TO BE ON THE SAFE SIDE

Exposing individuals in anonymized large datasets

Gábor György Gulyás, Levente Buttyán
Laboratory of Cryptography and System Security (CrySyS)
Budapest University of Technology and Economics
www.crysys.hu

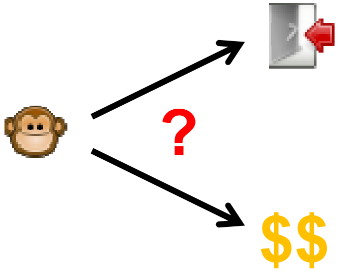
‘Natural’ sources of big data in (social) technology (e.g.)

Social networks & media

Recommender systems

Web tracking dbs (profiling)

Doc indexing & search



Predicting user behavior

Exposing trends

Trends in identification, deanonymization

Sweeney, 1990

87% of US population is identifiable by:
`{ZIP, gender, birth date}`

Golle, 2000

64% still.

Narayanan & Shmatikov, 2008

Netflix vs. IMDb

- rarely used features are identifying
- only 8 ratings identify 99% of users
- 2 erroneous, dates within 2 weeks

Using big data, things can get worse.

Golle & Partridge, 2009

Work-home location pairs (US):

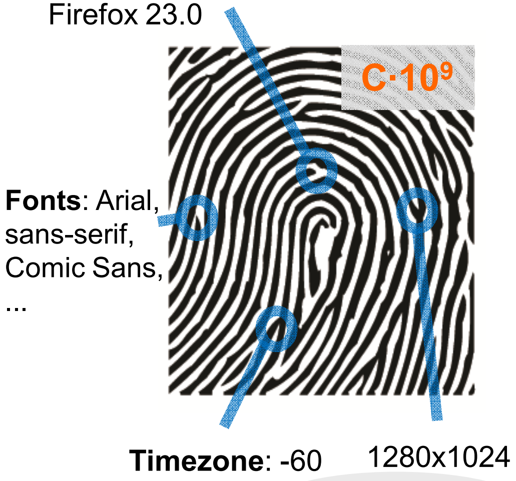
- ~ 1500 / loc. cells
- 5% identifiable
- avg. anonymity set size is ca. 20

Trends in identification, deanonymization (2)

Xing **group memberships**:

- ~8m users, ca. 42% unique
- 2.9 collisions for 90% of users

Wondracek et al., 2010



Firefox 23.0

$C \cdot 10^9$

Fonts: Arial, sans-serif, Comic Sans, ...

Timezone: -60

1280x1024

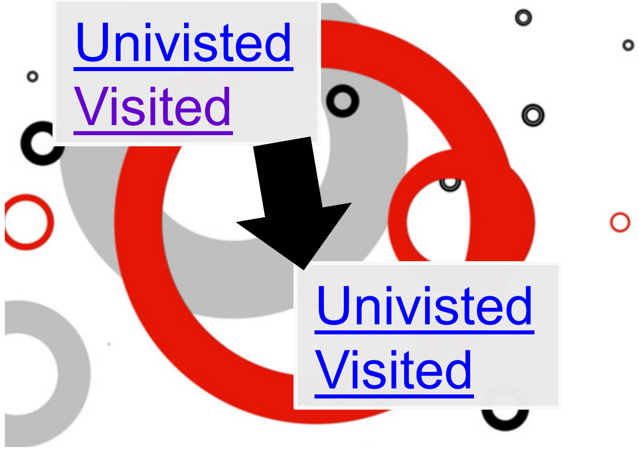
Boda et al., 2011

Stylometric profiling on blogs:

- unstructured data
- 100,000 blogs, cross-context
- ~33% TPR with avg. 20 post / author
- manual inspection!

Eckersley, 2010

Narayanan et al., 2012



Univisted

Visited

Univisted

Visited

Fingerprinting:

- 2010, browser fingerprint (e.g., accuracy: 94.2%)
- 2011, system fingerprint
- 2012, connecting personal devices
- Biometric fingerprinting?

Trends in identification, deanonymization (3)

Network alignment on **temporal location information and social networks** with 80% TPR.

Srivatsa & Hicks, 2012

GSM

Wi Fi

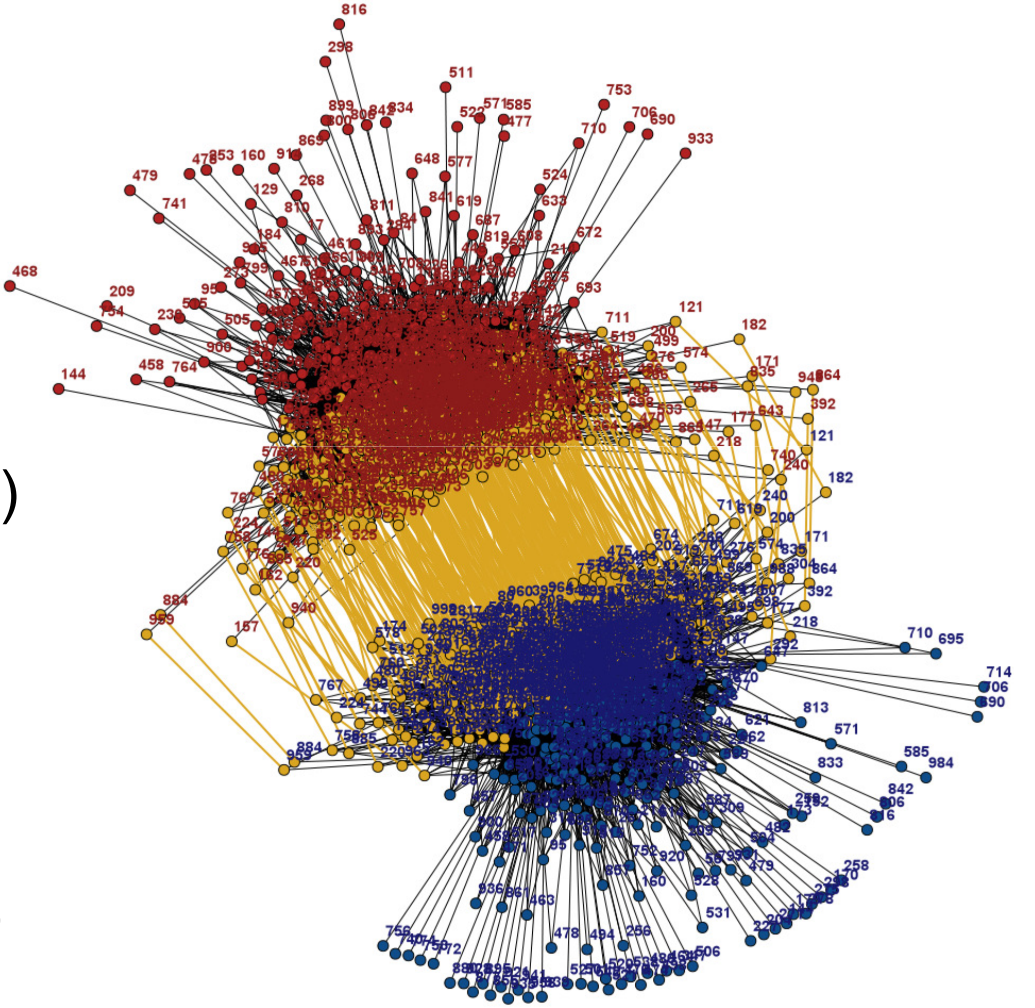
?

Sum of problems: privacy in large databases?

- Basic problem:
  - 7 billion -> 33 bit of information enough
- Problems in large databases:
  - Sparsity: k-anonymity fails
  - Low similarity of items: heavy tail distribution of used attributes
- Pro's and con's:
  - Publishing (anonymous) databases is good for research
  - Breakability of anonymization schemes? Provability?
  - We have some ideas, but not there yet (privacy vs. usability).
  - But we also have wholesale surveillance, thus one should prepare for attackers with strong auxiliary data!

Deanonymizing social networks

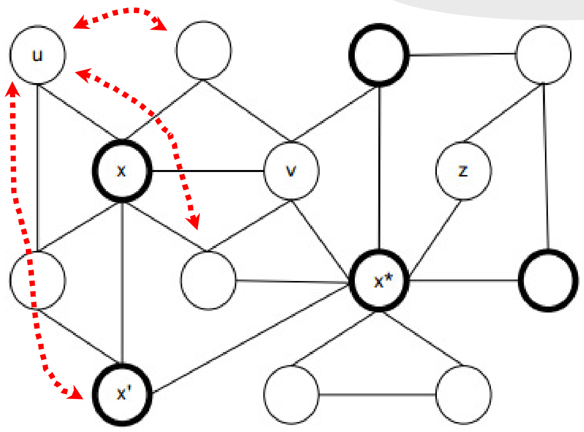
- Underlying concepts work on large social networks
  - Auxiliary data: Flickr (3,3m ns, 53m es)
  - Target (anon.) data: Twitter (224k ns, 8,5m es)
  - Ground truth: 27k nodes (name/user/loc.)
- Results
  - 30% TP, only 12% FP
  - (Init: 150 highdeg. seeds)



Narayanan & Shmatikov, 2009

How to defeat deanonymization?

Beato et al., 2013

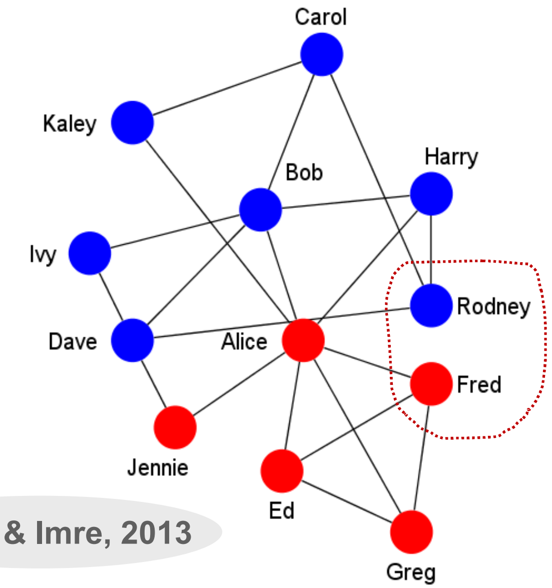


Identity separation:

- no cooperation
- info revealed ~ |Y|
- decoy identities: tricks the attack!
- with cooperation 3% of users are enough

Friend-in-the-middle model:

- requires cooperation of users
- 10% of users are enough
(or maybe less)



Gulyas & Imre, 2013

Thank you for your attention! Questions?

Gábor György Gulyás, Levente Buttyán

gulyas@crysys.hu, buttyan@crysys.hu

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.