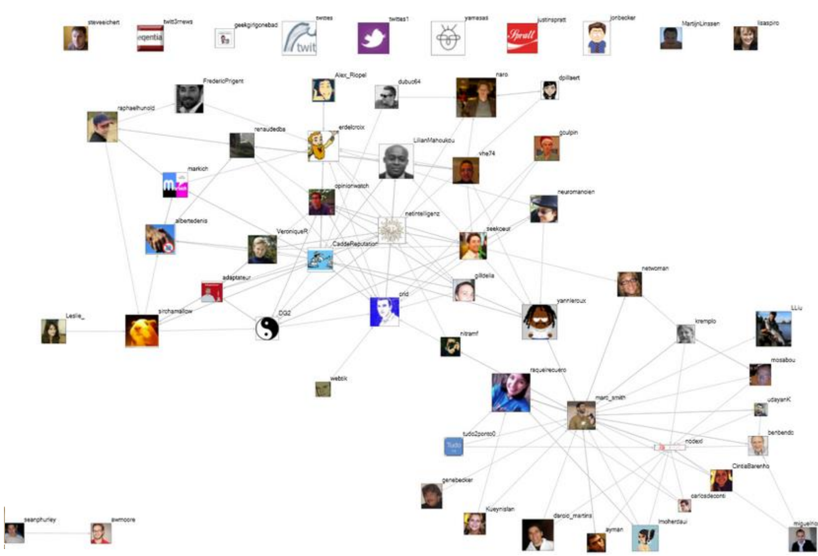
Flicker



Narayanan-Shmatikov attack (IEEE S&P 2009)

- A. Narayanan and V. Shmatikov, *De-anonymizing Social Networks*, **IEEE S&P 2009**.
- Anonymized data: Twitter (crawled in late 2007)
  - A microblogging service
  - 224K users, 8.5M edges
- Auxiliary data: Flicker (crawled in late 2007/early 2008)
  - A photo-sharing service
  - 3.3M users, 53M edges
- Result: 30.8% of the users are successfully de-anonymized

Twitter



Heuristics

Eccentricity
Edge directionality
Node degree
Revisiting nodes
Reverse match

User mapping

Flicker

Motivation

- Question 1: *Why social networks are vulnerable to structure based de-anonymization attacks?*

Motivation

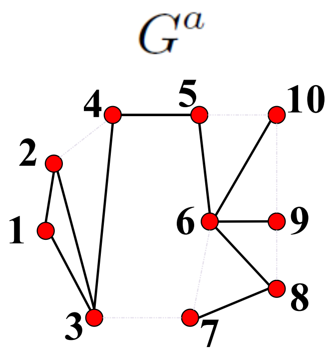
- Question 1: *Why social networks are vulnerable to structure based de-anonymization attacks?*
- Question 2: *How de-anonymizable a social network is?*

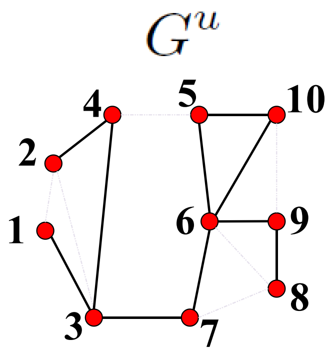
Motivation

- Question 1: *Why social networks are vulnerable to structure based de-anonymization attacks?*
- Question 2: *How de-anonymizable a social network is?*
- Question 3: *How many users within a social network can be successfully de-anonymized?*

System Model

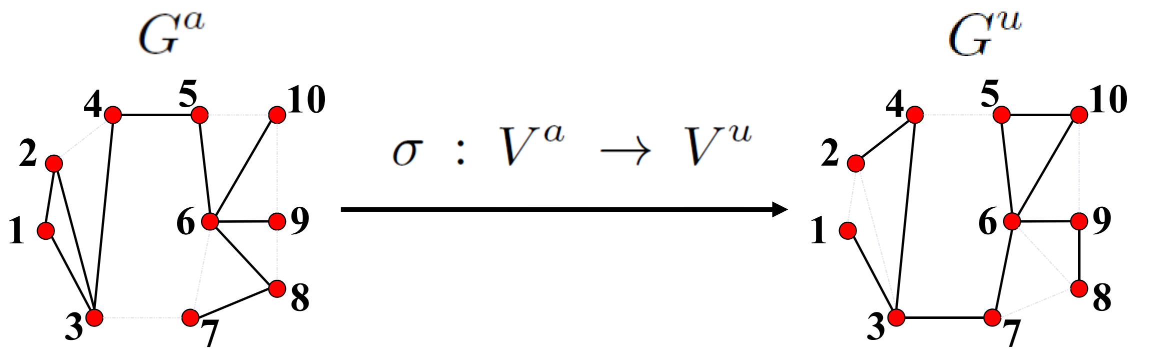
- Anonymized Data (G^a)
- Auxiliary Data (G^u)





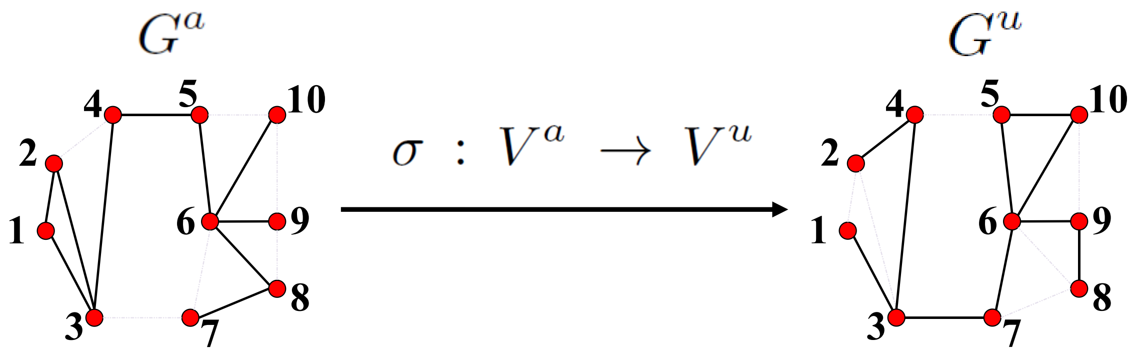
System Model

- Anonymized Data (G^a)
- Auxiliary Data (G^u)
- De-anonymization (σ)



System Model

- Anonymized Data (G^a)
- Auxiliary Data (G^u)
- De-anonymization (σ)
- Measurement(Δ_σ)



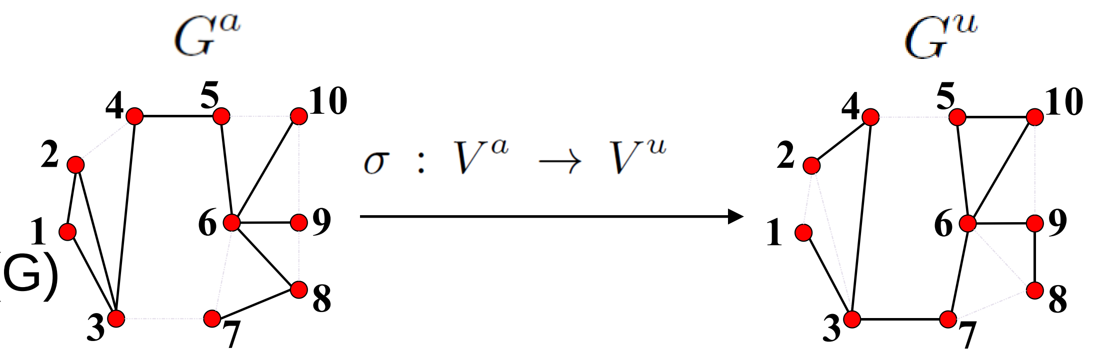
De-anonymization Error (DE) on a user mapping $(i, i') \in \sigma$

$$\Delta_{\sigma:(i,j)}(s) = \left|\sigma\left(E_i^a(S^a)\right)\backslash E_j^u(S^u)\right| + \left|\sigma^{-1}\left(E_j^u(S^u)\right)\backslash E_i^a(S^a)\right|$$

$$\Delta_\sigma(\mathrm{S}) = \sum_{(i,j)\in\sigma} \Delta_{\sigma:(i,j)}(S)$$

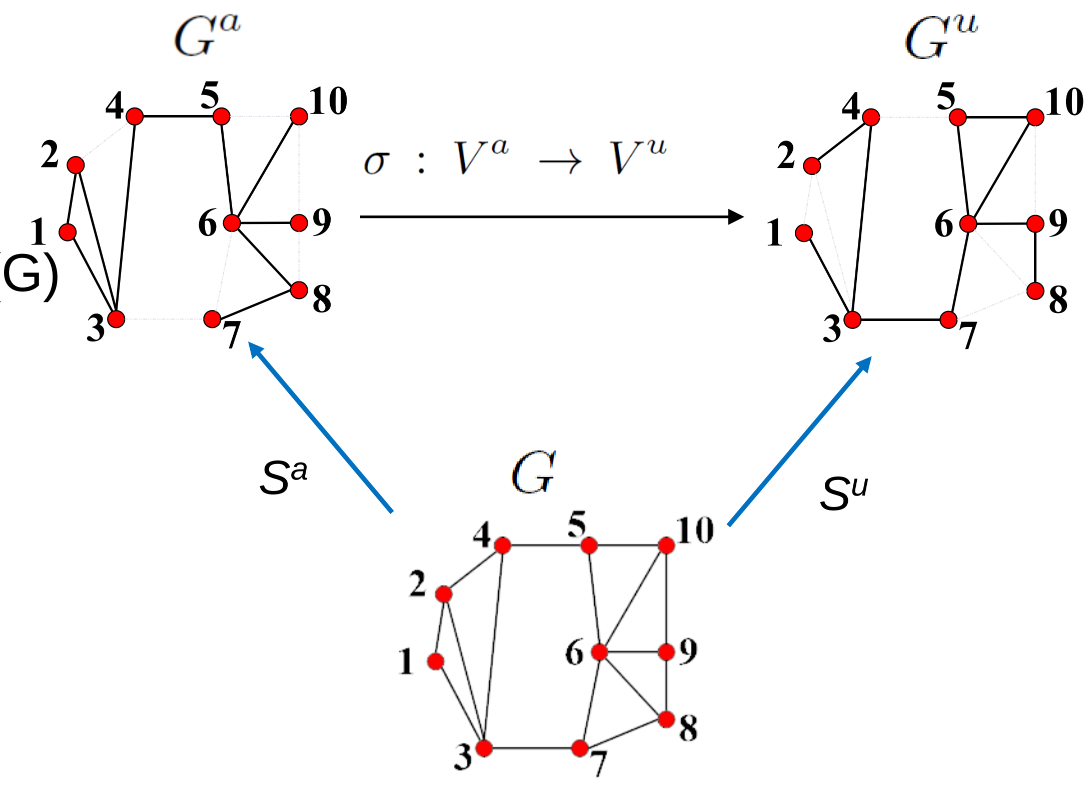
System Model

- Anonymized Data (G^a)
- Auxiliary Data (G^u)
- De-anonymization (σ)
- Conceptual Underlying Graph (G)



System Model

- Anonymized Data (G^a)
- Auxiliary Data (G^u)
- De-anonymization (σ)
- Conceptual Underlying Graph (G)



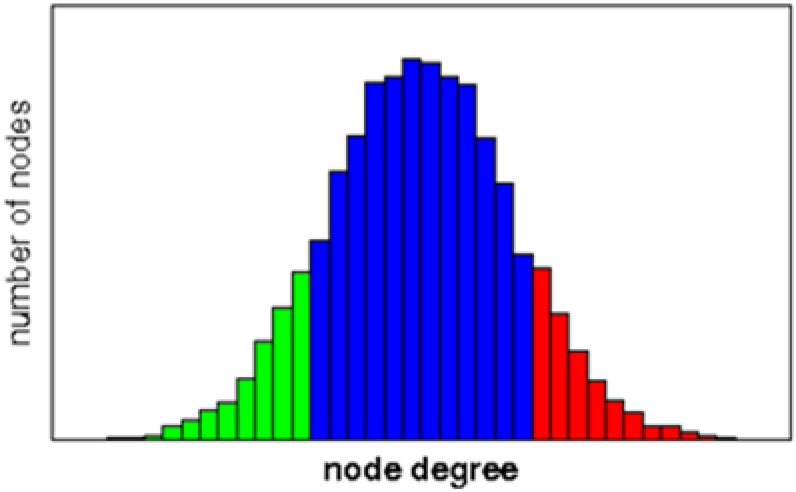
De-anonymization Quantification

- Conceptual Underlying Graph:
  - Erdős–Rényi(ER) model

De-anonymization Quantification

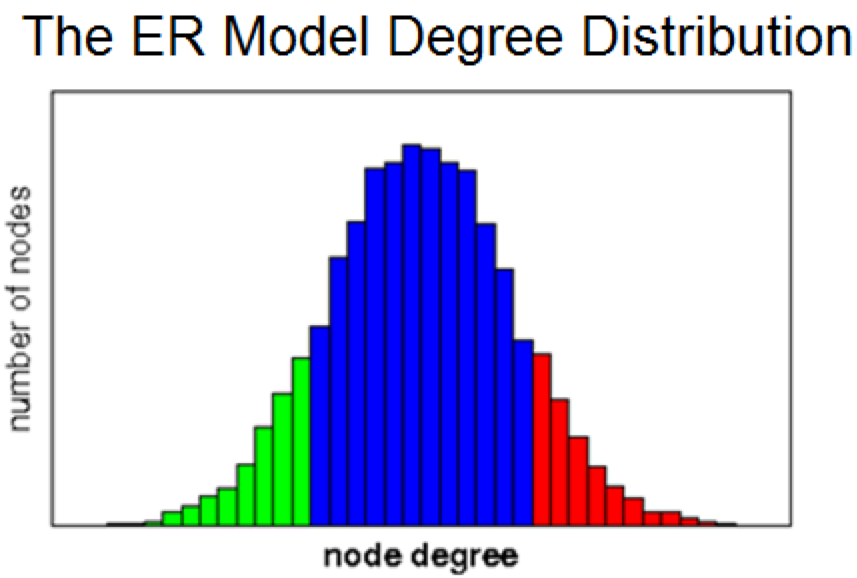
- Conceptual Underlying Graph:
  - Erdős–Rényi(ER) model

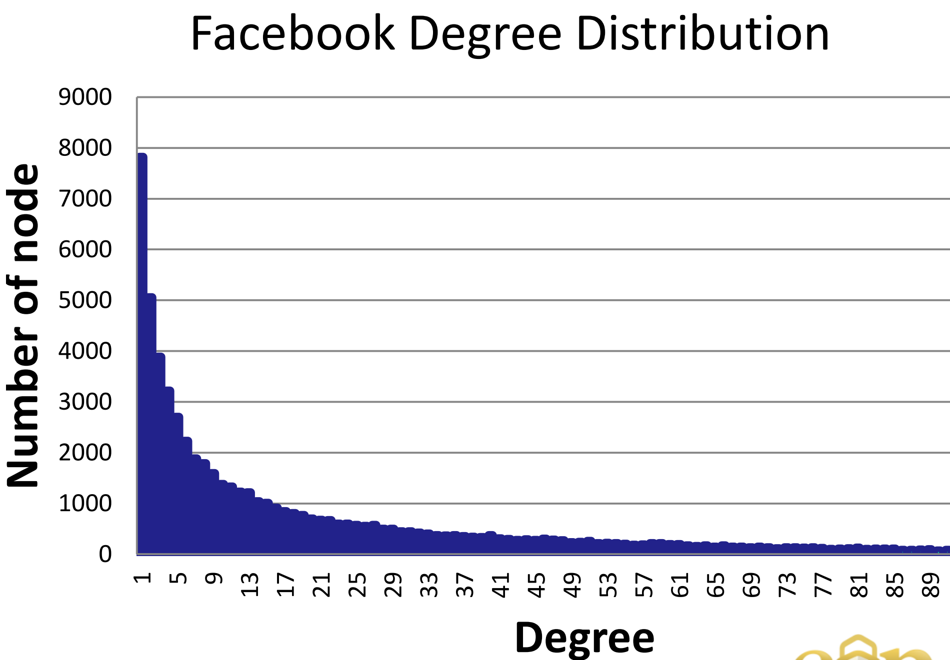
The ER Model Degree Distribution



De-anonymization Quantification

- Conceptual Underlying Graph:
  - Erdős–Rényi(ER) model





De-anonymization Quantification

- Conceptual Underlying Graph:
  - Erdős–Rényi(ER) model
  - General model

De-anonymization Quantification

- Conceptual Underlying Graph:
  - Erdős–Rényi(ER) model
  - General model
- Quantifications
  - Mathematically shown the conditions needed

De-anonymization Quantification

- Conceptual Underlying Graph:
  - Erdős–Rényi(ER) model
  - General model
- Quantifications
  - Mathematically shown the conditions needed
    - Seed based perfect *de-anonymization*
    - *Structure based* perfect *de-anonymization*
    - *Error tolerant de-anonymization*

Evaluation

- Datasets

TABLE I. DATASET STATISTICS.

Name	n	m	ρ	$\bar{d}$
Hyves	1,402,673	2,777,419	2.82E-06	3.96
Douban	154,908	327,162	2.73E-05	4.22
Friendster	5,689,498	14,067,887	8.69E-07	4.95
YouTube	1,138,499	2,990,443	4.61E-06	5.25
Flixster	2,523,386	7,918,801	2.49E-06	6.28
Last.fm	1,191,812	4,519,340	6.36E-06	7.58
FB-NO-wall	45,813	183,412	1.75E-04	8.01
Gowalla	196,591	950,327	4.92E-05	9.70
Foursquare	639,014	3,214,986	1.57E-05	10.06
Enron	33,696	180,811	3.19E-04	10.73
Skitter	1,694,616	11,094,209	7.73E-06	13.09
Slashdot	82,168	582,533	1.73E-04	14.18
Digg	771,229	5,907,413	1.99E-05	15.32
LiveJournal	4,843,953	43,362,750	3.70E-06	17.90
HepPh	11,204	117,649	1.87E-03	21.00
AstroPh	17,903	197,031	1.23E-03	22.01
FB-NO-links	63,731	817,090	4.02E-04	25.64
Pokec	1,632,803	22,301,964	1.67E-05	27.32
BlogCatalog	97,884	1,668,647	3.48E-04	34.10
Google+	4,692,671	90,751,480	8.24E-06	38.68
Livemocha	104,103	2,193,083	4.05E-04	42.13
Twitter	456,293	12,508,272	1.20E-04	54.83
Orkut	3,072,441	117,185,083	2.48E-05	76.28
Flickr	80,513	5,899,882	1.82E-03	146.56

Evaluation

- Datasets
  - Number of nodes (n)

TABLE I. DATASET STATISTICS.

Name	n	m	ρ	$\bar{d}$
Hyves	1,402,673	2,777,419	2.82E-06	3.96
Douban	154,908	327,162	2.73E-05	4.22
Friendster	5,689,498	14,067,887	8.69E-07	4.95
YouTube	1,138,499	2,990,443	4.61E-06	5.25
Flixster	2,523,386	7,918,801	2.49E-06	6.28
Last.fm	1,191,812	4,519,340	6.36E-06	7.58
FB-NO-wall	45,813	183,412	1.75E-04	8.01
Gowalla	196,591	950,327	4.92E-05	9.70
Foursquare	639,014	3,214,986	1.57E-05	10.06
Enron	33,696	180,811	3.19E-04	10.73
Skitter	1,694,616	11,094,209	7.73E-06	13.09
Slashdot	82,168	582,533	1.73E-04	14.18
Digg	771,229	5,907,413	1.99E-05	15.32
LiveJournal	4,843,953	43,362,750	3.70E-06	17.90
HepPh	11,204	117,649	1.87E-03	21.00
AstroPh	17,903	197,031	1.23E-03	22.01
FB-NO-links	63,731	817,090	4.02E-04	25.64
Pokec	1,632,803	22,301,964	1.67E-05	27.32
BlogCatalog	97,884	1,668,647	3.48E-04	34.10
Google+	4,692,671	90,751,480	8.24E-06	38.68
Livemocha	104,103	2,193,083	4.05E-04	42.13
Twitter	456,293	12,508,272	1.20E-04	54.83
Orkut	3,072,441	117,185,083	2.48E-05	76.28
Flickr	80,513	5,899,882	1.82E-03	146.56

Evaluation

- Datasets
  - Number of nodes (n)
  - Number of edges(m)

TABLE I. DATASET STATISTICS.

Name	n	m	ρ	$\overline{d}$
Hyves	1,402,673	2,777,419	2.82E-06	3.96
Douban	154,908	327,162	2.73E-05	4.22
Friendster	5,689,498	14,067,887	8.69E-07	4.95
YouTube	1,138,499	2,990,443	4.61E-06	5.25
Flixster	2,523,386	7,918,801	2.49E-06	6.28
Last.fm	1,191,812	4,519,340	6.36E-06	7.58
FB-NO-wall	45,813	183,412	1.75E-04	8.01
Gowalla	196,591	950,327	4.92E-05	9.70
Foursquare	639,014	3,214,986	1.57E-05	10.06
Enron	33,696	180,811	3.19E-04	10.73
Skitter	1,694,616	11,094,209	7.73E-06	13.09
Slashdot	82,168	582,533	1.73E-04	14.18
Digg	771,229	5,907,413	1.99E-05	15.32
LiveJournal	4,843,953	43,362,750	3.70E-06	17.90
HepPh	11,204	117,649	1.87E-03	21.00
AstroPh	17,903	197,031	1.23E-03	22.01
FB-NO-links	63,731	817,090	4.02E-04	25.64
Pokec	1,632,803	22,301,964	1.67E-05	27.32
BlogCatalog	97,884	1,668,647	3.48E-04	34.10
Google+	4,692,671	90,751,480	8.24E-06	38.68
Livemocha	104,103	2,193,083	4.05E-04	42.13
Twitter	456,293	12,508,272	1.20E-04	54.83
Orkut	3,072,441	117,185,083	2.48E-05	76.28
Flickr	80,513	5,899,882	1.82E-03	146.56

Georgia Institute of Technology

Evaluation

- Datasets
  - Number of nodes (n)
  - Number of edges(m)
  - Graph density (ρ)

TABLE I. DATASET STATISTICS.

Name	n	m	ρ	$\bar{d}$
Hyves	1,402,673	2,777,419	2.82E-06	3.96
Douban	154,908	327,162	2.73E-05	4.22
Friendster	5,689,498	14,067,887	8.69E-07	4.95
YouTube	1,138,499	2,990,443	4.61E-06	5.25
Flixster	2,523,386	7,918,801	2.49E-06	6.28
Last.fm	1,191,812	4,519,340	6.36E-06	7.58
FB-NO-wall	45,813	183,412	1.75E-04	8.01
Gowalla	196,591	950,327	4.92E-05	9.70
Foursquare	639,014	3,214,986	1.57E-05	10.06
Enron	33,696	180,811	3.19E-04	10.73
Skitter	1,694,616	11,094,209	7.73E-06	13.09
Slashdot	82,168	582,533	1.73E-04	14.18
Digg	771,229	5,907,413	1.99E-05	15.32
LiveJournal	4,843,953	43,362,750	3.70E-06	17.90
HepPh	11,204	117,649	1.87E-03	21.00
AstroPh	17,903	197,031	1.23E-03	22.01
FB-NO-links	63,731	817,090	4.02E-04	25.64
Pokec	1,632,803	22,301,964	1.67E-05	27.32
BlogCatalog	97,884	1,668,647	3.48E-04	34.10
Google+	4,692,671	90,751,480	8.24E-06	38.68
Livemocha	104,103	2,193,083	4.05E-04	42.13
Twitter	456,293	12,508,272	1.20E-04	54.83
Orkut	3,072,441	117,185,083	2.48E-05	76.28
Flickr	80,513	5,899,882	1.82E-03	146.56

Georgia Institute of Technology

Evaluation

- Datasets
  - Number of nodes (n)
  - Number of edges(m)
  - Graph density (ρ)
  - Average degree ($\bar{d}$)

TABLE I. DATASET STATISTICS.

Name	n	m	ρ	$\bar{d}$
Hyves	1,402,673	2,777,419	2.82E-06	3.96
Douban	154,908	327,162	2.73E-05	4.22
Friendster	5,689,498	14,067,887	8.69E-07	4.95
YouTube	1,138,499	2,990,443	4.61E-06	5.25
Flixster	2,523,386	7,918,801	2.49E-06	6.28
Last.fm	1,191,812	4,519,340	6.36E-06	7.58
FB-NO-wall	45,813	183,412	1.75E-04	8.01
Gowalla	196,591	950,327	4.92E-05	9.70
Foursquare	639,014	3,214,986	1.57E-05	10.06
Enron	33,696	180,811	3.19E-04	10.73
Skitter	1,694,616	11,094,209	7.73E-06	13.09
Slashdot	82,168	582,533	1.73E-04	14.18
Digg	771,229	5,907,413	1.99E-05	15.32
LiveJournal	4,843,953	43,362,750	3.70E-06	17.90
HepPh	11,204	117,649	1.87E-03	21.00
AstroPh	17,903	197,031	1.23E-03	22.01
FB-NO-links	63,731	817,090	4.02E-04	25.64
Pokec	1,632,803	22,301,964	1.67E-05	27.32
BlogCatalog	97,884	1,668,647	3.48E-04	34.10
Google+	4,692,671	90,751,480	8.24E-06	38.68
Livemocha	104,103	2,193,083	4.05E-04	42.13
Twitter	456,293	12,508,272	1.20E-04	54.83
Orkut	3,072,441	117,185,083	2.48E-05	76.28
Flickr	80,513	5,899,882	1.82E-03	146.56

Evaluation

- Datasets Types
  - Social media (e.g. LiveJournal 4 million nodes, 43 million edges)

TABLE I. DATASET STATISTICS.

Name	n	m	ρ	$\bar{d}$
Hyves	1,402,673	2,777,419	2.82E-06	3.96
Douban	154,908	327,162	2.73E-05	4.22
Friendster	5,689,498	14,067,887	8.69E-07	4.95
YouTube	1,138,499	2,990,443	4.61E-06	5.25
Flixster	2,523,386	7,918,801	2.49E-06	6.28
Last.fm	1,191,812	4,519,340	6.36E-06	7.58
FB-NO-wall	45,813	183,412	1.75E-04	8.01
Gowalla	196,591	950,327	4.92E-05	9.70
Foursquare	639,014	3,214,986	1.57E-05	10.06
Enron	33,696	180,811	3.19E-04	10.73
Skitter	1,694,616	11,094,209	7.73E-06	13.09
Slashdot	82,168	582,533	1.73E-04	14.18
Digg	771,229	5,907,413	1.99E-05	15.32
LiveJournal	4,843,953	43,362,750	3.70E-06	17.90
HepPh	11,204	117,649	1.87E-03	21.00
AstroPh	17,903	197,031	1.23E-03	22.01
FB-NO-links	63,731	817,090	4.02E-04	25.64
Pokec	1,632,803	22,301,964	1.67E-05	27.32
BlogCatalog	97,884	1,668,647	3.48E-04	34.10
Google+	4,692,671	90,751,480	8.24E-06	38.68
Livemocha	104,103	2,193,083	4.05E-04	42.13
Twitter	456,293	12,508,272	1.20E-04	54.83
Orkut	3,072,441	117,185,083	2.48E-05	76.28
Flickr	80,513	5,899,882	1.82E-03	146.56

Evaluation

- Datasets Types
  - Social media (e.g. LiveJournal 4 million nodes, 43 million edges).
  - Email network (e.g. Enron 33 thousand nodes, 180 thousand edges).

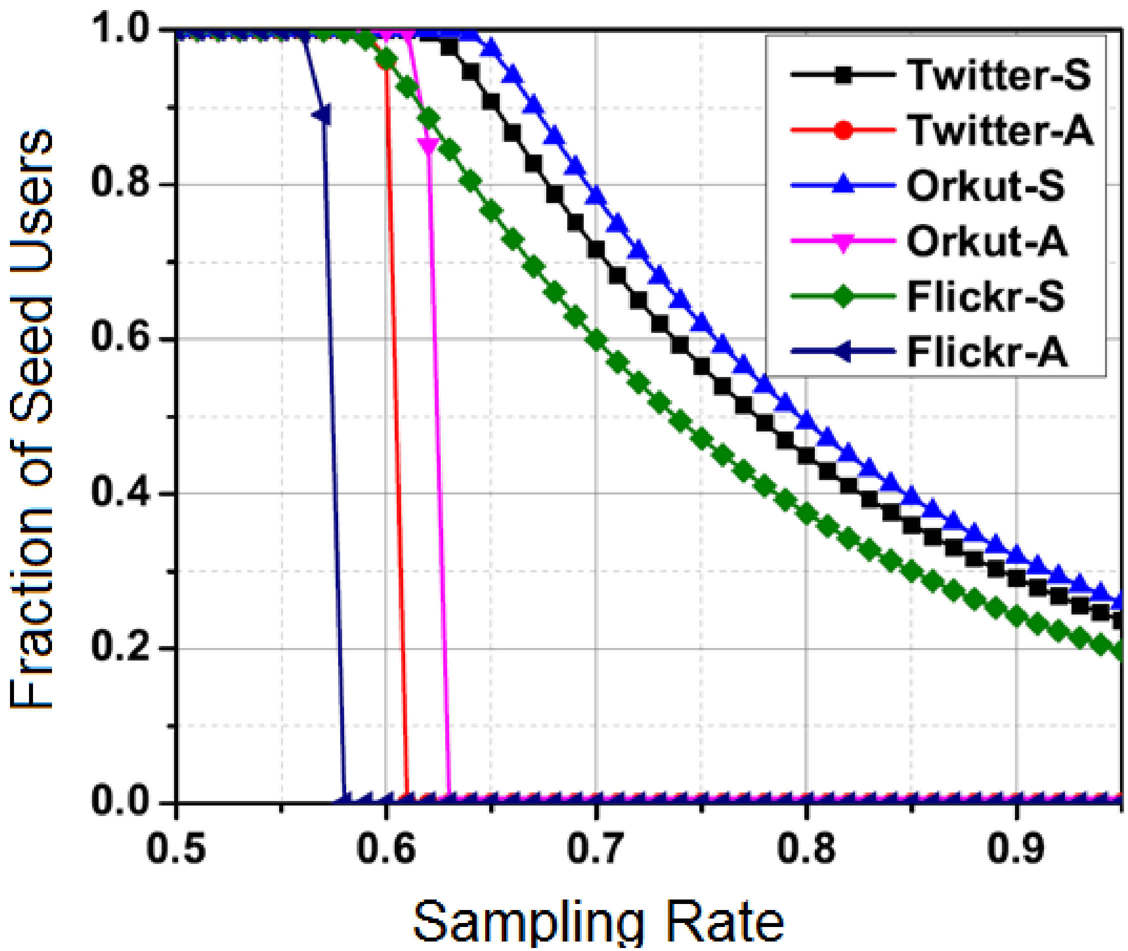
TABLE I. DATASET STATISTICS.

Name	n	m	ρ	$\bar{d}$
Hyves	1,402,673	2,777,419	2.82E-06	3.96
Douban	154,908	327,162	2.73E-05	4.22
Friendster	5,689,498	14,067,887	8.69E-07	4.95
YouTube	1,138,499	2,990,443	4.61E-06	5.25
Flixster	2,523,386	7,918,801	2.49E-06	6.28
Last.fm	1,191,812	4,519,340	6.36E-06	7.58
FB-NO-wall	45,813	183,412	1.75E-04	8.01
Gowalla	196,591	950,327	4.92E-05	9.70
Foursquare	639,014	3,214,986	1.57E-05	10.06
Enron	33,696	180,811	3.19E-04	10.73
Skitter	1,694,616	11,094,209	7.73E-06	13.09
Slashdot	82,168	582,533	1.73E-04	14.18
Digg	771,229	5,907,413	1.99E-05	15.32
LiveJournal	4,843,953	43,362,750	3.70E-06	17.90
HepPh	11,204	117,649	1.87E-03	21.00
AstroPh	17,903	197,031	1.23E-03	22.01
FB-NO-links	63,731	817,090	4.02E-04	25.64
Pokec	1,632,803	22,301,964	1.67E-05	27.32
BlogCatalog	97,884	1,668,647	3.48E-04	34.10
Google+	4,692,671	90,751,480	8.24E-06	38.68
Livemocha	104,103	2,193,083	4.05E-04	42.13
Twitter	456,293	12,508,272	1.20E-04	54.83
Orkut	3,072,441	117,185,083	2.48E-05	76.28
Flickr	80,513	5,899,882	1.82E-03	146.56

Georgia Institute of Technology

Evaluation

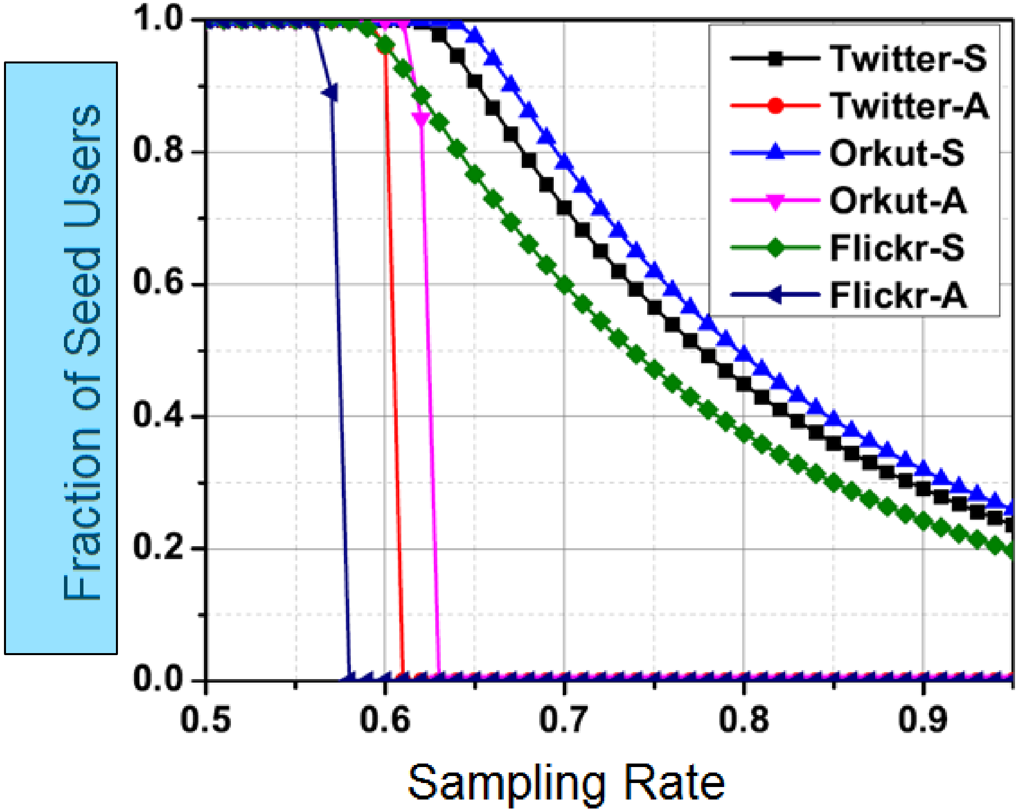
Perfect De-anonymization Analysis on seed users versus sampling rate



Georgia Institute of Technology

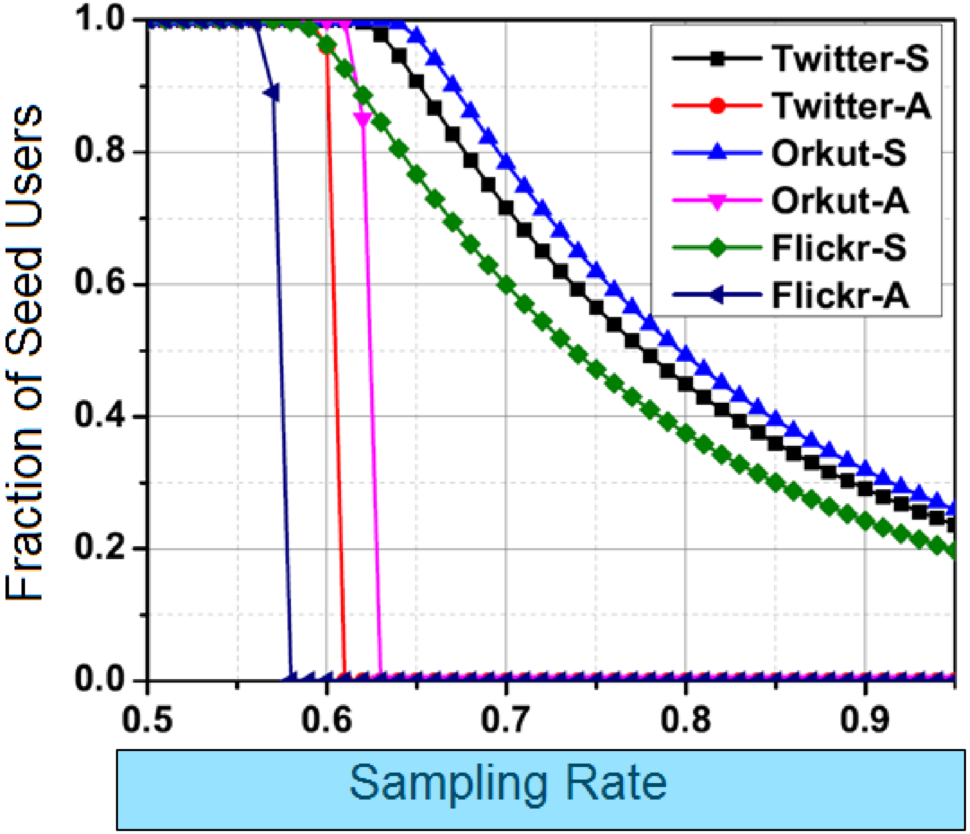
Evaluation

- The y-axis: number of pre-determined users needed.



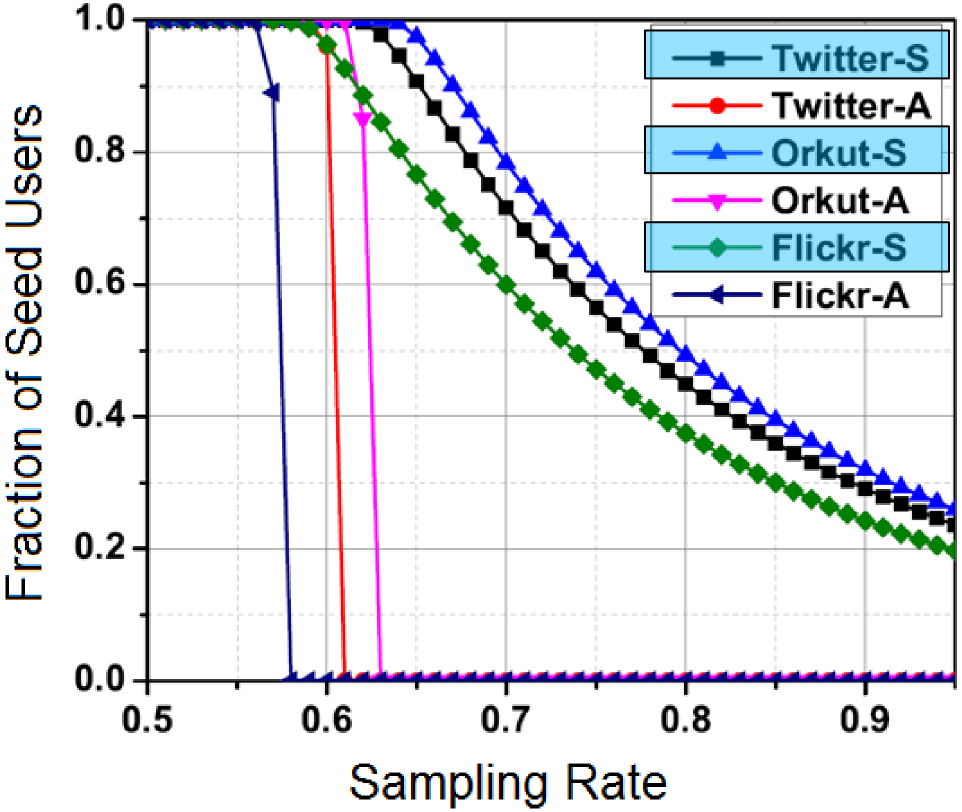
Evaluation

- The y-axis: number of pre-determined users needed.
- The x-axis: sampling rate needed.



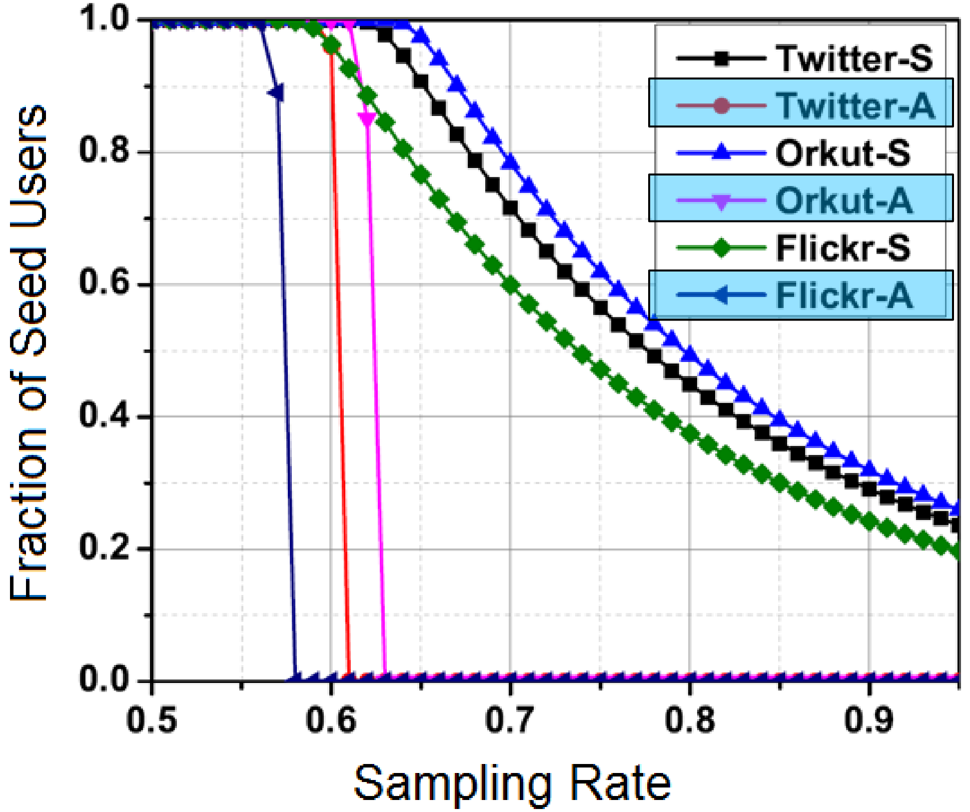
Evaluation

- The y-axis: number of pre-determined users needed.
- The x-axis: sampling rate needed.
- *-S: de-anonymization with only seed information.



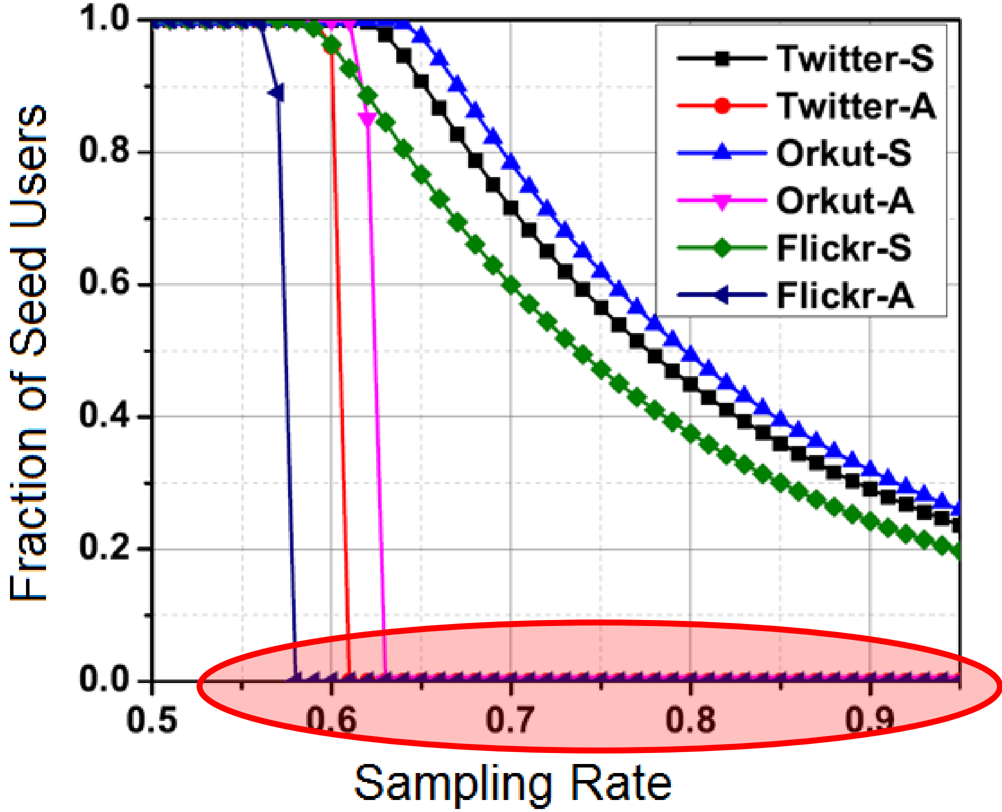
Evaluation

- The y-axis: number of pre-determined users needed.
- The x axis: sampling rate needed.
- *-S: de-anonymization with only seed information.
- *-A: de-anonymization with seed and structural information



Evaluation

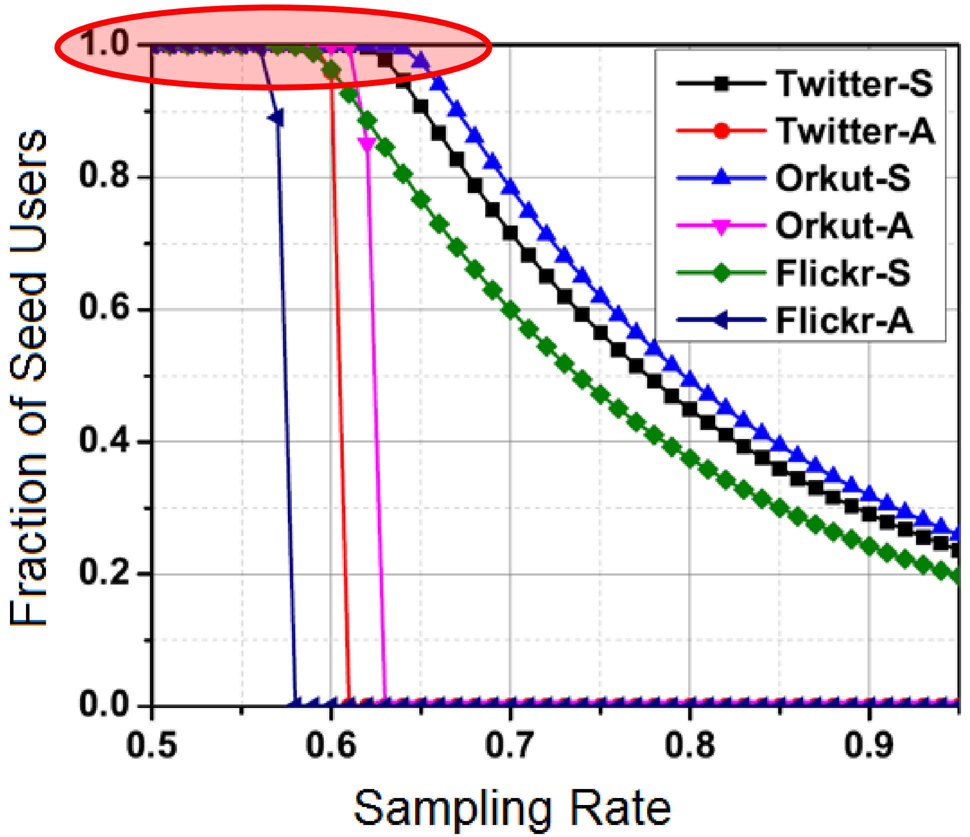
- When sampling rate is large no seeds needed for perfect de-anonymization.



Evaluation

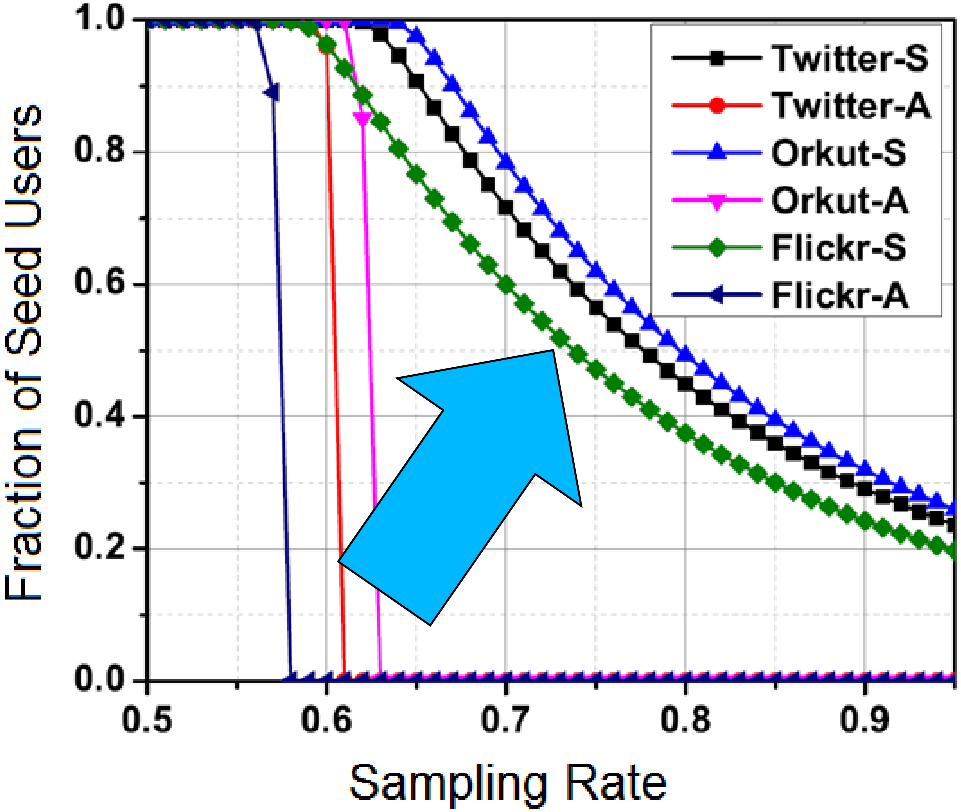
- When sampling rate is large no seeds needed for perfect de-anonymization.

- When sampling rate is small all users need to be seed users for perfect de-anonymization.



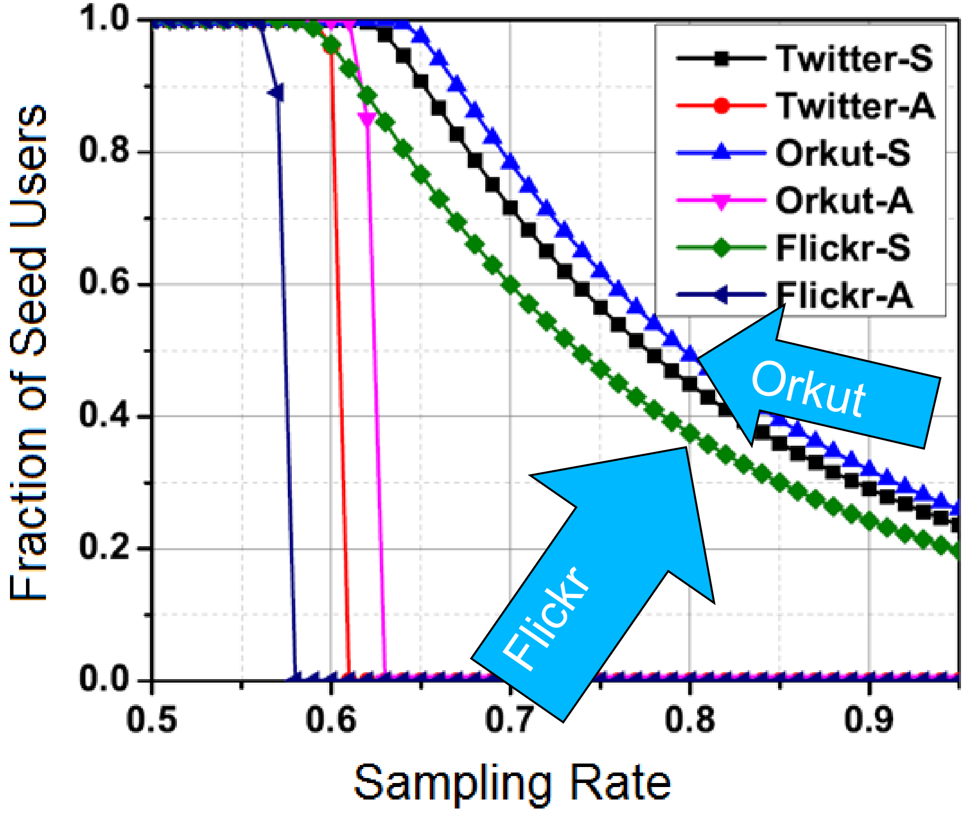
Evaluation

- Using only seed information, fewer seed users are required with lower sampling rate.



Evaluation

- Graphs with higher density requires fewer seed users.
  - Orkut density of 2.48E-5
  - Flicker density of 1.82E-3



Conclusion and Future Works

- Conclusion
  - Theoretical quantification with no degree distribution requirement.
  - Large-scale de-anonymizability evaluation.
- Future works
  - New mathematical model
  - Defense techniques

Thank you!

Questions?

Acknowledgement

We thank the anonymous reviewers very much for their valuable comments!