Big Data Privacy

Gábor György Gulyás
gulyas.info // @GulyasGG
Laboratory of Cryptography and System Security (CrySyS)
Budapest University of Technology and Economics
www.crysys.hu

Overview and goal of the talk

- Privacy in large datasets
- Possible privacy solutions
- Structural de-anonymization in social networks
  - Attacks
  - Defenses
  - Next generation of attacks
- Conclusion

PRIVACY IN LARGE DATASETS

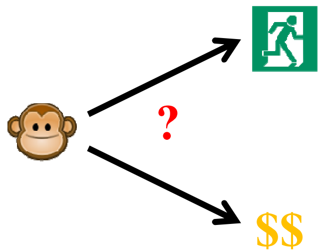
‘Natural’ sources of big data in (social) technology (e.g.)

Social networks & media

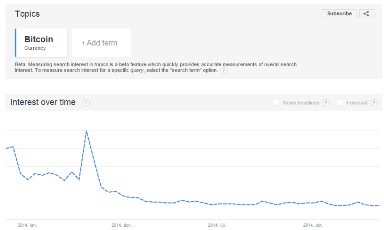
Recommender systems

Web tracking dbs (profiling)

Doc indexing & search



Predicting user behavior

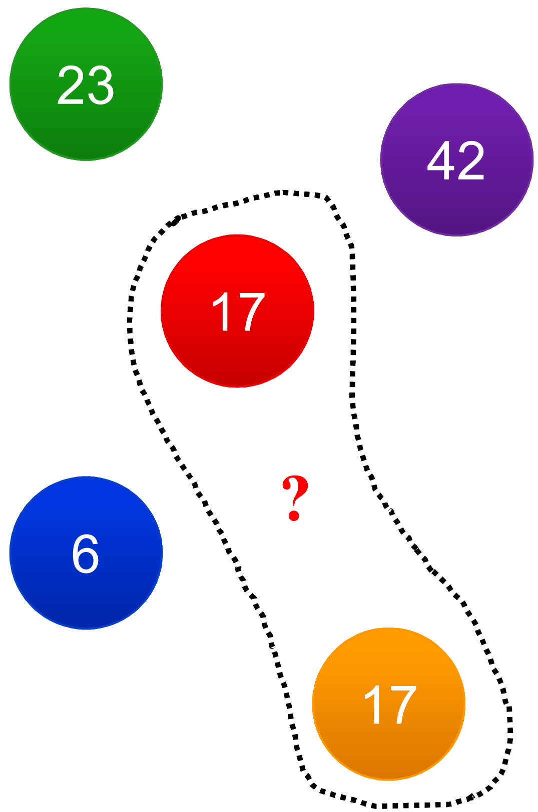


Exposing trends

What is anonymity?

- **One is anonymous,** who can not be identified within a set of subjects.
  - **Anonymity set!**
  - Identifying attributes are the same
  - Point of view can be local or global
  - Determined by the attacker model

Participants and their age



The A_1 anonymity set:

Bob is the one who

is 17 year old. Which one?

How identifiable are we?

Sweeney, 1990

87% of US population is identifiable by (216 million of 248 million):
`{5 digit ZIP, gender, date of birth}`

Revisiting study: 64% of US population is identifiable by:
`{ZIP-code, gender, date of birth}`

Golle, 2000

How identifiable are we? (2)

Work-home location pairs as identifying information (US):

- avg. 1500 person / location cells
- 5% totally identifiable.
- avg. anonymity set size is ca. 20

Location based services?!

Golle & Partridge, 2009

How identifiable are we? (3)

Anonymized NetFlix dataset

```
28 89 40 10 e5 f9 41 07 3f 8c
ee 09 3d 71 54 85 83 43 4e 04
1f 64 71 a5 14 ca dd 95 4e bb
2a 35 dc 89 f8 99 dd 56 ca 42
1f 93 f5 d1 dc f1 b0 34 e8 b1
f6 43 5a 28 49 5c f3 40 fa ba
aa cf bc 49 80 26 71 29 66 f6
5a d9 10 7a b8 27 ea 74 6f 72
50 b3 ce 8b ee d9 65 92 17 f5
01 89 2c a0 c4 60 53 88 a1 e1
```

Public IMDb ratings

```
2d 1d ed d1 39 b8 f9 fb 20 53
de 14 96 cb a3 0b 80 52 ff 52
39 65 84 61 d3 50 a7 d3 aa 80
93 cc ca 4f 8e 3a 47 0a e6 de
fa 05 64 be 4c 59 0e 04 91 85
4c ba ba 30 91 a9 34 47 0d 2e
0f 51 26 23 fd 5c 43 1e e5 9f
37 8a d4 7d 64 0a 8a 60 e1 26
d0 31 38 a0 eb 7d bd 52 2a a6
8a 30 0a c3 86 dd 4d 16 20 76
```

Netflix vs. IMDb

- rarely used features are identifying
- only 8 ratings identify 99% of users (2 erroneous),
- dates within a 2 week timeframe

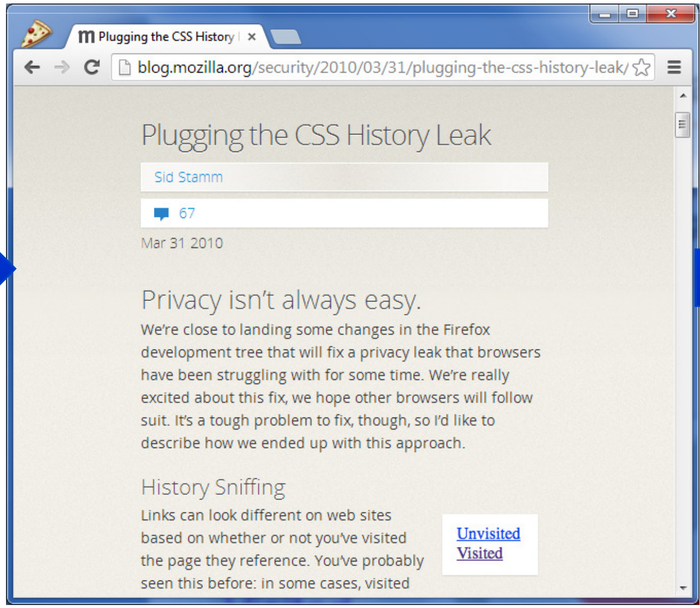
Narayanan & Shmatikov, 2008

How identifiable are we? (4)

An experiment on Xing indicates that **group memberships** are identifying:

- ~8m users at the time
- ca. 42% uniquely identified
- extremely small anonymity sets: 2.912 collisions for 90% of users!

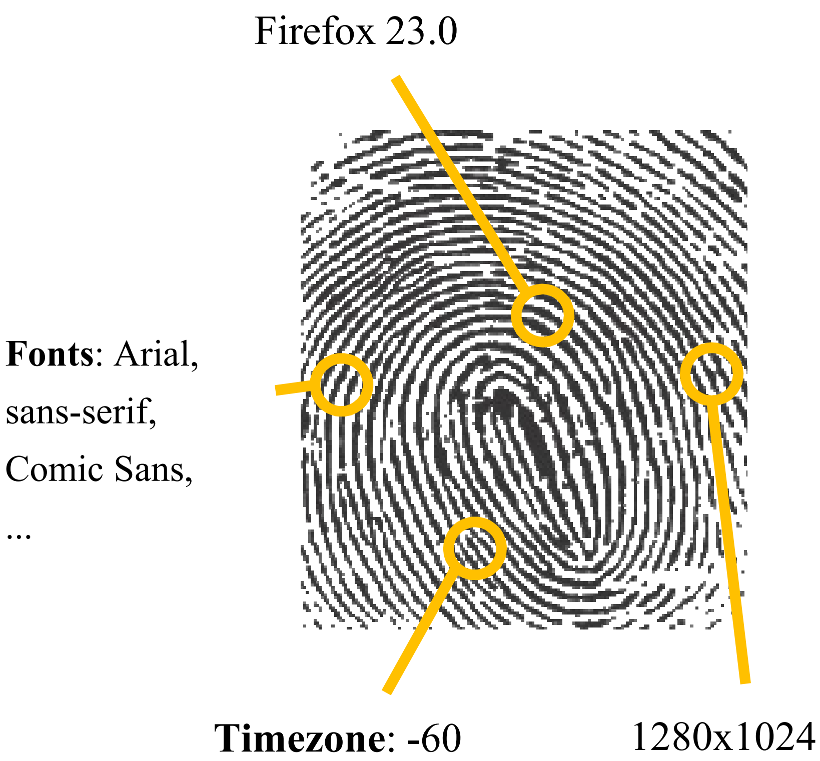
Univisted
Visited



Univisted
Visited

Wondracek et al., 2010

How identifiable are we? (5)



- Fingerprinting evolves:
  - 2010: Browser fingerprint (e.g., accuracy: 94.2%)
  - 2011: System fingerprint (works well on Windows)
  - 2012: Connecting personal devices
  - Future: biometric fingerprinting?
- Billions of (device) fingerprints in databases
  - Based on simple characteristics

Eckersely, 2010
Boda et al., 2011

How identifiable are we? (6)

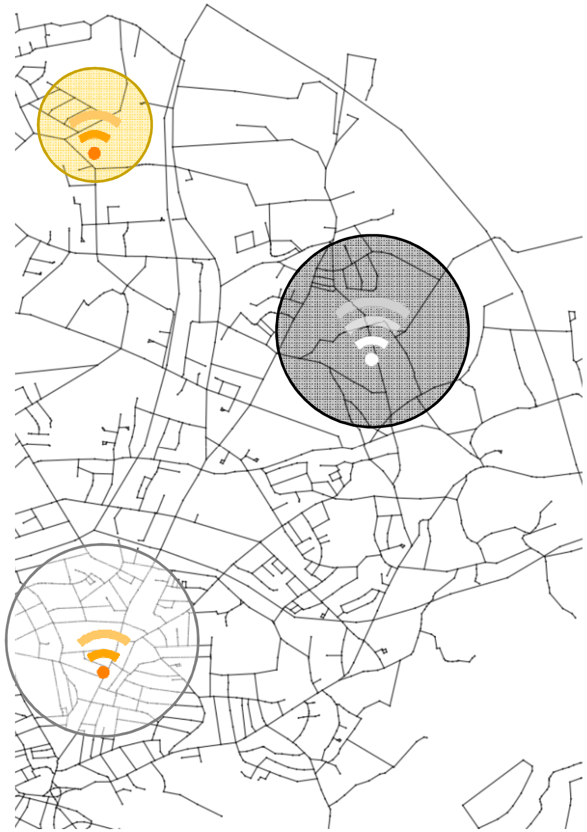
- Unstructured data!
- Writing style can be structured:
  - e.g., inspecting the relative frequency of ‘since’ and ‘because’
  - many of these can enable stylometric profiling

Results on in searching the author of a few posts:
- On 100,000 blogs, cross-context validation
- 20% of correct identification (of 3 posts)
- Improvements:
  - Manual inspection of top 20 results → 35% success rate
  - 30-35% corr. id. with 20 posts

Narayanan et al., 2012

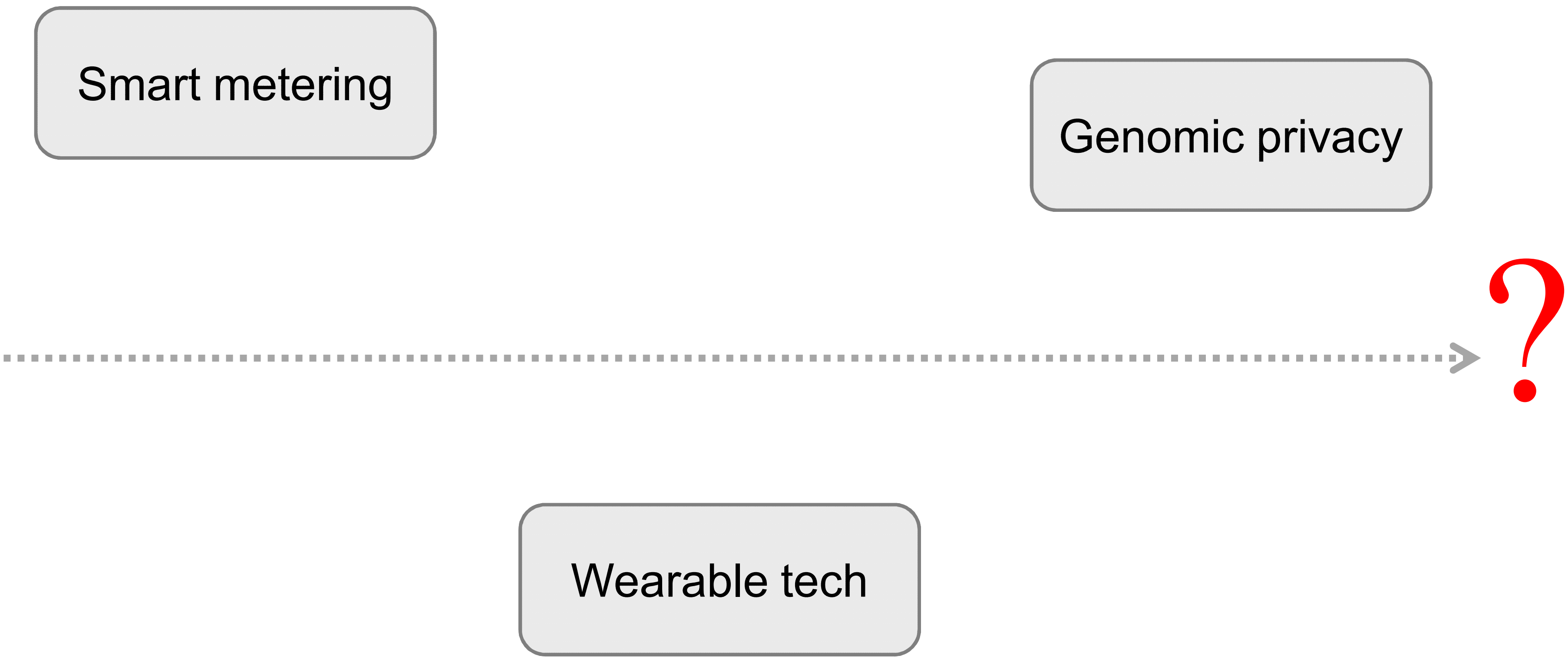
How identifiable are we? (7)

Network alignment on **temporal location information and social networks**
- with ca. 80% recall in small nets (2012)
- up to 84% recall in ~200k users (2014)



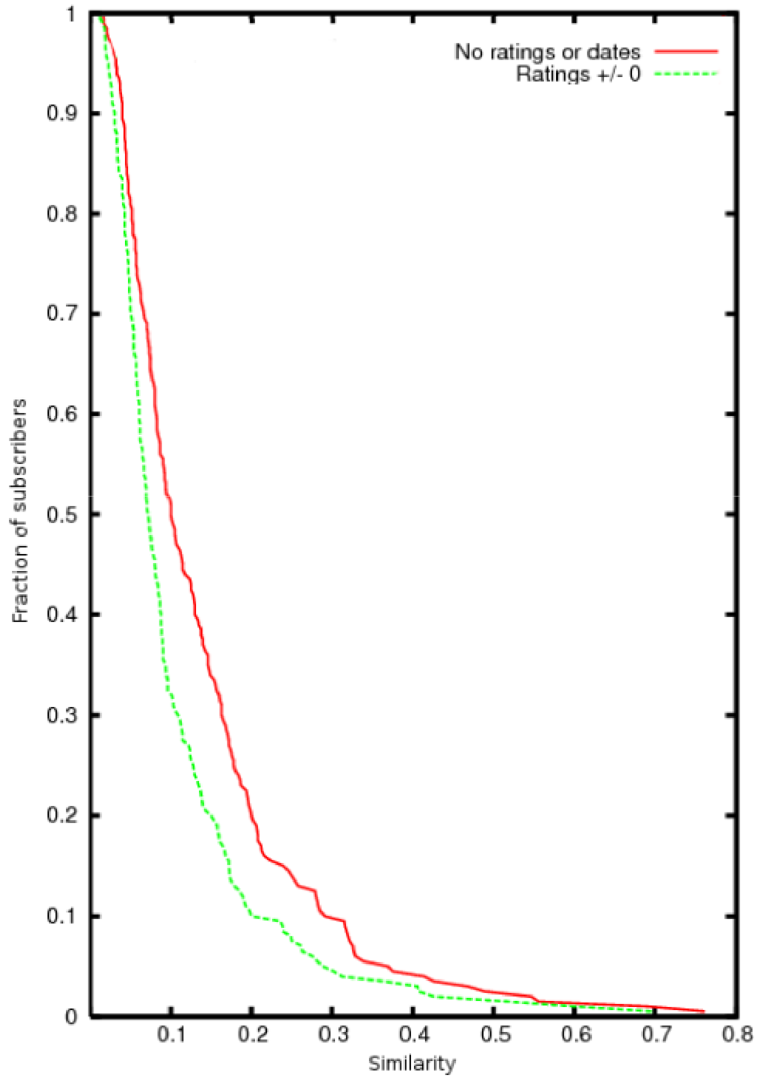
Srivatsa & Hicks, 2012
Ji et al., 2014

How identifiable are we? (8)



Sum of these problems

- Basic problem: population of 7 billion → 33 bits of information
- Low similarity of items
  - Large dimensionality of data
  - Heavy tail distribution of used attributes
  - Easy feature selection!



Narayanan & Shmatikov, 2008

Sum of these problems (2)

Pros

- Publishing (anonymous) databases is good for research
  - We have types and sizes of data never before.

Cons

- Previous techniques fail (because of sparsity)
- Breakability of anonymization schemes? Provability?
- What about wholesale surveillance?
  - One should prepare for attackers with strong auxiliary data!

ANY SOLUTION CANDIDATES? K-ANONYMITY AND DIFFERENTIAL PRIVACY

K-anonymity

- Definition
  - In a database a set of attributes can be considered as quasi identifiers. The database achieves k-anonymity if for all records there are at least (k-1) other rows with the same quasi identifier.
- Methods: supression or generalization
- Attributes can be: explicit id, quasi id, sensitive

Employee database

Name	Birth date	City
John	1980-01-31	New York
Emily	1976-06-25	Flint
Bob	1985-09-05	New York
Dave	1973-02-07	South Bend
...		

Healthcare database

Birth date	City	Diagnosis
1985-09-05	New York	Stroke
1973-02-07	South Bend	-
1980-01-31	New York	Flu
1976-06-25	Flint	HIV
...		

K-anonymity (2)

Employee database

Name	Birth date	City
John	1980-01-31	New York
Emily	1976-06-25	Flint
Bob	1985-09-05	New York
Dave	1973-02-07	South Bend

Healthcare database

Birth date	City	Diagnosis
198*	New York	Stroke
197*	South Bend	-
198*	New York	Flu
197*	Flint	HIV

?

Better: P(„John has flu”)=1 → P(„John has flu”)= ½

Employee database

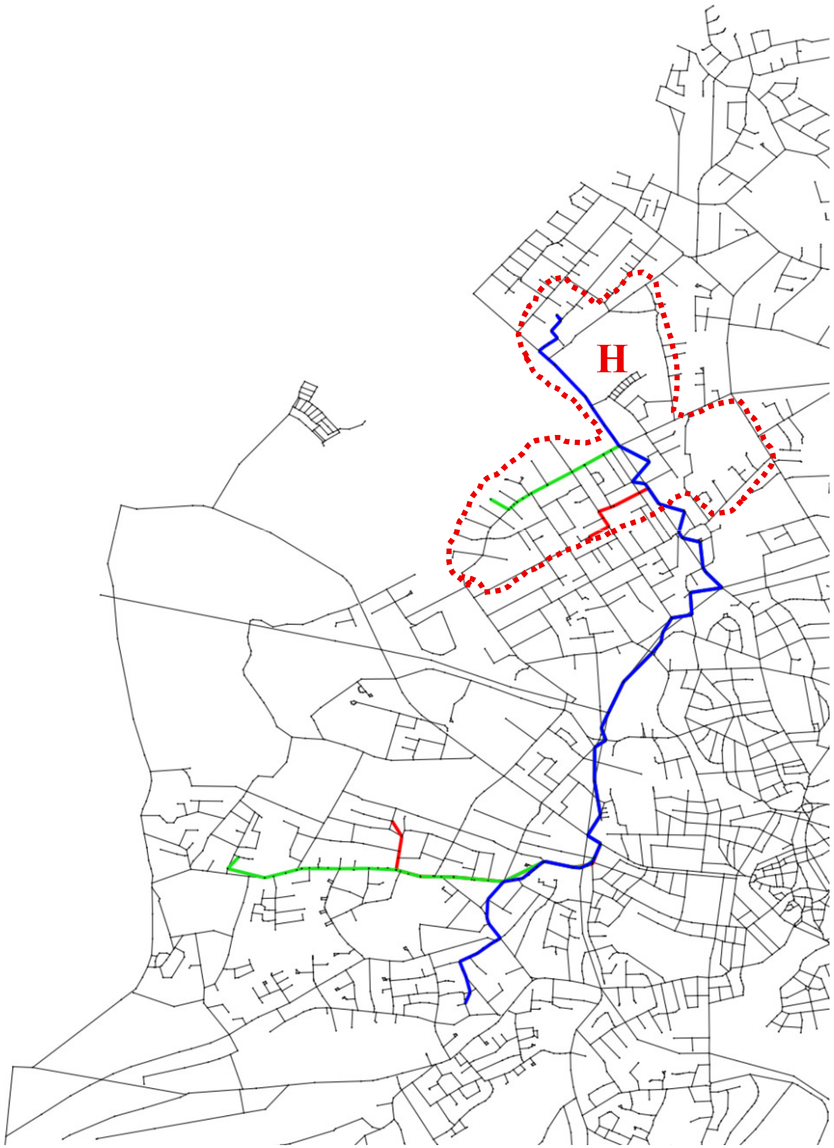
Name	Birth date	City
John	1980-01-31	New York
Emily	1976-06-25	Flint
Bob	1985-09-05	New York
Dave	1973-02-07	South Bend

Healthcare database

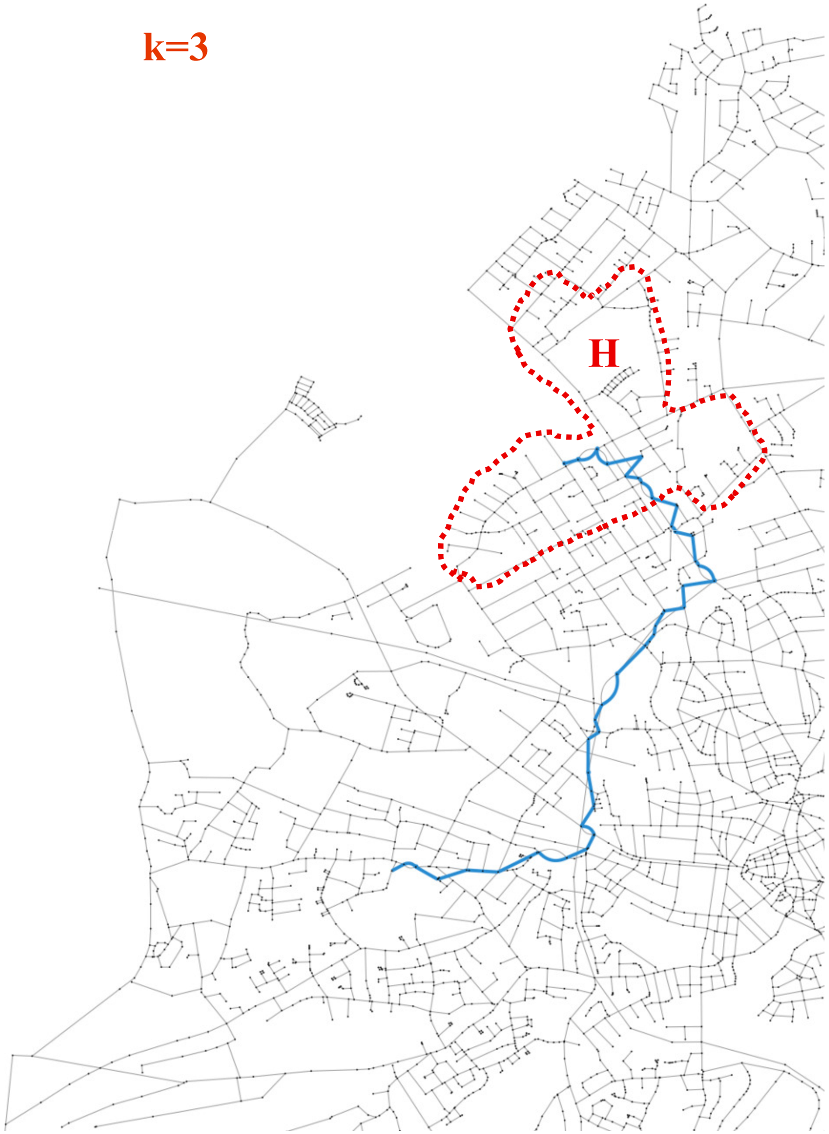
Birth date	City	Diagnosis
198*	New York	Stroke
197*	[small city]	-
198*	New York	Flu
197*	[small city]	HIV

Even better: probs are now ½ for all! (2-anonymity)

K-anonymity (3) – homogeneity attack



k=3



ε-differential privacy

- Definition
  - A randomized algorithm *A* is ε-diferentially private if for all two datasets D_1 and D_2 that differ in single row, for all *S* outcomes of *A* the following holds:

$$P\left(A\left(D_1\right) \in S\right) \leq e^{\varepsilon} \cdot P\left(A\left(D_2\right) \in S\right)$$

- In practice?
  - Changing one element in the datasets will not change the outcome significantly, that someone could tell the differing value.
    - E.g., by adding noise to results.
  - Provable privacy!
  - Not very good with some types of data, some types of uses, or with small datasets.

ε-differential privacy (2)

Query #1
avg blood sugar level
of the group?

Alice	4.2
Bob	5.9
Cathy	5.2
Diana	6.9
Ellen	5.7
Avg:	**5.58**

Query #2
avg blood sugar level
of female members?

Alice	4.2
-	
Cathy	5.2
Diana	6.9
Ellen	5.7
Avg:	**5.50**

Differentially private approach:
let's add some noise of `unif(-2,2)`

Alice	4.5
Bob	5.1
Cathy	4.41
Diana	6.2
Ellen	5.7
Avg:	**5.23**

Err. ~7%

Alice	3.0
-	
Cathy	3.7
Diana	7.5
Ellen	7.5
Avg:	**5.46**

Err. <1%

Blood sugar level of Bob?
`5*5.58-4*5.5 = 5.9`

Blood sugar level of Bob?
`5*5,23-4*5,46 = 4,3`

Err. ~27%

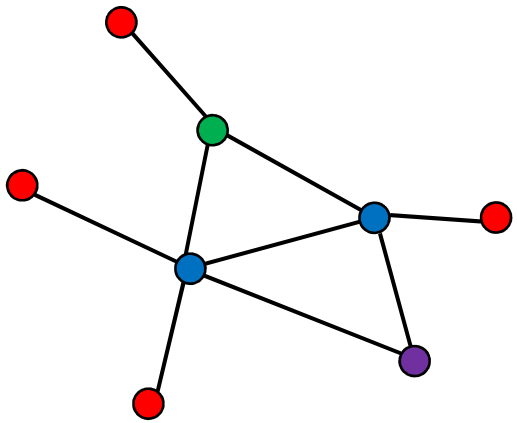
Differential privacy sounds cool, right?

Arvind Narayanan
@random_walker

Rappor is the 2nd real-life differential privacy deployment I've heard of cnet.com/news/how-googl… (after Onthemap onthemap.ces.census.gov)

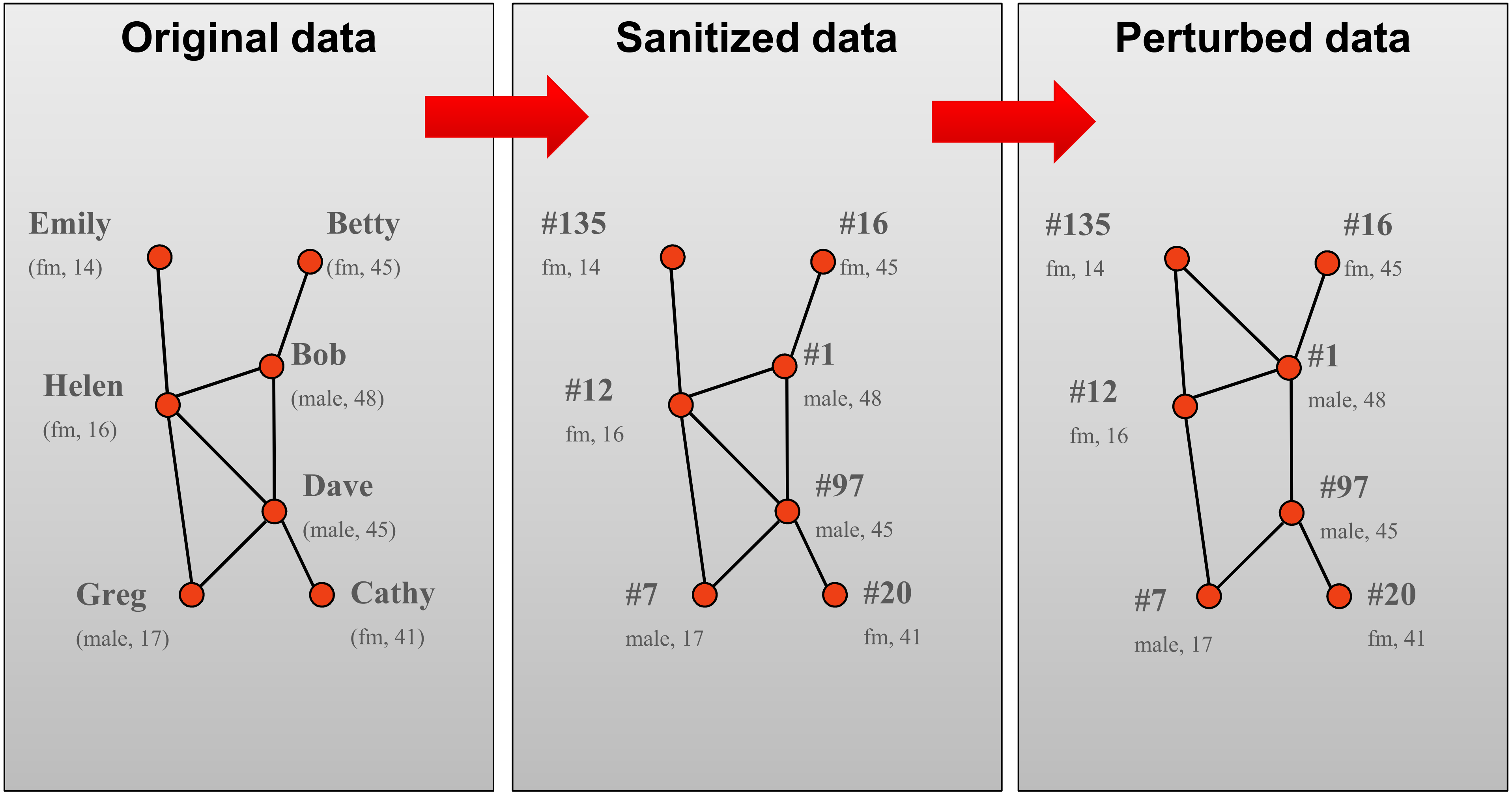
CNET

How Google tricks itself to protect Chrome user privacy
By CNET @CNET
An open-source project called Rappor uses randomly muddled data to let Google gather information about people's software usage while keeping individuals' behavior private.



STRUCTURAL DE-ANONYMIZATION IN SOCIAL NETWORKS

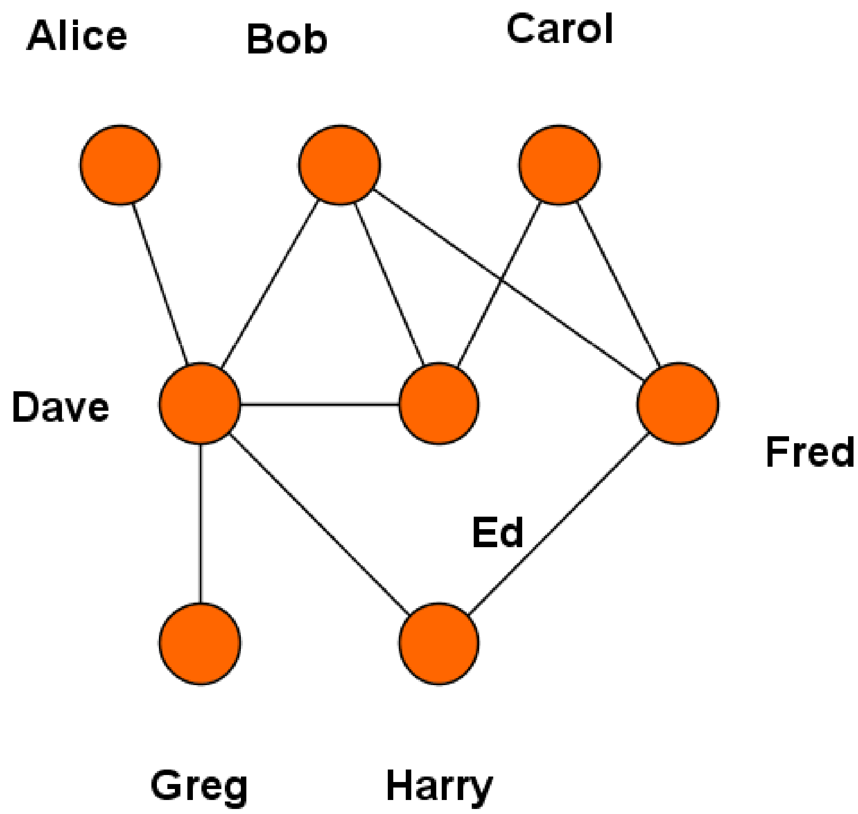
Data perturbation and sanitization



Attacker model

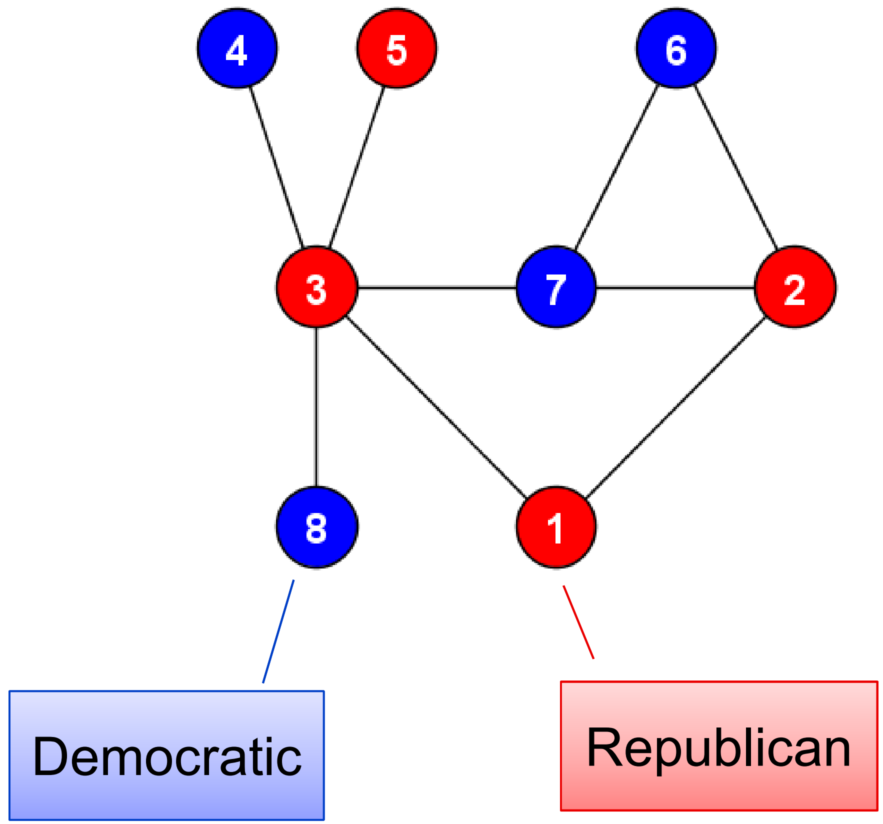
Auxiliary information, G_{src}

(a public crawl, e.g., Flickr)



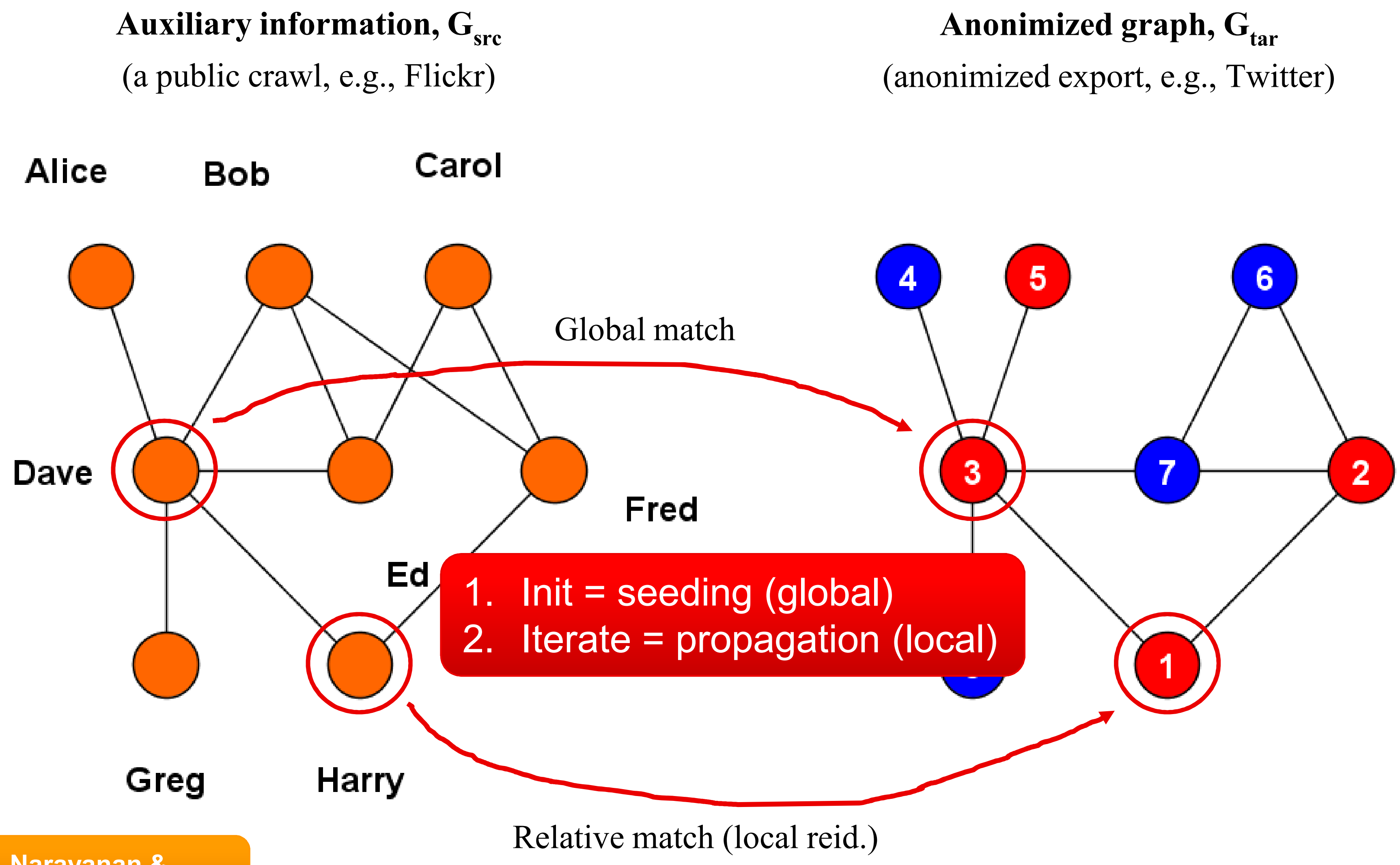
Anonimized graph, G_{tar}

(anonimized export, e.g., Twitter)



Narayanan & Shmatikov, 2009

Attacker model (2)

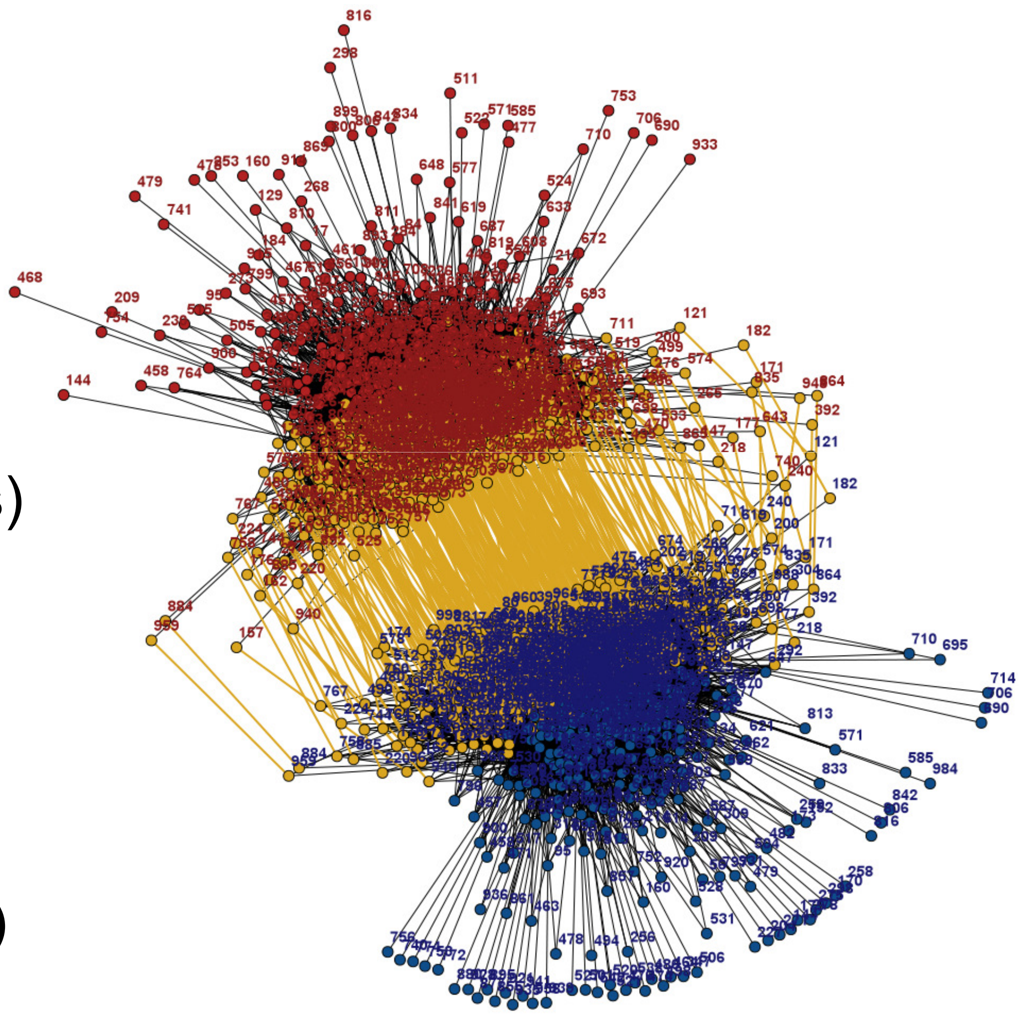


1. Init = seeding (global)
2. Iterate = propagation (local)

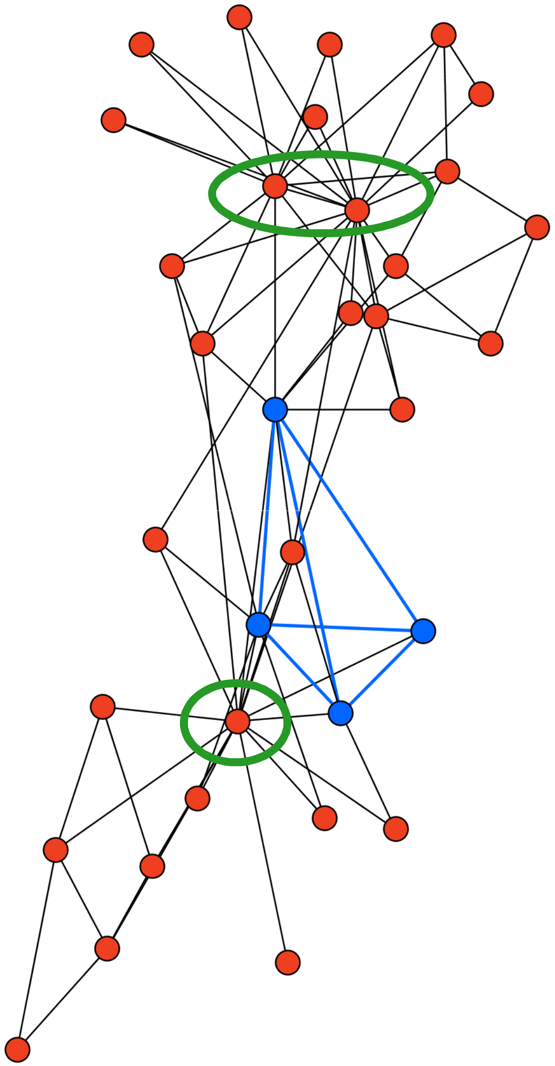
Narayanan & Shmatikov, 2009

Large-scale re-identification

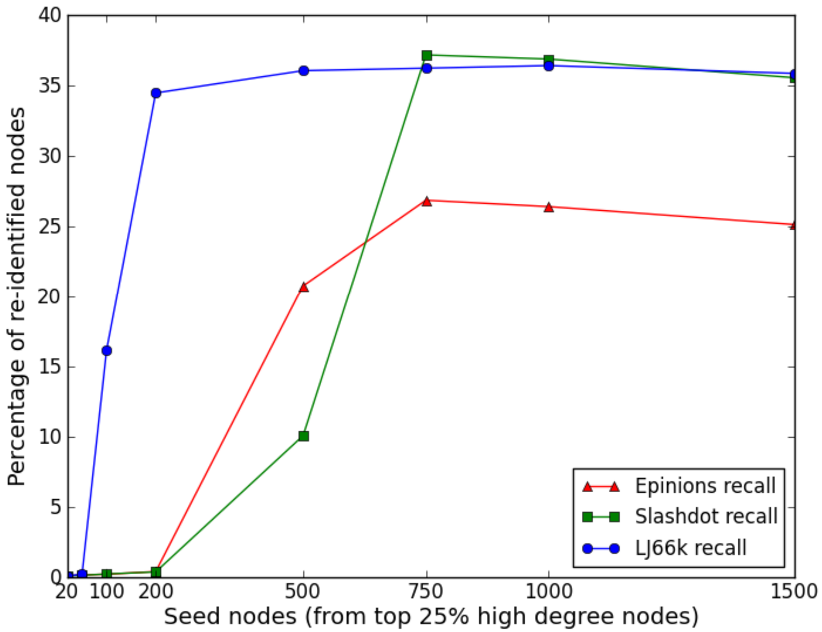
- Underlying concepts work on large social networks
  - Auxiliary data:
    Flickr (3,3m ns, 53m es)
  - Target (anon.) data:
    Twitter (224k ns, 8,5m es)
  - Ground truth: 27k nodes (name/user/loc.)
- Results:
  - 30% TP, only 12% FP
  - (Init: 150 highdeg. seeds)

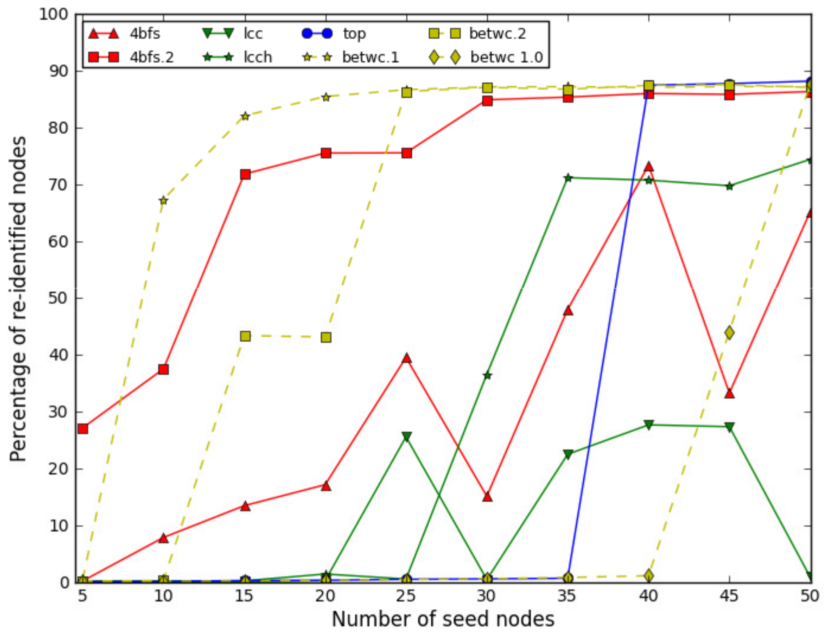


Initialization?



Initialization? (2)



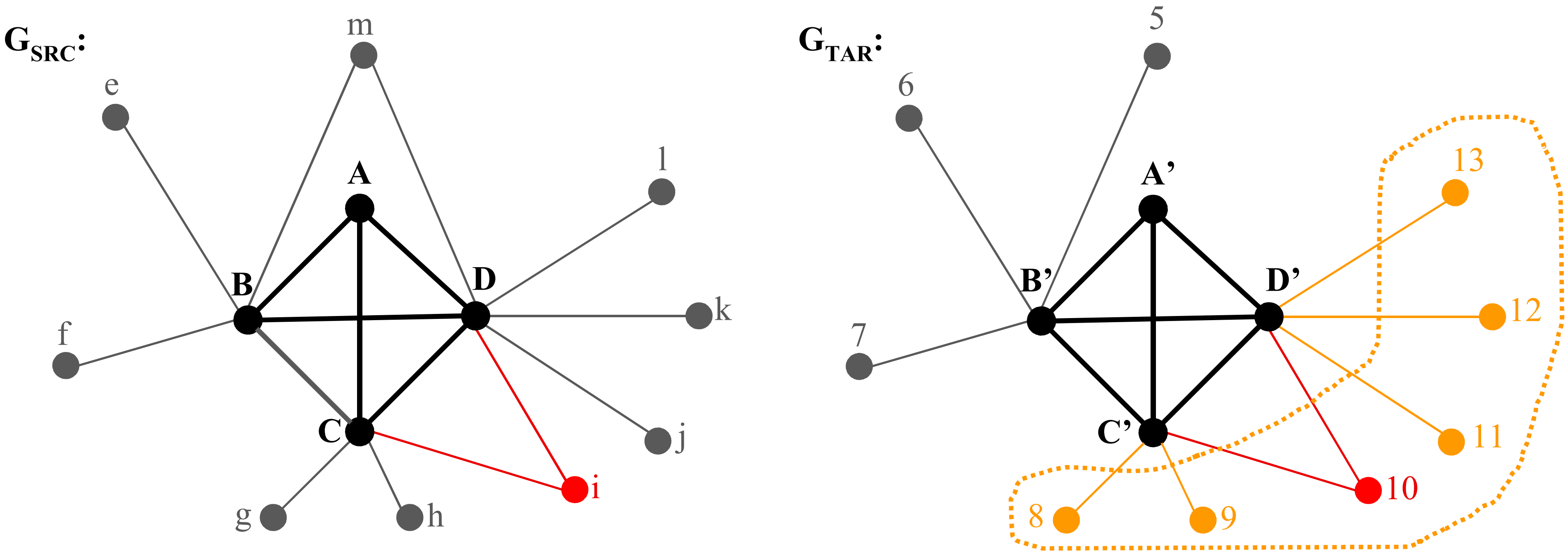


http://gulyas.info/upload/GulyasG_SESOC14.pdf

Details on the propagation phase

Narayanan & Shmatikov, 2009

- Do $\forall v_i \in V_{SRC}$ until we have convergence:
  1. Identified neighbors: $\{v_1,\dots,v_k\} \in V_{SRC}$, mapped to $\{v_{1'},\dots,v_{k'}\} \in V_{TAR}$, e.g. $\mu(v_1)=v_{1'}$
     a. Select $N=\{v_{u_1},\dots,v_{u_m}\} \in V_{TAR}$ from nbrs($\{v_{1'},\dots,v_{k'}\}$)
     b. Calculate score: $S=\{s_{u_1},\dots,s_{u_m}\}$
  2. If $v_{i'}$ is an outstanding candidate in S, do a reverse match checking by swaping the datasets G_{TAR} and G_{SRC} (and the mapping)
  3. If v_i is the reverse best-match, set $\mu(v_i)=v_{i'}$



Details on the propagation phase (2)

Narayanan & Shmatikov, 2009

- Score calculation:

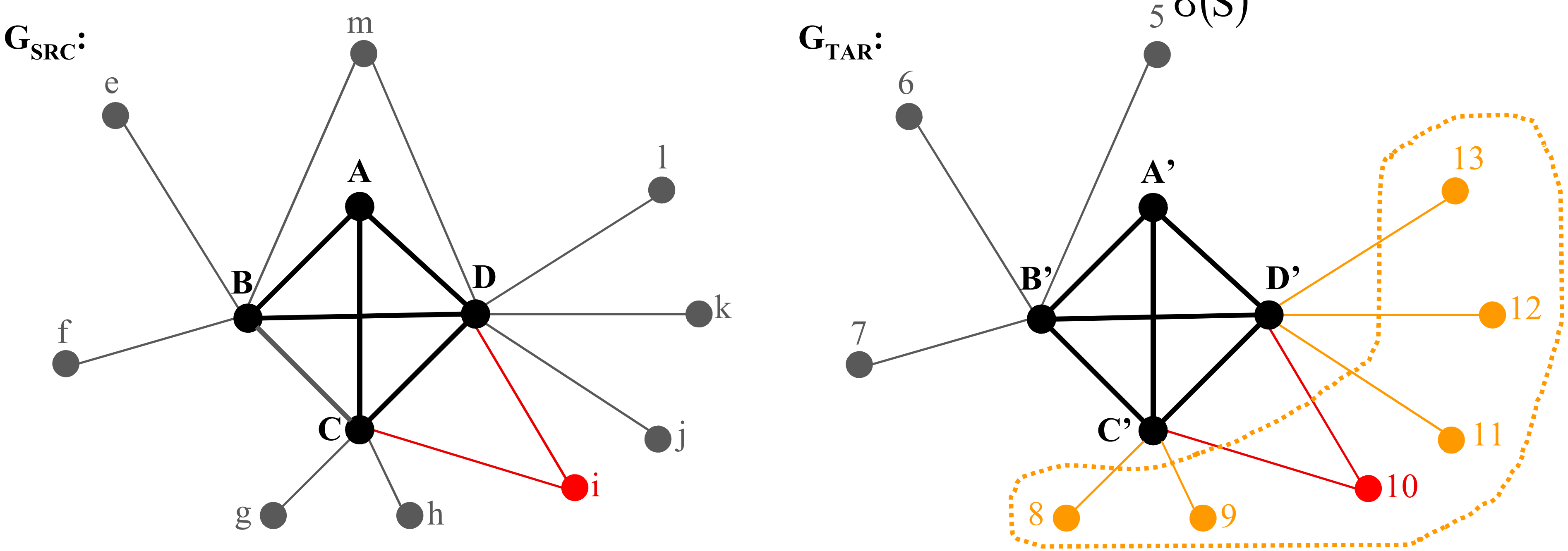
$$\mathrm{Score}(v_i, v_j) = \frac{|V_i \cap V_j|}{\sqrt{|V_j|}}$$

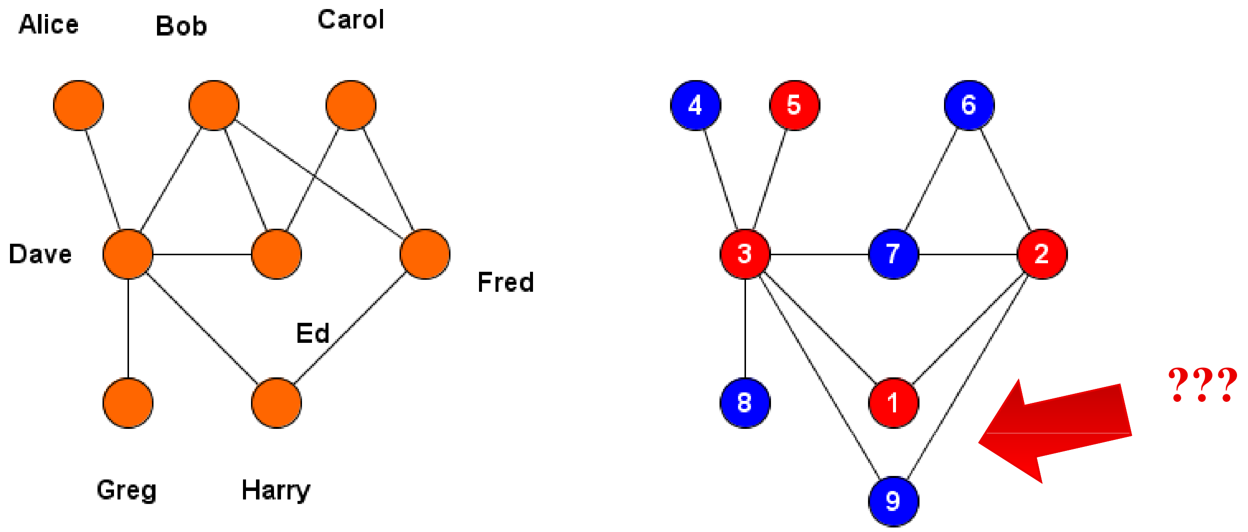
- Cosine similarity:

$$\mathrm{CosSim}(v_i, v_j) = \frac{|V_i \cap V_j|}{\sqrt{|V_i| \cdot |V_j|}}$$

- Eccentricity check:

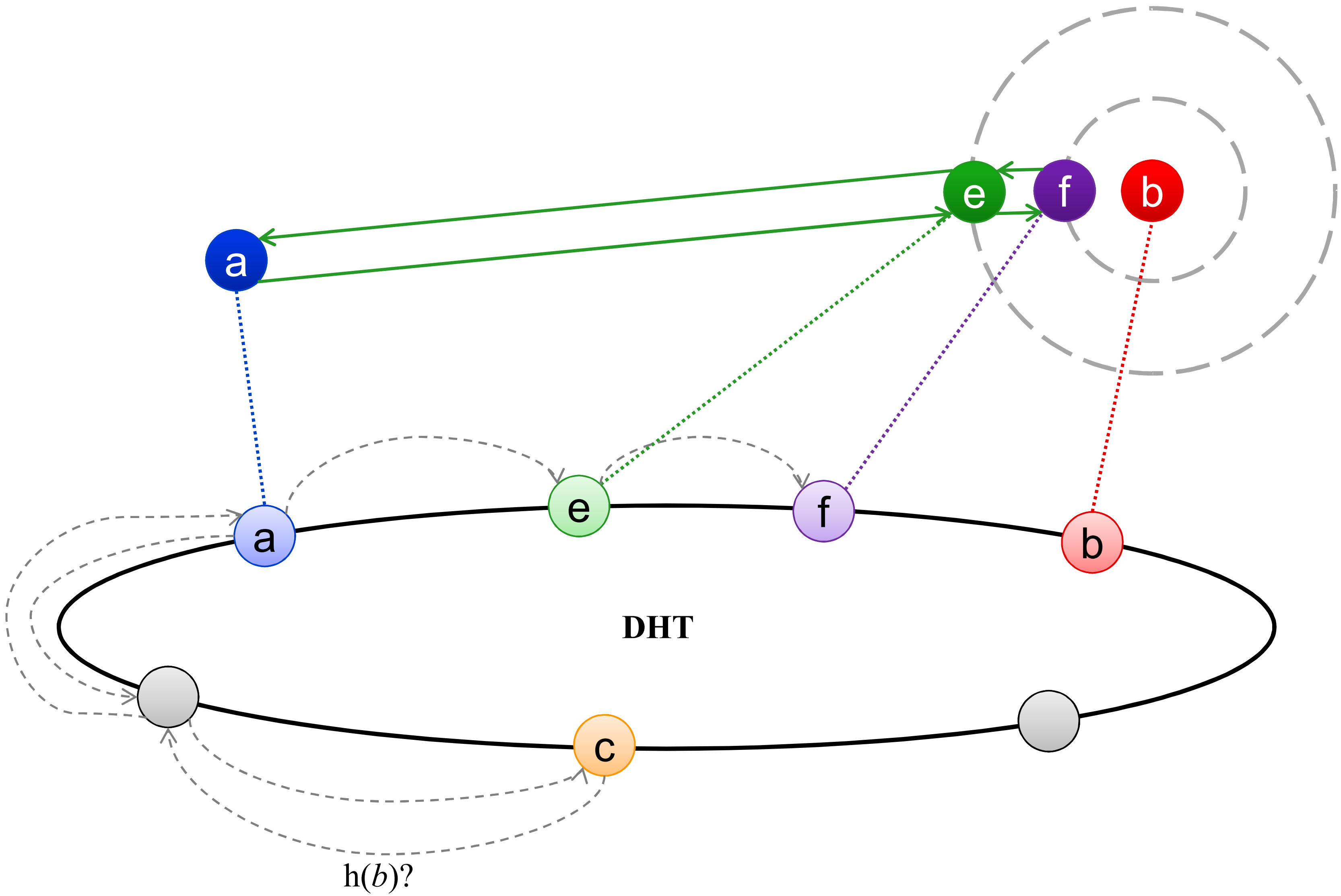
$$\mathrm{Eccentricity}(S) = \frac{\max(S) - \max(\{S \setminus \max(S)\})}{\sigma(S)}$$



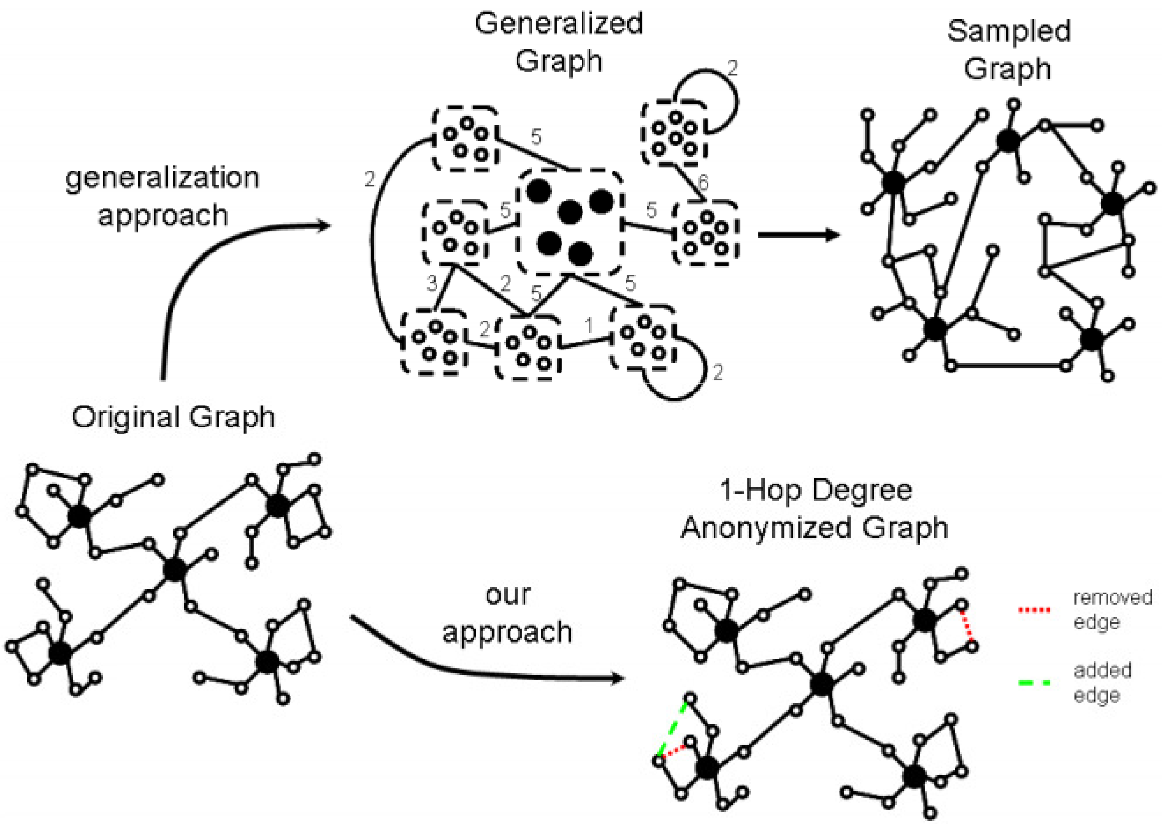


TACKLING STRUCTURAL DE-ANONYMIZATION

Possible solutions? Safebook.



Possible solutions? Data sanitization. (2)

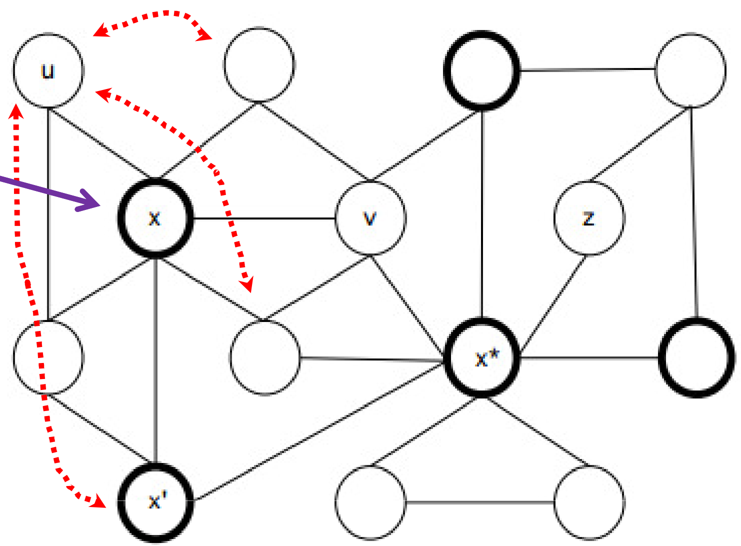


http://people.cs.vt.edu/danfeng/papers/social-anon.pdf

The friend-in-the-middle model

Beato et al., 2013

the proxying friend

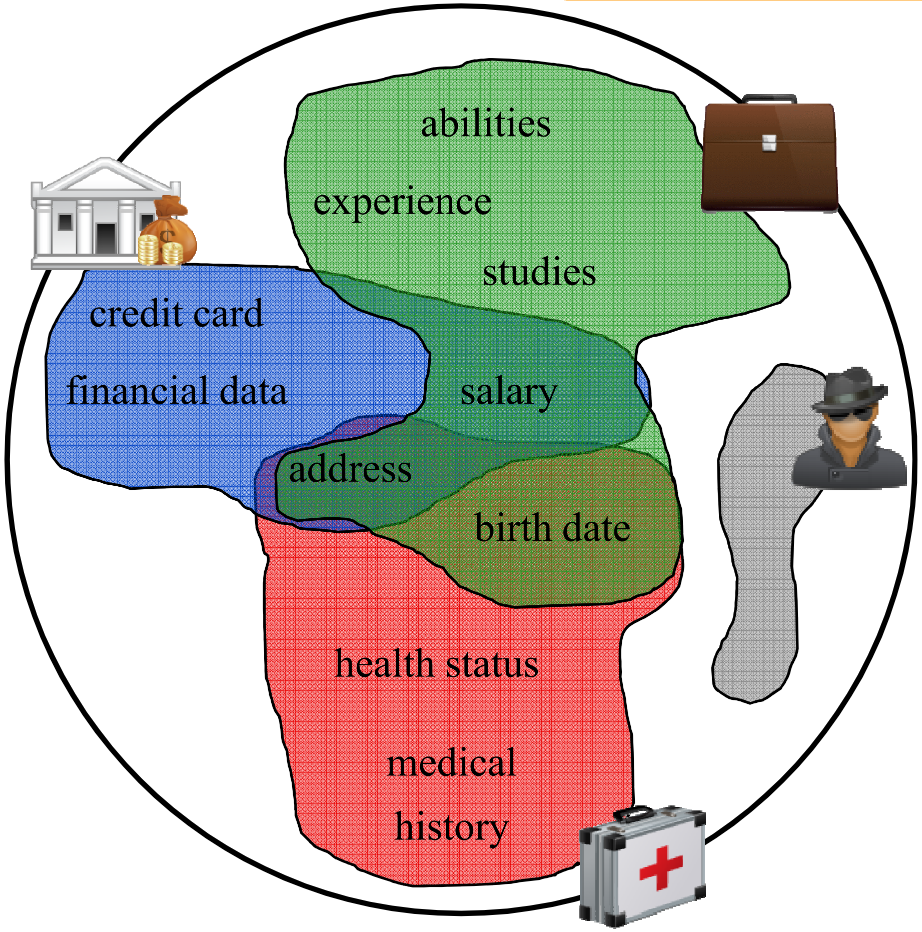


- Basic principle: some nodes act as a proxy (hiding edges)
- Cooperative: users choose proxy nodes (both trusted)
- Results:
  - Proves 10% of users are enough (perhaps less)
  - On a quite sparse network (easier to defend ☹)
  - Requires cooperation: 3 nodes need to agree per edge

(Privacy-Enhancing) Identity management

Clauß et al., 2005

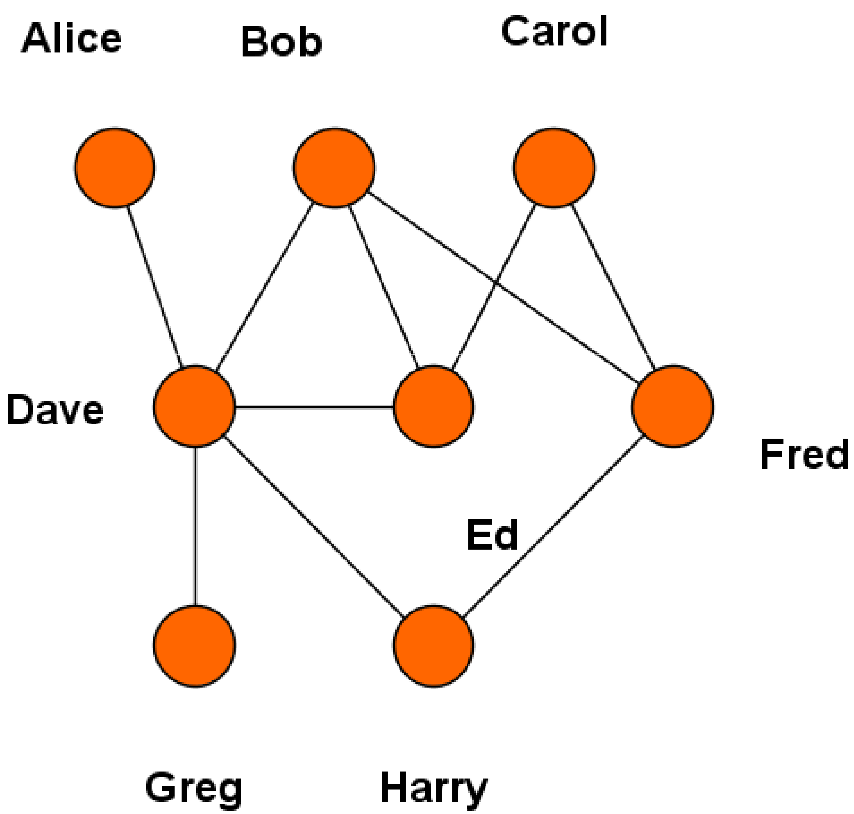
- Partial identity:
  - Subset of the attributes of the global identity
  - Invoked by different roles and contexts
  - Can have pseudonyms
  - Linkability of partial identites and actions



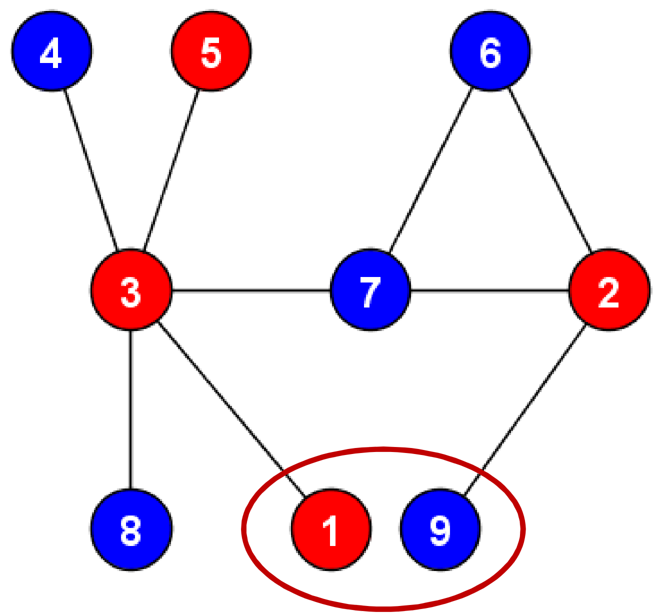
Global and partial identities of John Doe

Idea: using identity management? (2)

Auxiliary information, G_{src}
(a public crawl, e.g., Flickr)



Anonimized graph, G_{tar}
(anonimized export, e.g., Twitter)



Identity separation

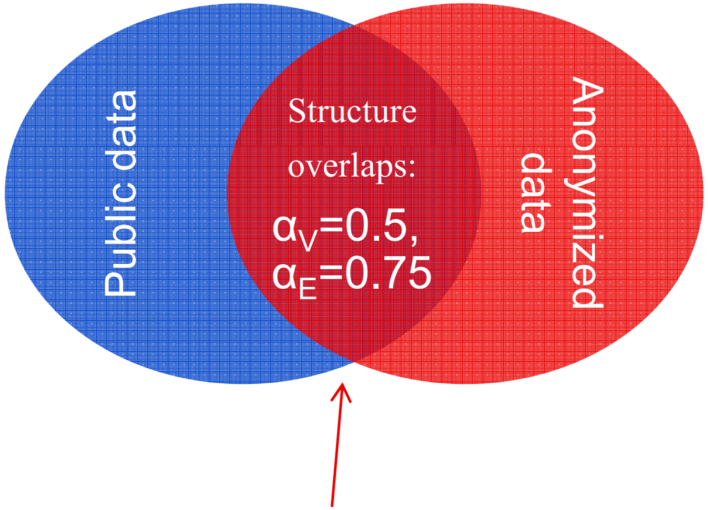
Gulyas, 2014

Idea: using identity management? (3)

Step 1: anonymized network

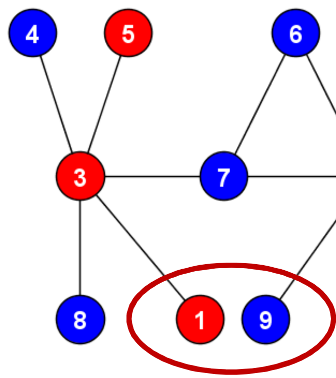
Epinions (75k),
Slashdot (82k),
LiveJournal (66k)

Step 2: perturbation



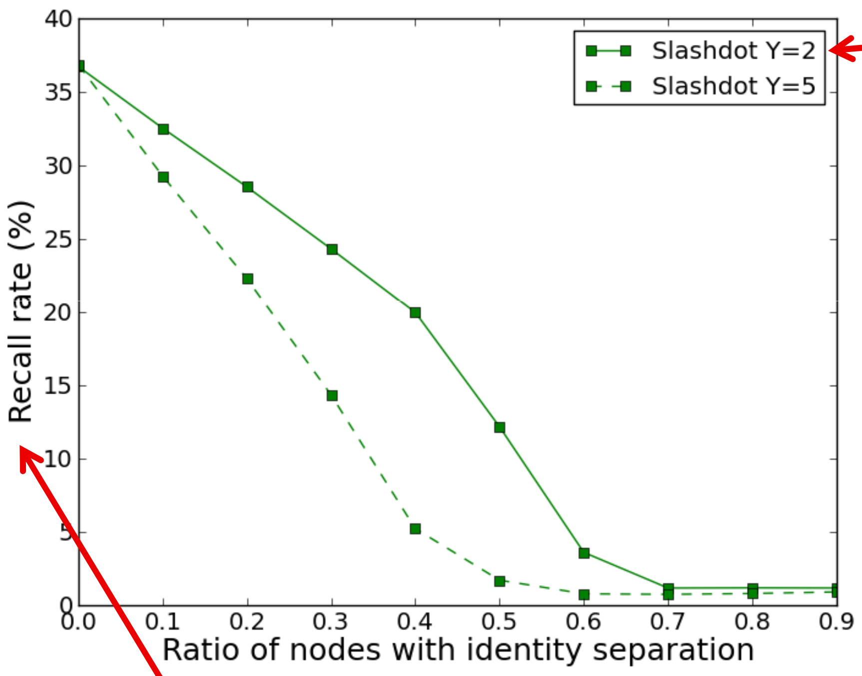
ground truth

Step 3: simulating identity separation



Non-cooperative identity separation?

- Splitting nodes and redistributing edges uniformly (basic model)



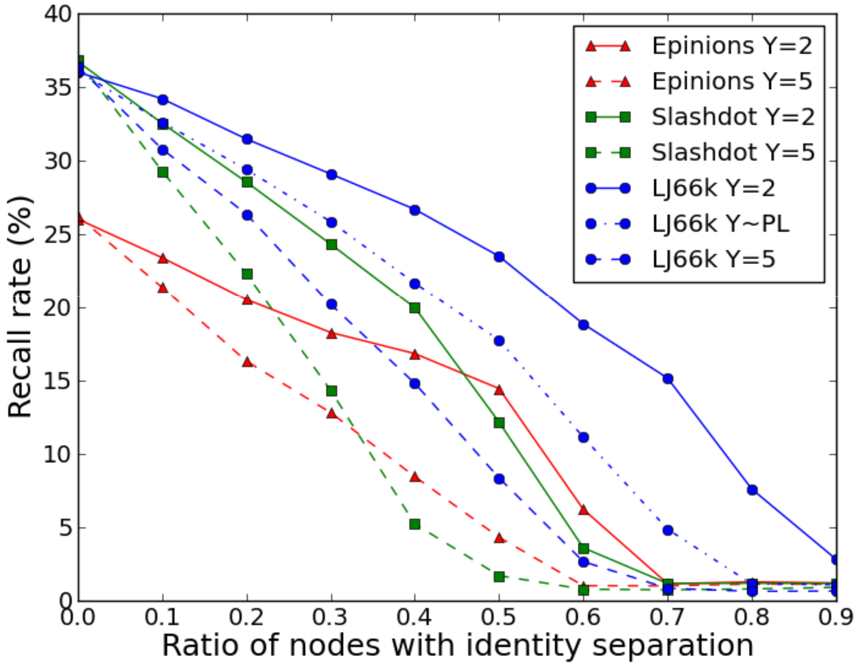
Creating Y=2 new vertices from one, and sorting edges with ½ probability to each.

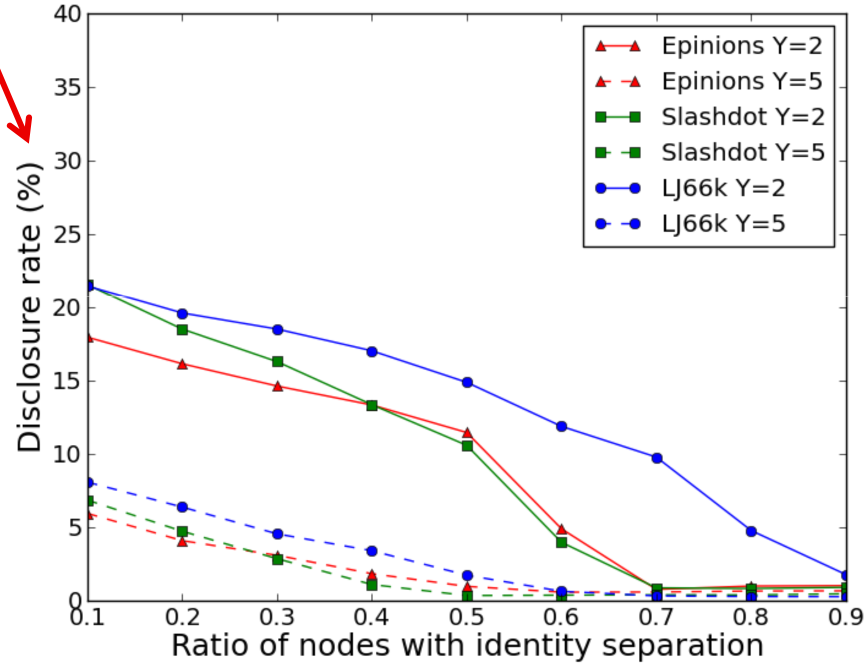
Recall rate: percent of correctly re-identified nodes.

Non-cooperative identity separation? (2)

- Splitting nodes and redistributing edges uniformly (basic model)

Disclosure rate: what the attacker learns.
(i.e., amount of edges currently)



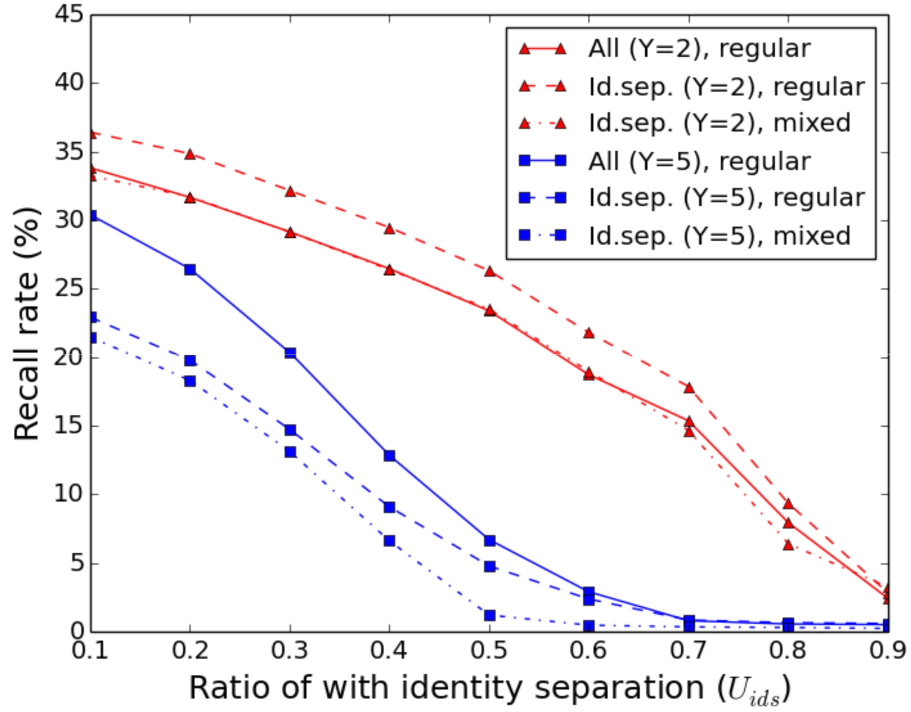


Over all nodes!

Over nodes with identity separation!

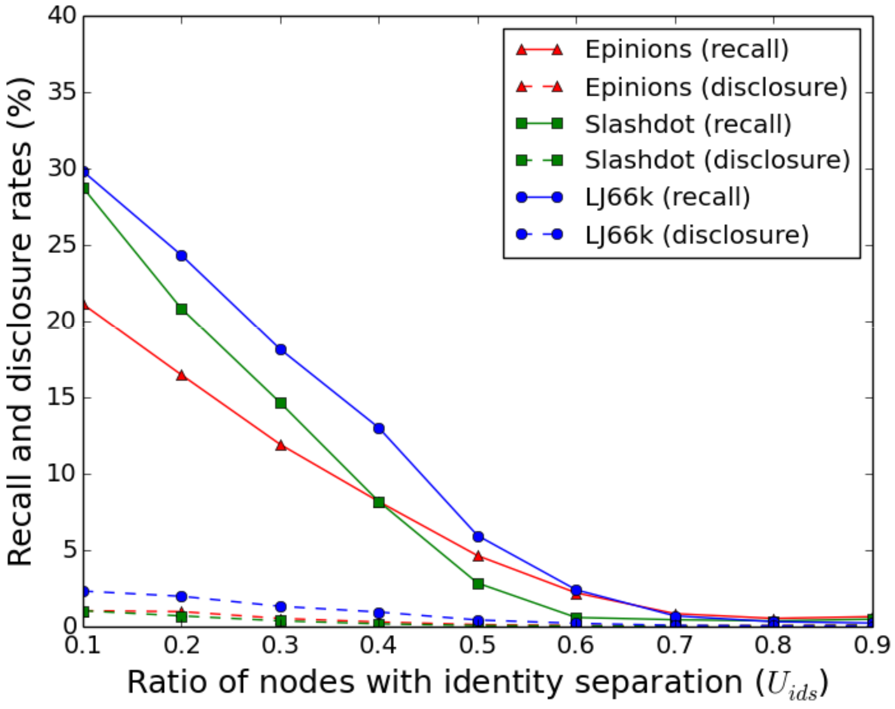
Non-cooperative identity separation? (3)

- Interesting finding:
  - Only for Y=2
  - Nodes with identity separation had higher recall rate than others
  - Caused by using non-idsep nodes for seeding
- Conclusion:
  - Natural choice → bad implications on privacy
  - Use Y=2+ ☺

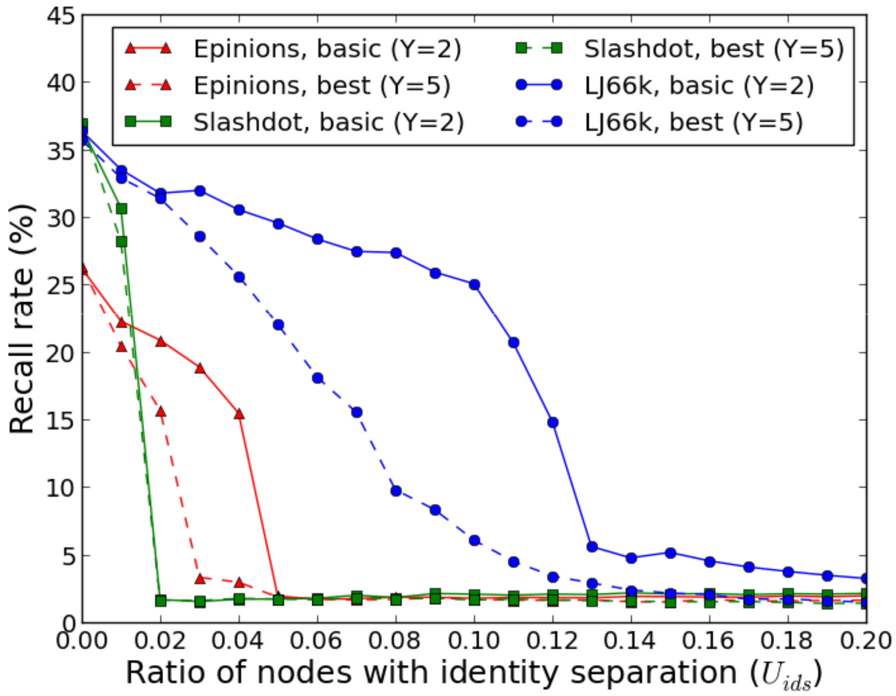


Tackling the attack: on the network level

- Splitting nodes, redistributing edges uniformly, while some may be subjected to deletion (best model)

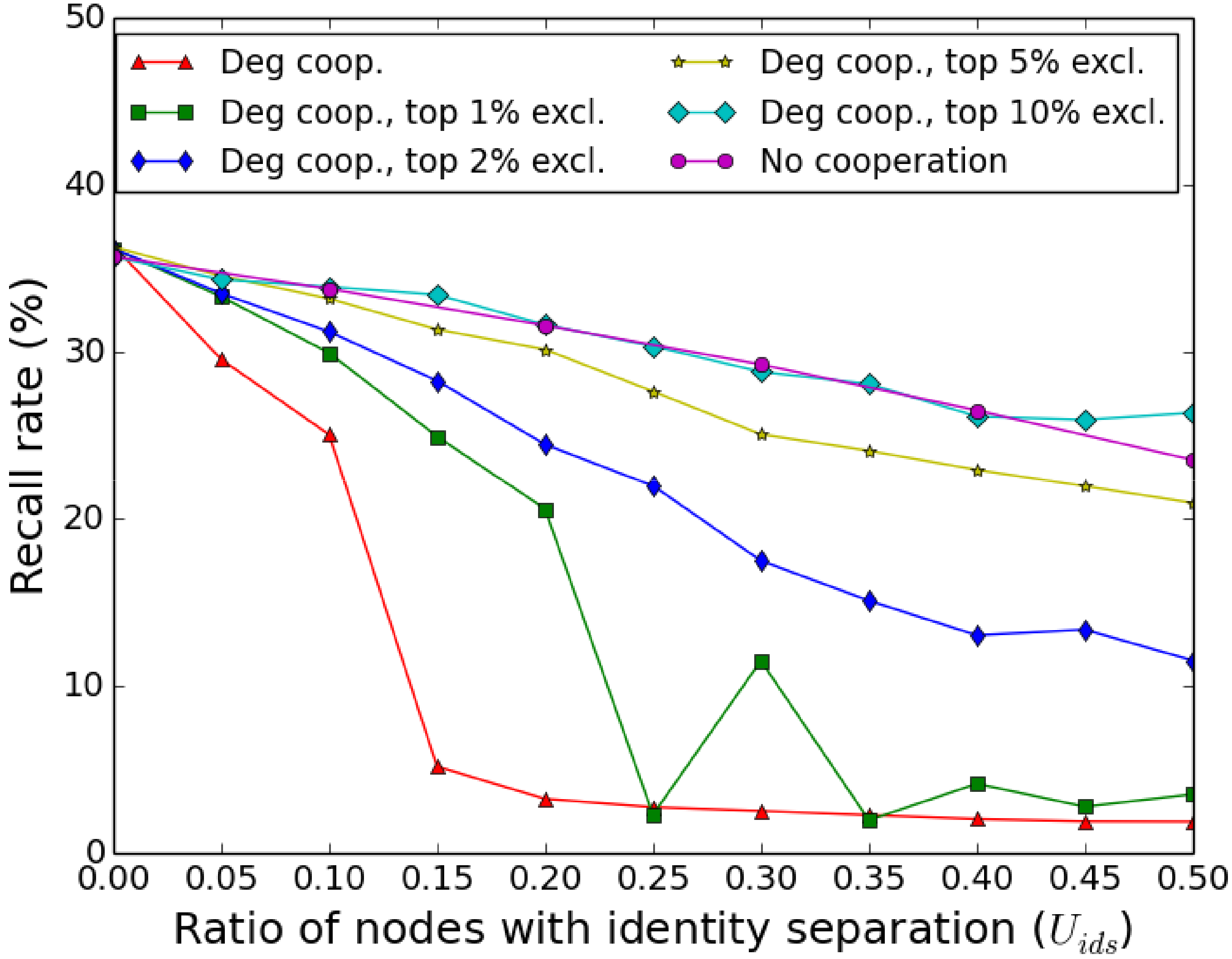


No cooperation between users



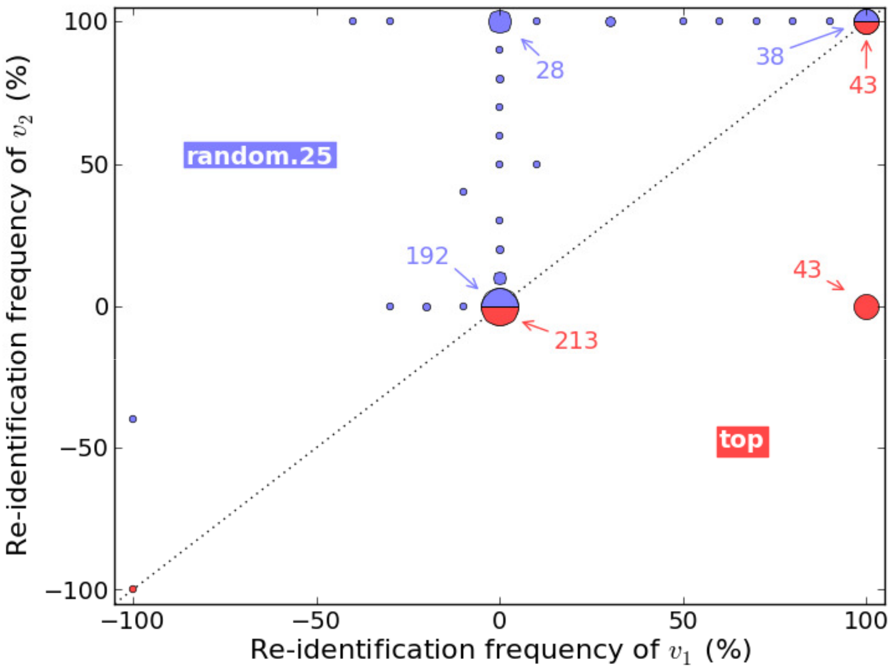
Users with highest degree cooperate

Network level protection: there is a problem!

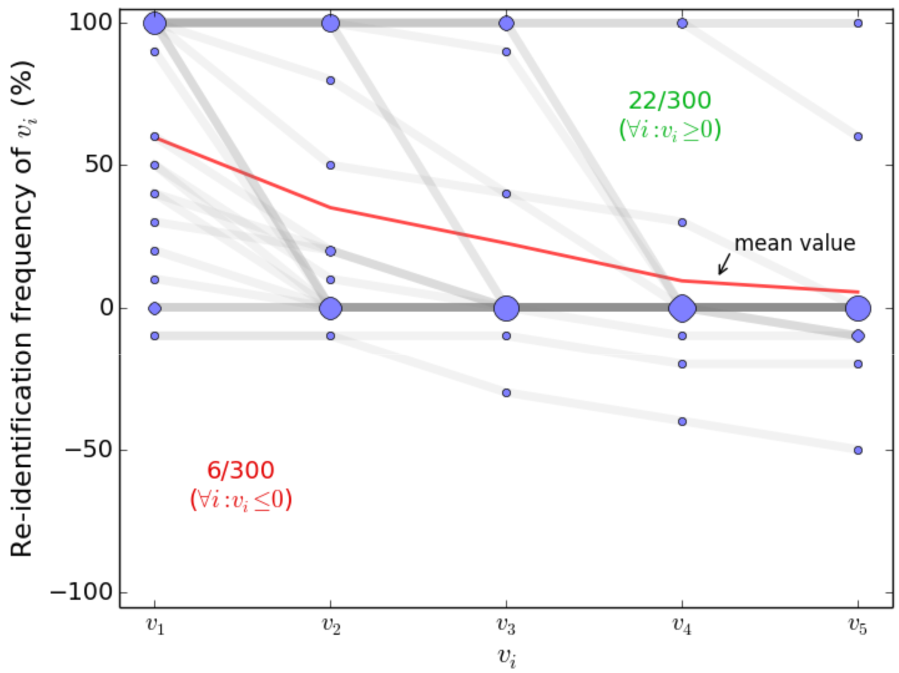


Tackling the attack: on the personal level

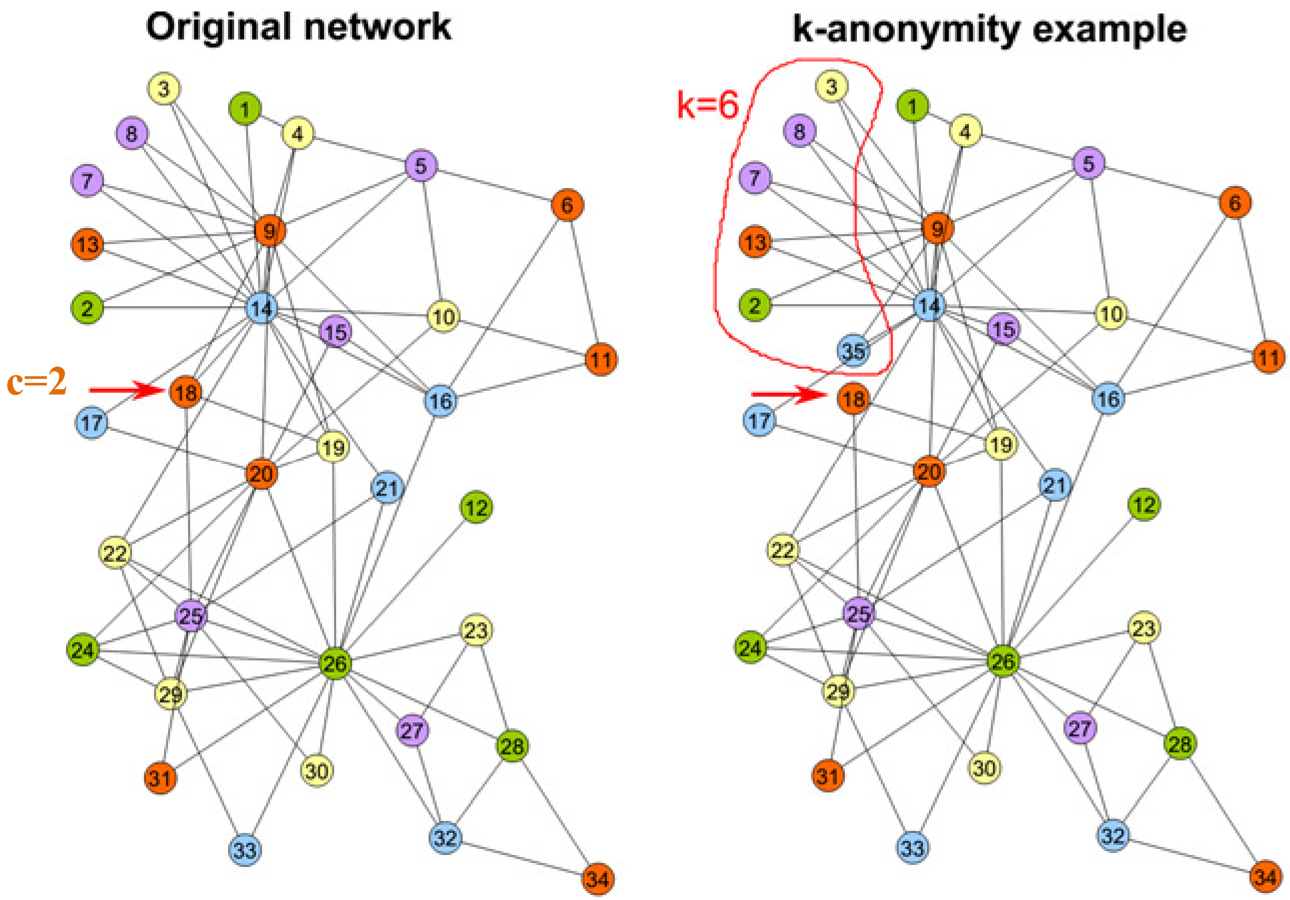
Basic model, 2 identities



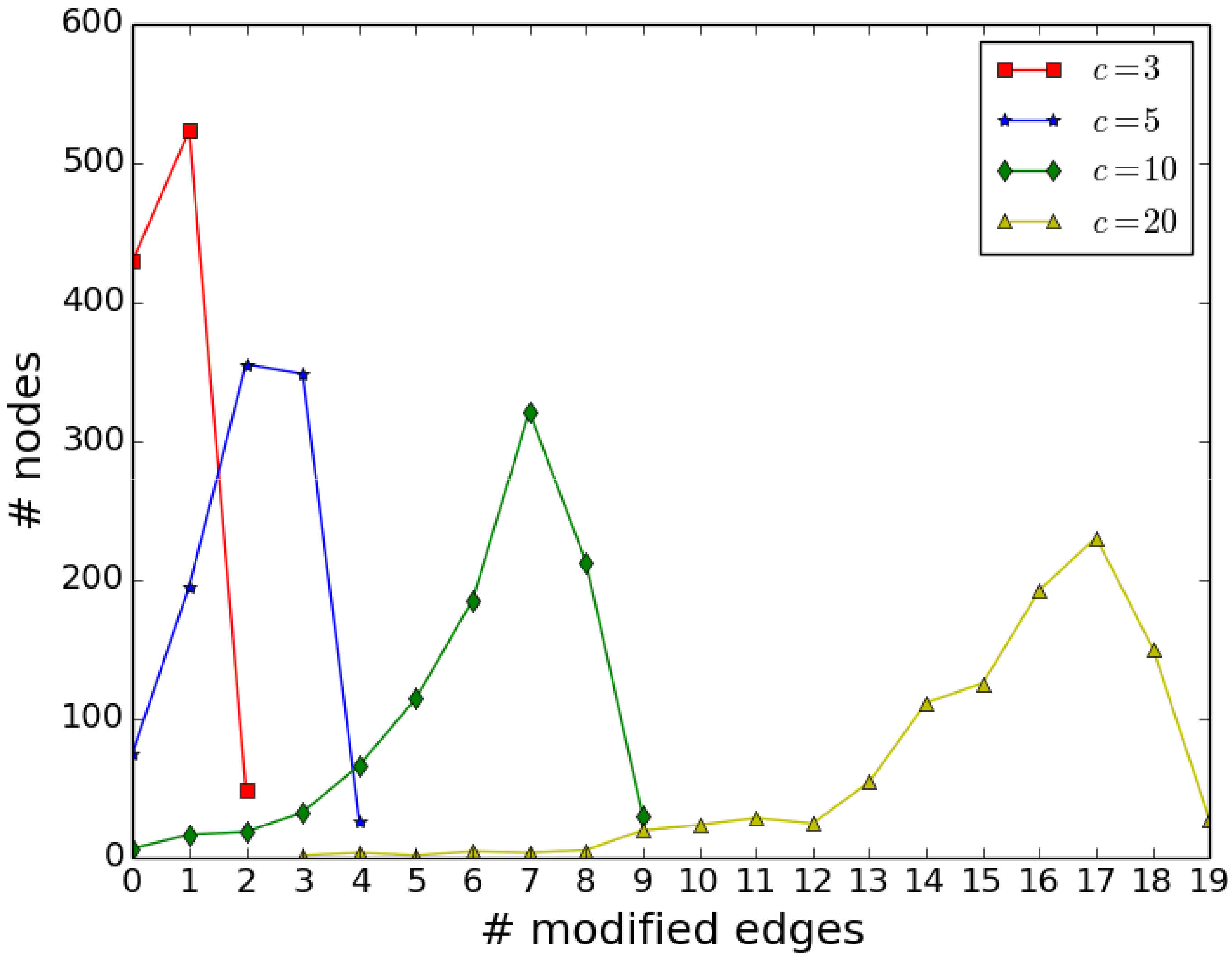
Basic model, 5 identities
(results ordered by frequency)



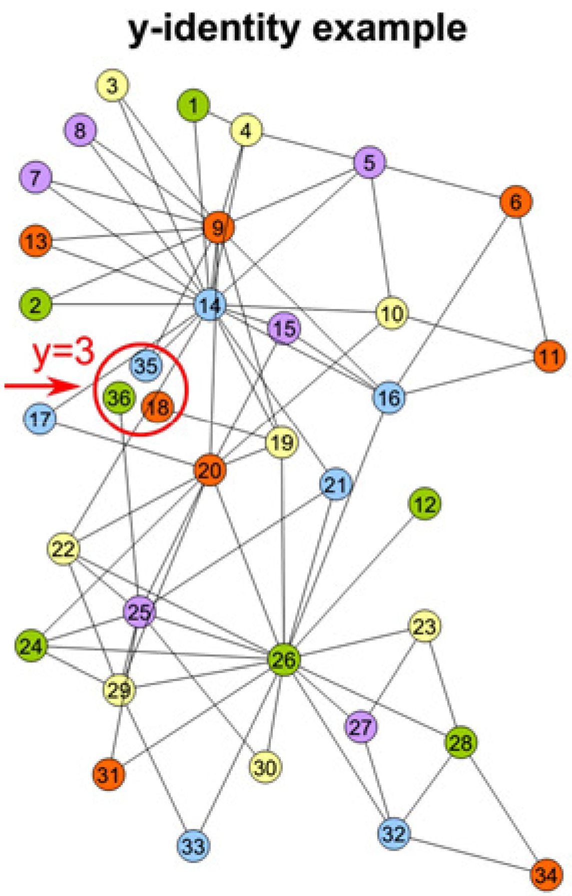
K-anonymity?


Original network
k-anonymity example
k=6
c=2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35

K-anonymity? (2)



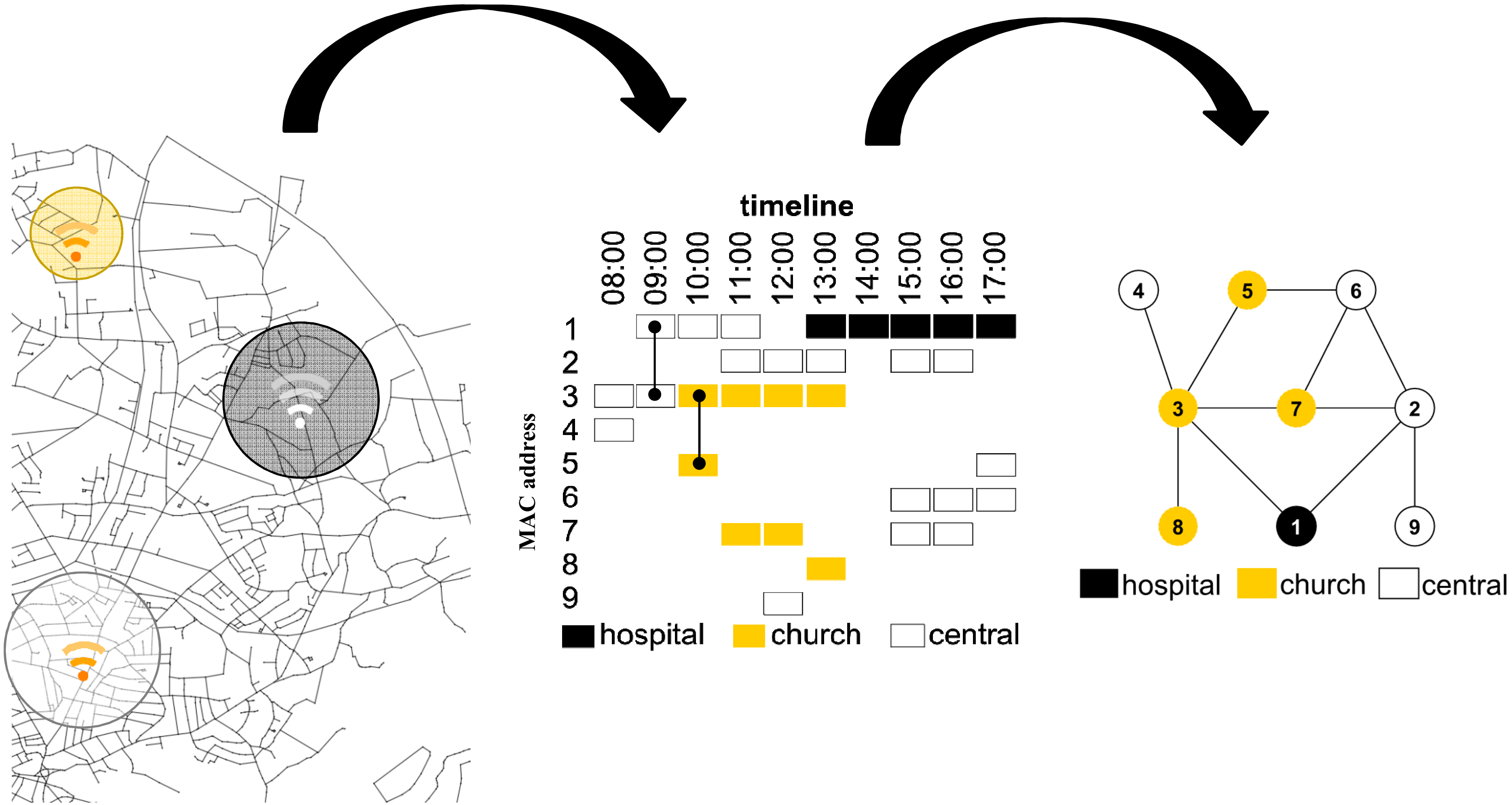
y-identity model



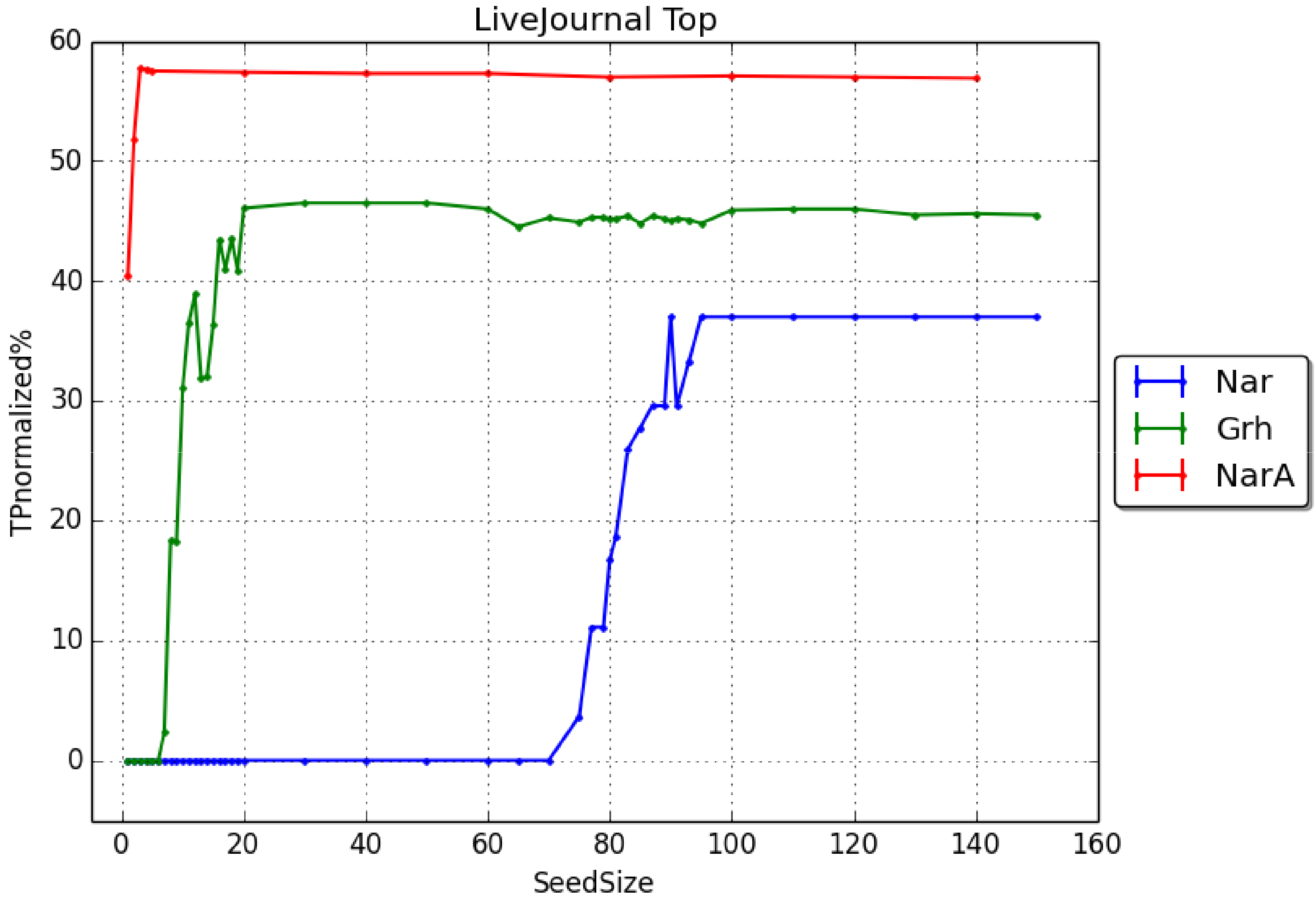
- It works simply, but:
  - tackling different attackers need different strategies
- It can be proven there is a one-fits-all strategy:
  - use 1/y probs,
  - there are some extension,
  - and some constraints.

NEXT ATTACKS ON SOCIAL DE-ANONYMIZATION?

Principles apply to other contexts also



Is this the top? (3)



CONCLUSION

Conclusions

- Technology providing vast amount of data is here – but we are not ready
  - How do we detect privacy leakeges?
  - How to design privacy friendly services?

    (and how to convince busniess men to do so ☺)
  - How do we protect privacy?
  - How can we evaluate protection schemes?
  - ...
- Can we handle big data technology somehow? Or have we yet passed the point of safe return?

Thank you for your attention!

Any questions?

Gábor György Gulyás
gulyas.info // @GulyasGG
Laboratory of Cryptography and System Security (CrySyS)
Budapest University of Technology and Economics
www.crysys.hu

References

- Latanya Sweeney: Uniqueness of simple demographics in the US population. *LIDAP-WP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA* (2000).
- Philippe Golle: Revisiting the uniqueness of simple demographics in the US population. *Proceedings of the 5th ACM workshop on Privacy in electronic society*. ACM, 2006.
- Philippe Golle, Kurt Partridge: On the anonymity of home/work location pairs. *Pervasive Computing*. Springer Berlin Heidelberg, 2009. 390-397.
- Arvind Narayanan, Vitaly Shmatikov: Robust de-anonymization of large sparse datasets. *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 2008.
- Arvind Narayanan, Vitaly Shmatikov: De-anonymizing social networks. *Security and Privacy, 2009 30th IEEE Symposium on*. IEEE, 2009.
- Gilbert Wondracek et al.: A practical attack to de-anonymize social network users. *Security and Privacy (SP), 2010 IEEE Symposium on*. IEEE, 2010.

References (2)

- Peter Eckersley: How unique is your web browser? *Privacy Enhancing Technologies*. Springer Berlin Heidelberg, 2010.
- Károly Boda et al.: User tracking on the Web via cross-browser fingerprinting. *Information Security Technology for Applications*. Springer Berlin Heidelberg, 2012. 31-46.
- Mudhakar Srivatsa, Mike Hicks: Deanonymizing mobility traces: Using social network as a side-channel. *Proceedings of the 2012 ACM conference on Computer and communications security*. ACM, 2012.
- Arvind Narayanan et al.: On the feasibility of internet-scale author identification. *Security and Privacy (SP), 2012 IEEE Symposium on*. IEEE, 2012.
- Filipe Beato, Mauro Conti, Bart Preneel: Friend in the Middle (FiM): Tackling De-Anonymization in Social Networks, 2013.
- Gábor György Gulyás, Sándor Imre: Hiding Information in Social Networks from De-anonymization Attacks by Using Identity Separation, 2013.

References (3)

- Clauß et al.: Privacy enhancing identity management: protection against re-identification and profiling, 2005.
- Gábor György Gulyás, Sándor Imre: Analysis of Identity Separation Against a Passive CliqueBased De-anonymization Attack, 2011.
- Gábor György Gulyás, Sándor Imre: Measuring Importance of Seeding for Structural De-anonymization Attacks in Social Networks, 2014.
- Ji et al.: Poster: Optimization based Data De-anonymization, 2014.
- Cutillo, Leucio Antonio, Refik Molva, and Thorsten Strufe. "Safebook: A privacy-preserving online social network leveraging on real-life trust."*Communications Magazine, IEEE* 47.12 (2009): 94-101.
- Gábor György Gulyás: Protecting Privacy Against Structural De-anonymization Attacks in Social Networks, dissertation, 2014.

Most images from: pixabay.com licenced under CC0 Public Domain