HIDING INFORMATION IN SOCIAL NETWORKS FROM DE-ANONYMIZATION ATTACKS BY USING IDENTITY SEPARATION

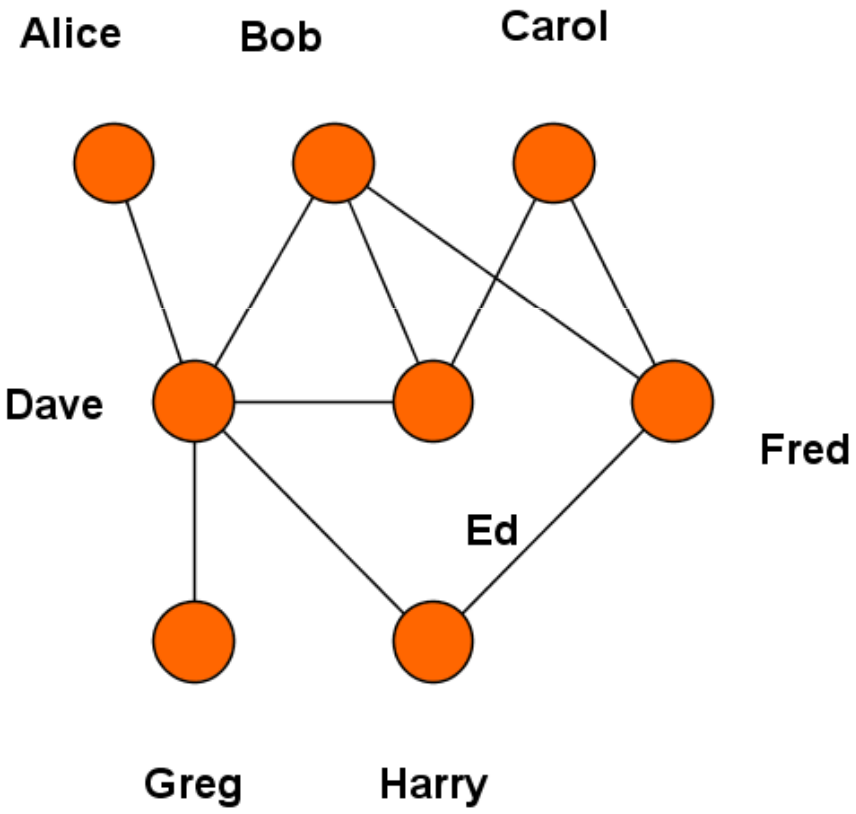
Gábor György Gulyás and Sándor Imre
Dept. Networked Systems and Services (BME)
{gulyasg, imre}@hit.bme.hu

Conference on Communications and Multimedia Security 2013

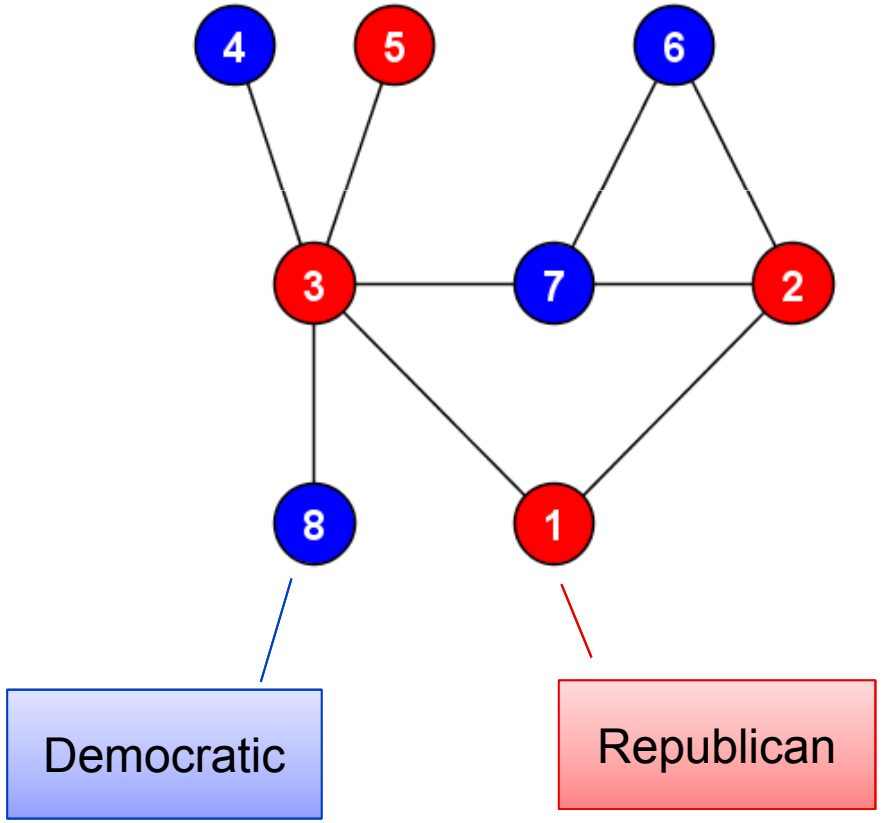
Magdeburg,
September 26, 2013.

Problem statement

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



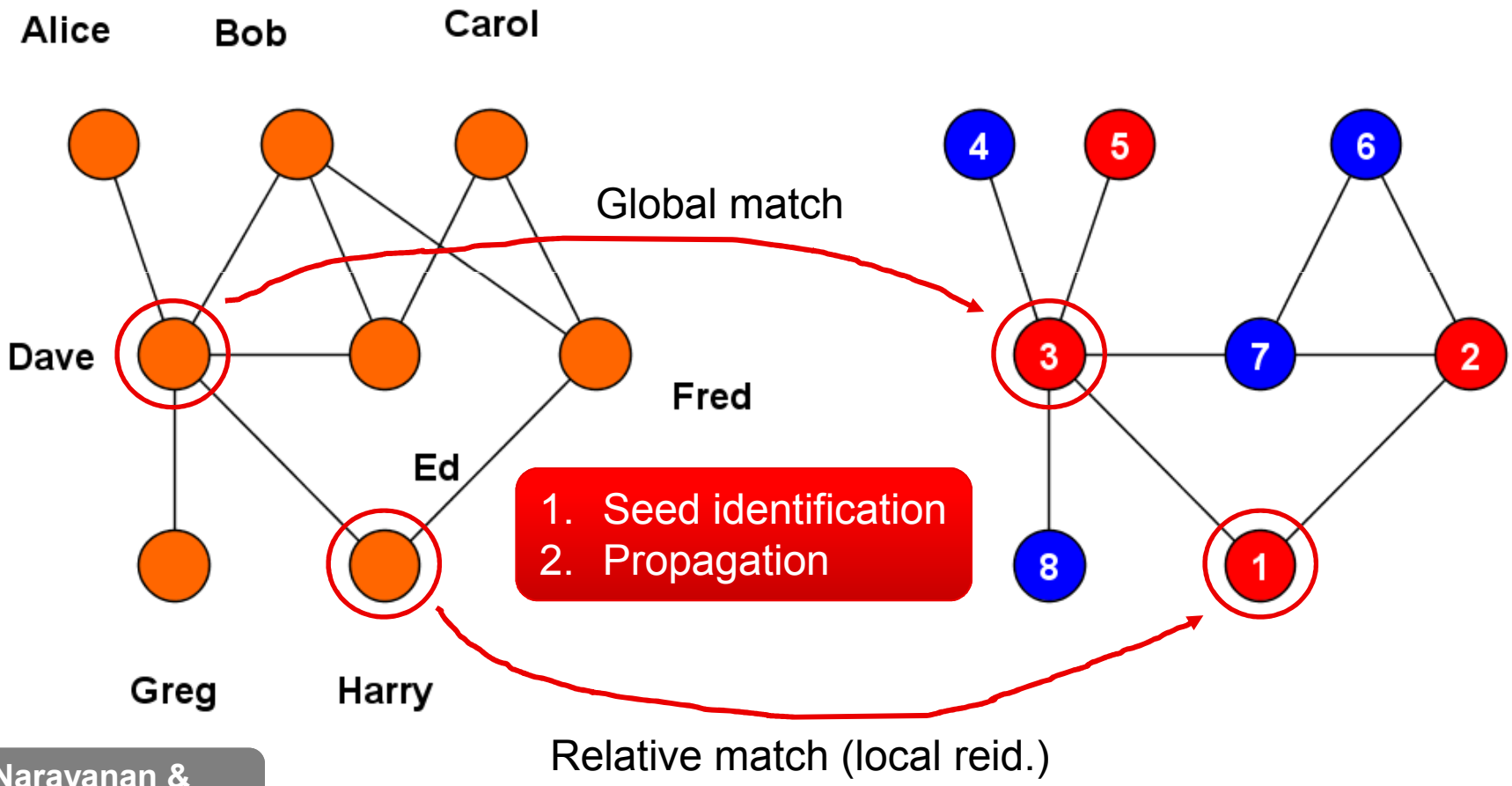
Anonimized graph, G_{tar}
(anonimized export, e.g., Twitter)



Problem statement (2)

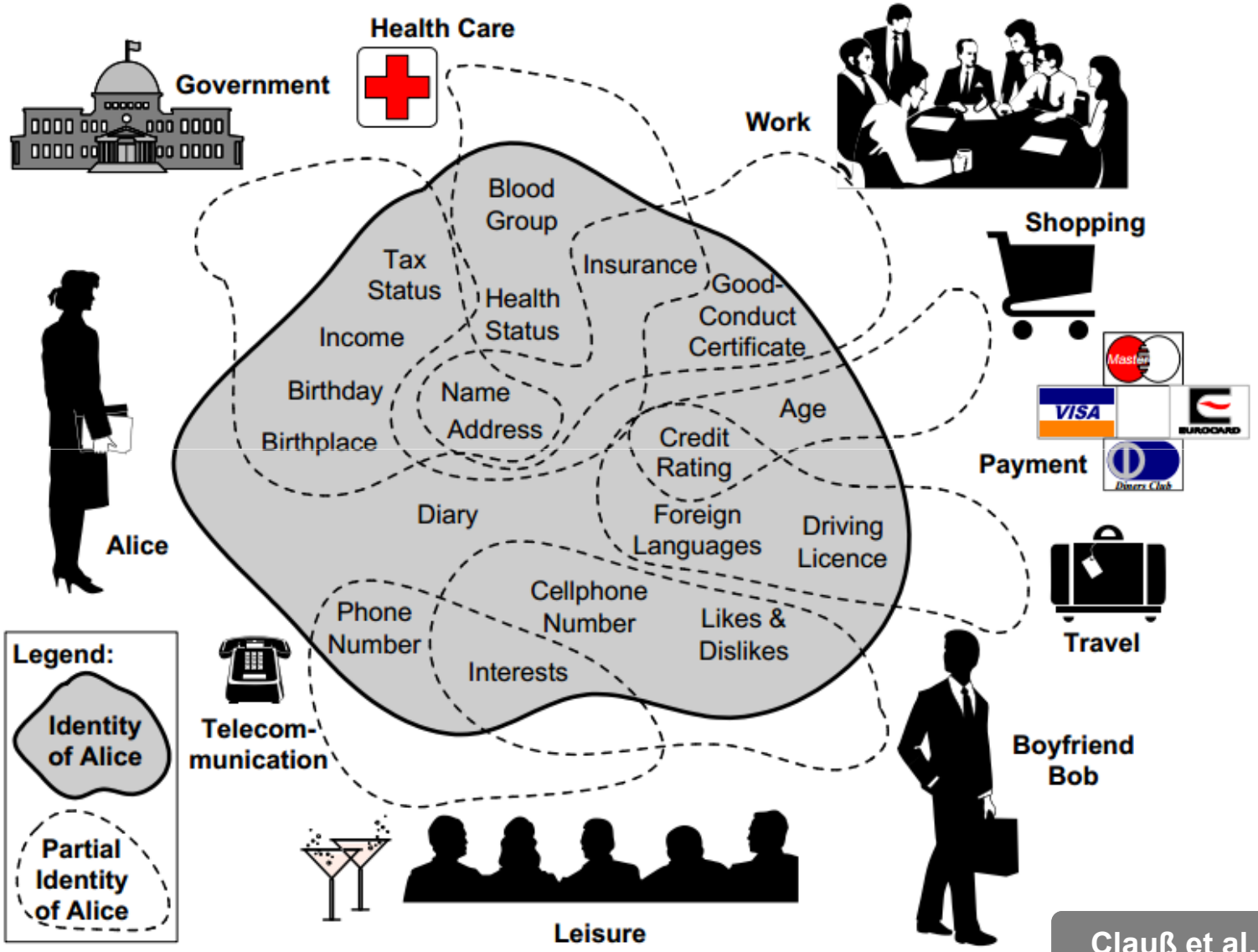
Auxiliary information, G_{src}
(a public crawl, e.g., Flickr)

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



Narayanan & Shmatikov, 2009

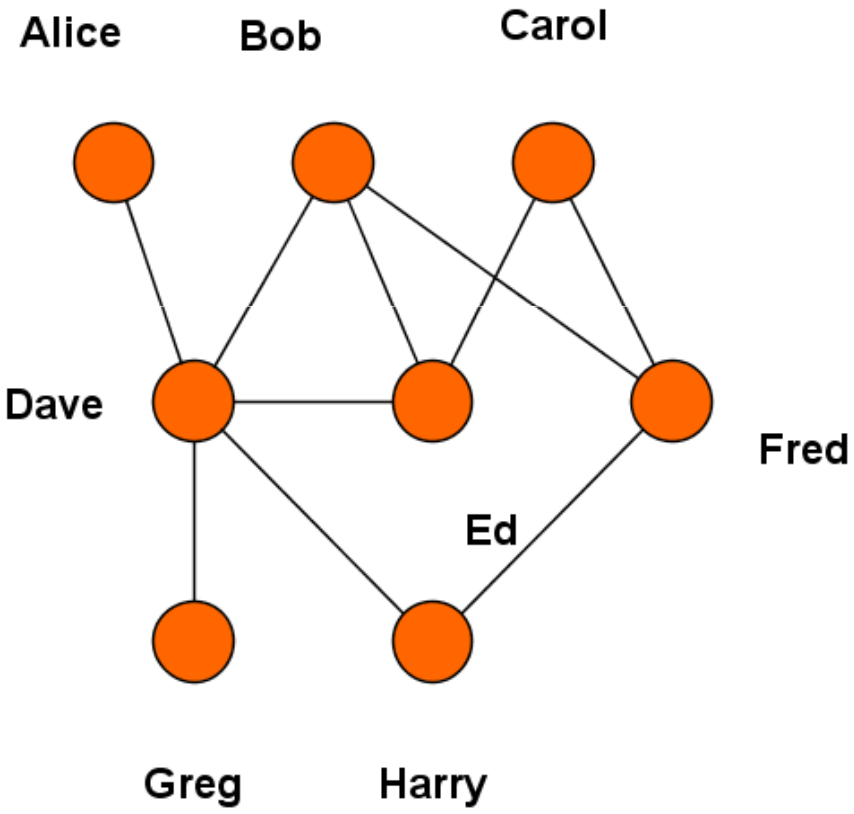
Problem statement (3)



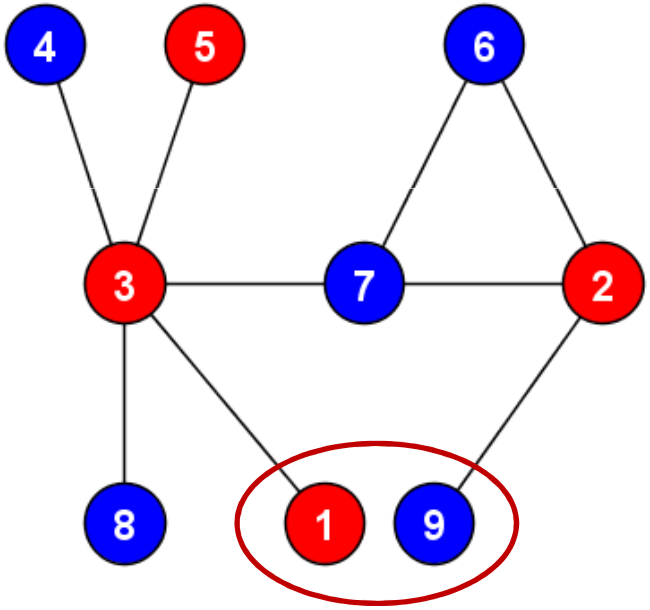
Clauß et al., 2005

Problem statement (4)

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



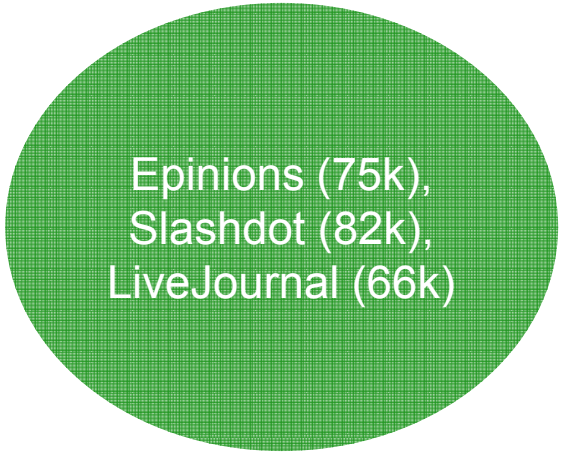
Anonimized graph, G_{tar}
(anonimized export, e.g., Twitter)



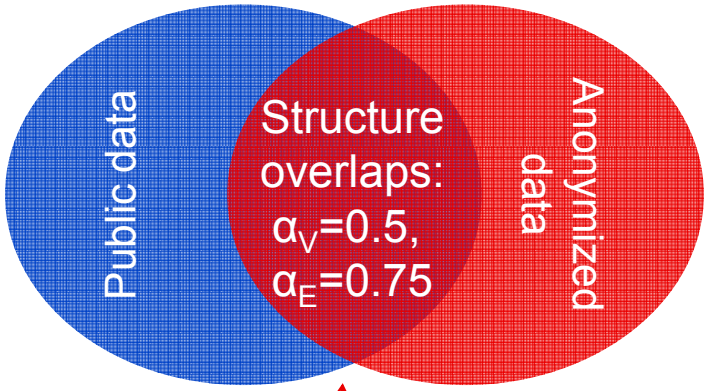
Using identity separation

Simulation: data preparation

Step 1: anonymized network

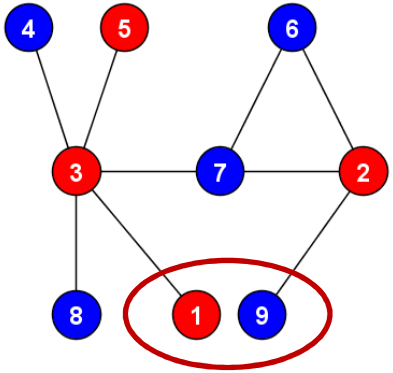


Step 2: perturbation



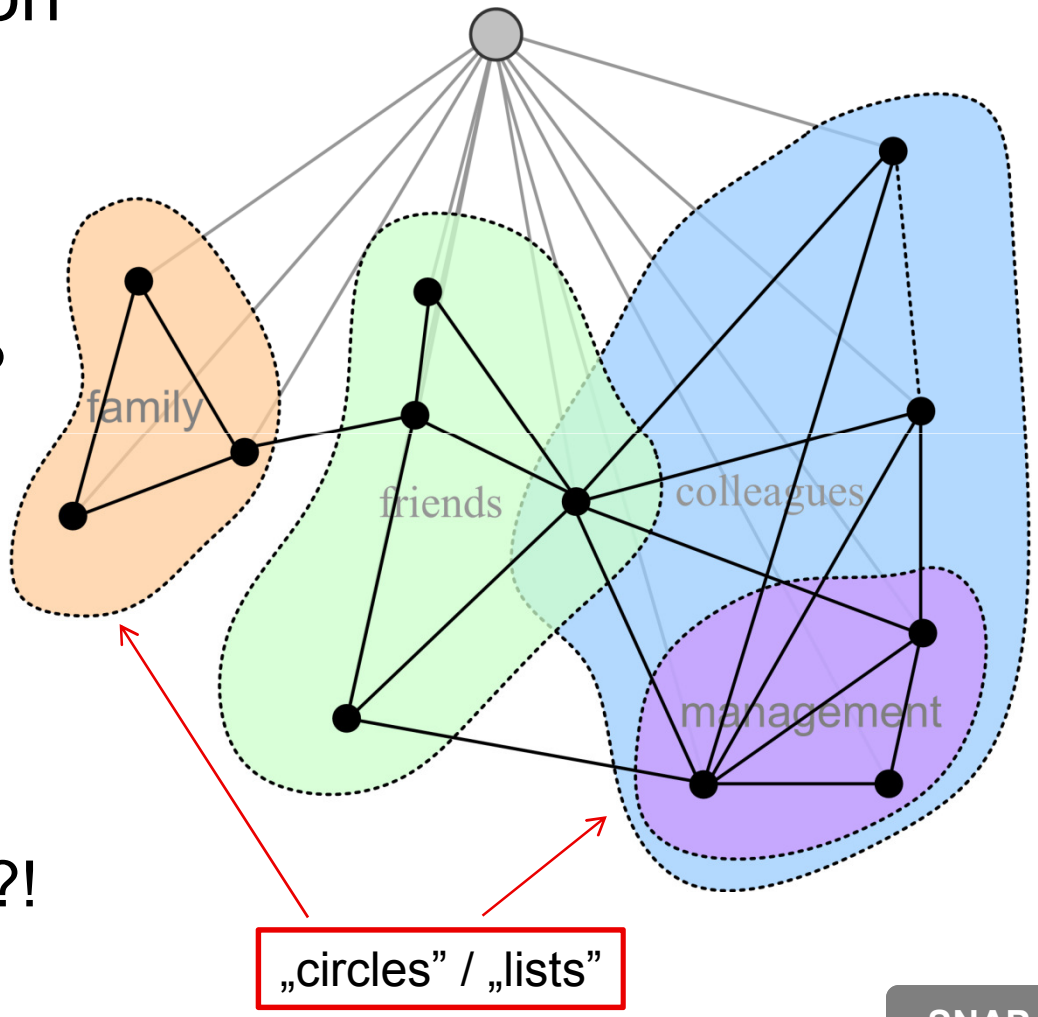
ground truth

Step 3: simulating identity separation



How to model identity separation?

- We have data only on *ego networks* from: Google+, twitter, facebook
- What can we know?
  - Num of circles is power-law (α=2.31)
  - Friend duplication:
    - None: 44.6%
    - 2x+: 6.07%
  - Hidden connections?!



SNAP

How to model identity separation? (2)

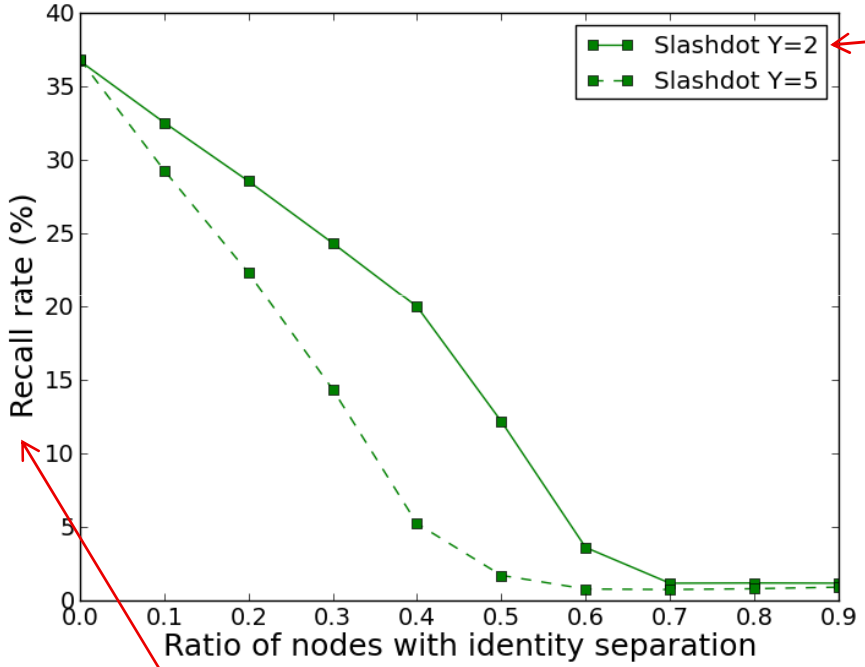
Gulyas & Imre, 2011

- Modeled as *splitting vertices* into partial identities
  - Number of new ids.: random variable Y
  - Edge sorting distribution:
    - Can connections overlap?
    - Can connection be deleted?

	Overlap	No overlap
Edge deletion	Realistic model	Best model
No edge deletion	Worst model	Basic model

Measuring sensitivity to the number of identities

- Basic model with uniform edge sorting probability



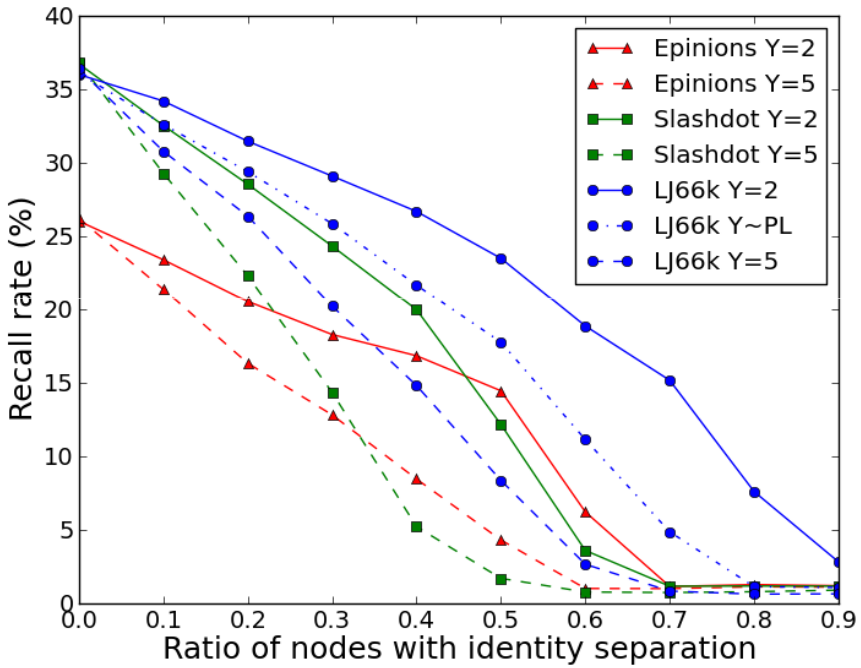
Creating Y=2 new vertices from one, and sorting edges with ½ probability to each.

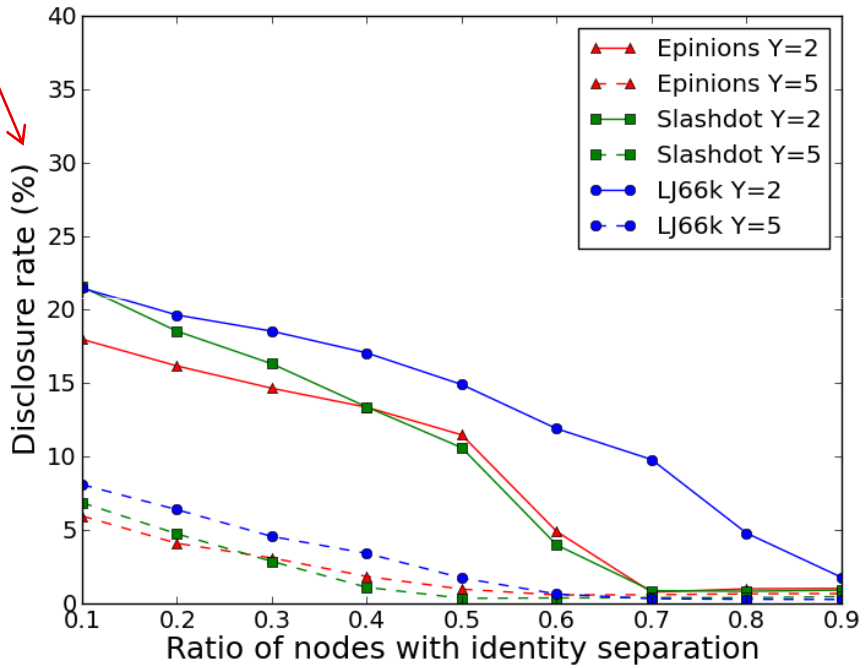
Recall rate: percent of correctly re-identified nodes.

Measuring sensitivity to the number of identities (2)

- Basic model with uniform edge sorting probability

Disclosure rate: what the attacker learns.



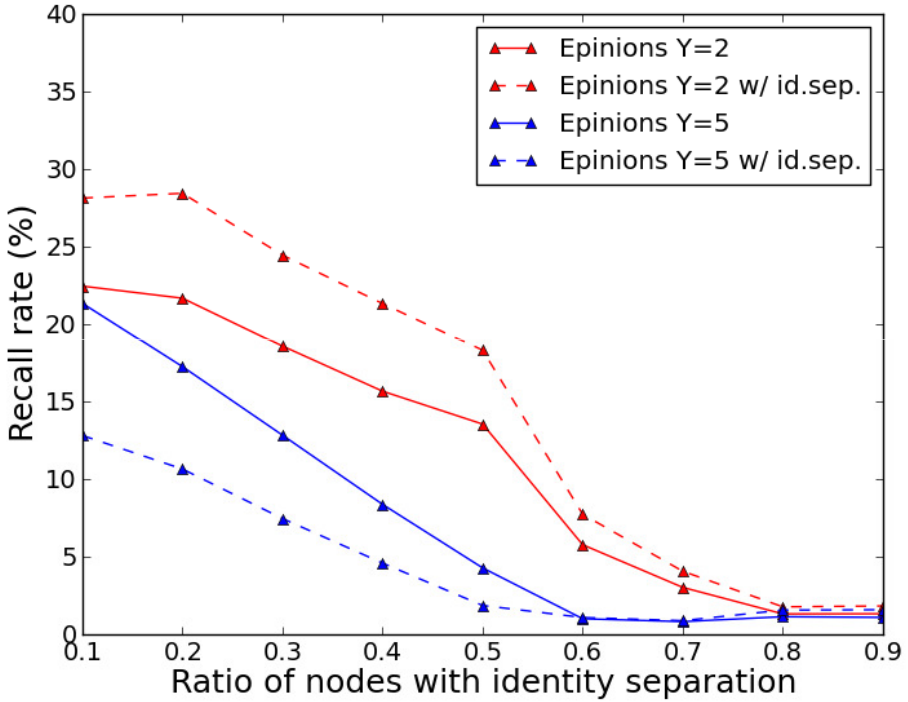


Over all nodes!

Over nodes with identity separation!

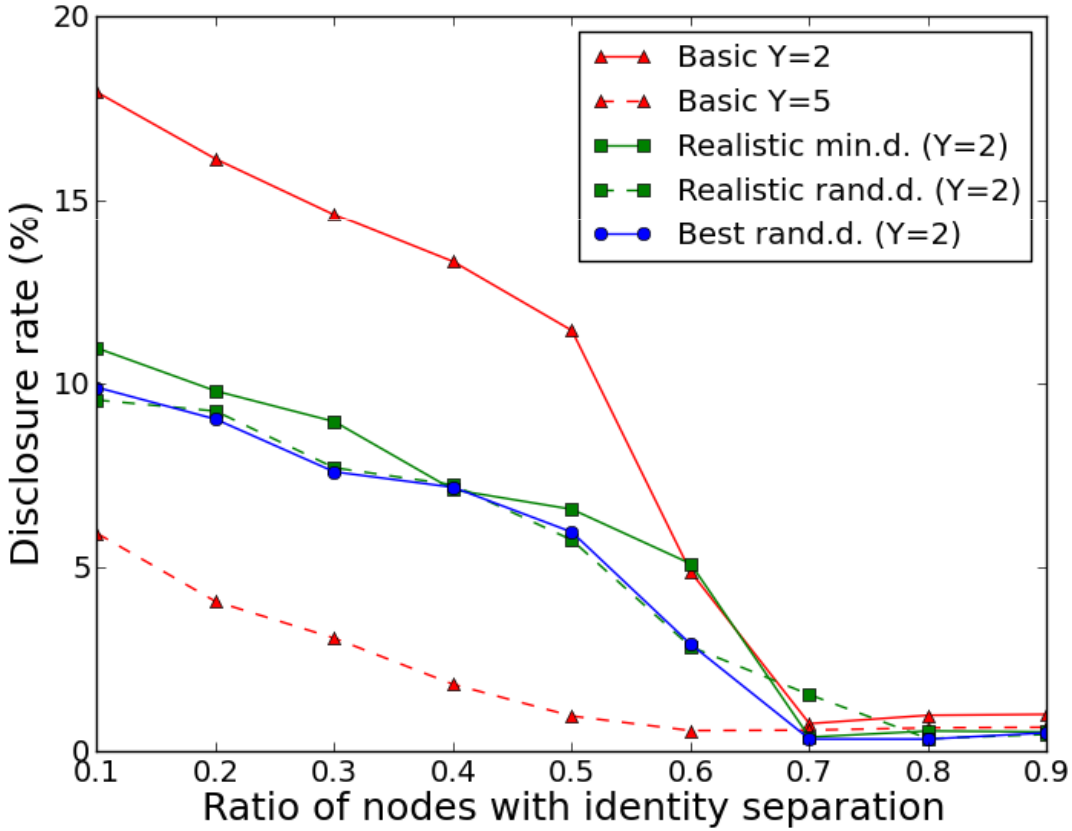
Measuring sensitivity to the number of identities (3)

- 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 →but has bad implications on privacy



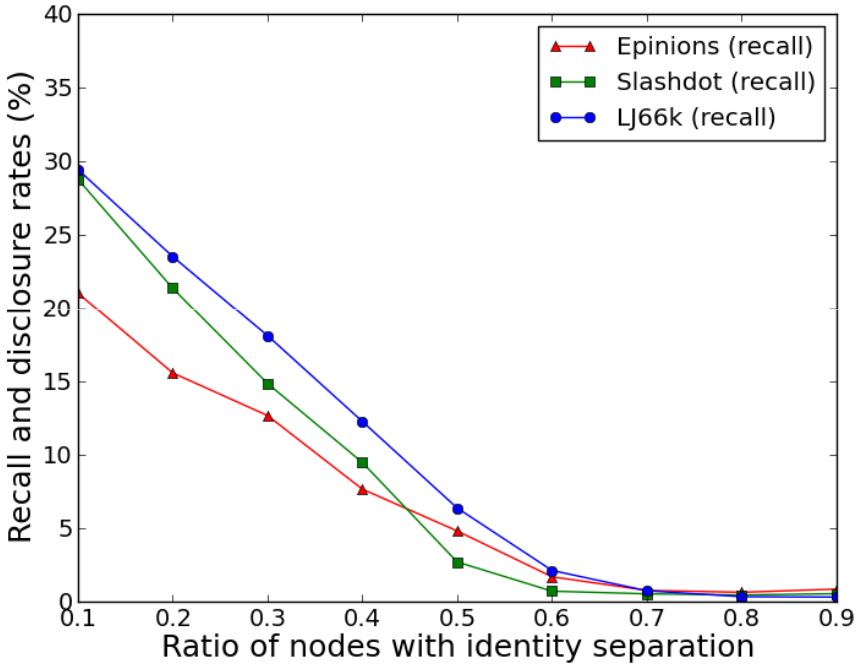
Measuring sensitivity to deletion of edges

- Used models:
  - realistic model with minimal deletion / random deletion
  - basic model with random deletion



In the search of privacy-enhancing methods

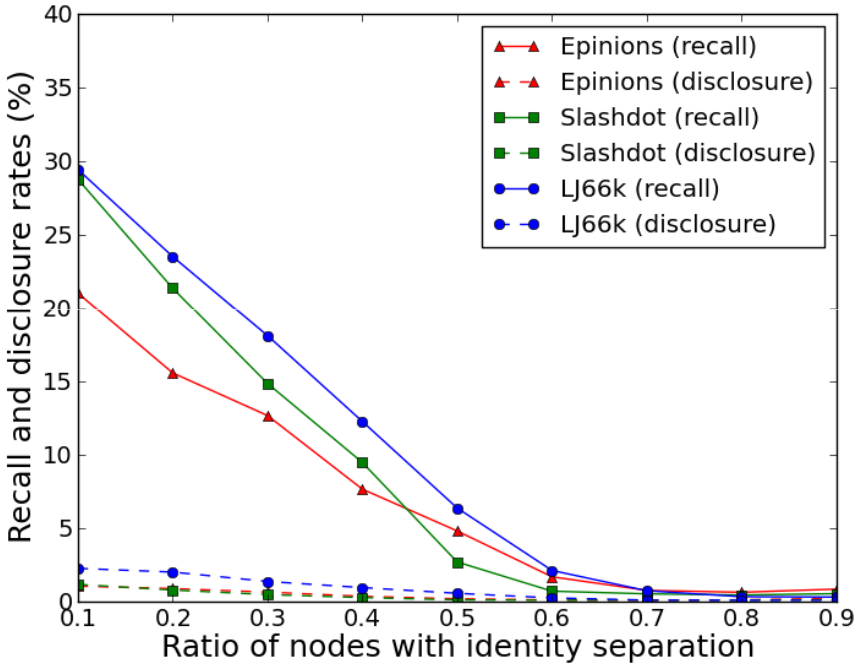
Tackling the attack on a network level?



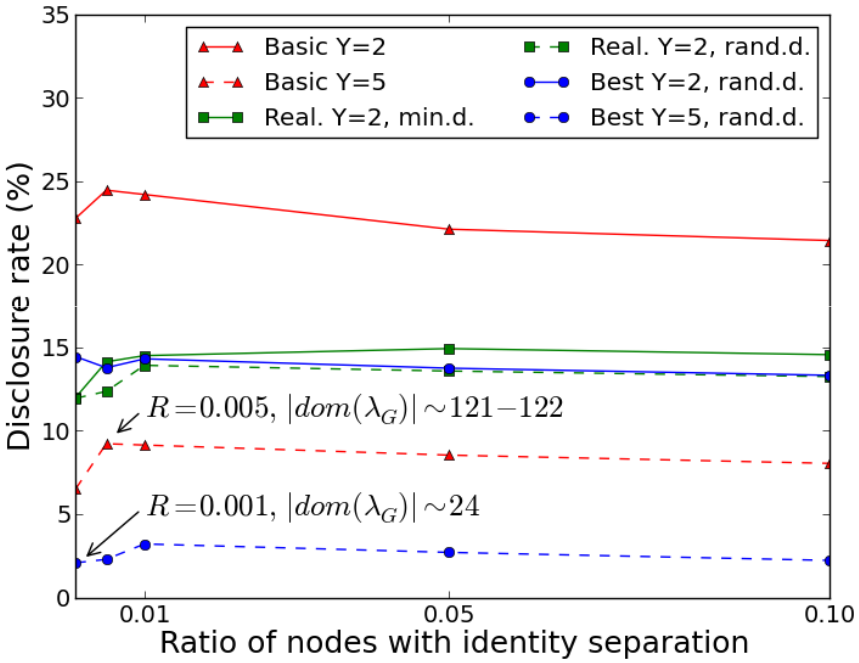
Best model, Y=5, random deletion

In the search of privacy-enhancing methods (2)

Tackling the attack on a network level?

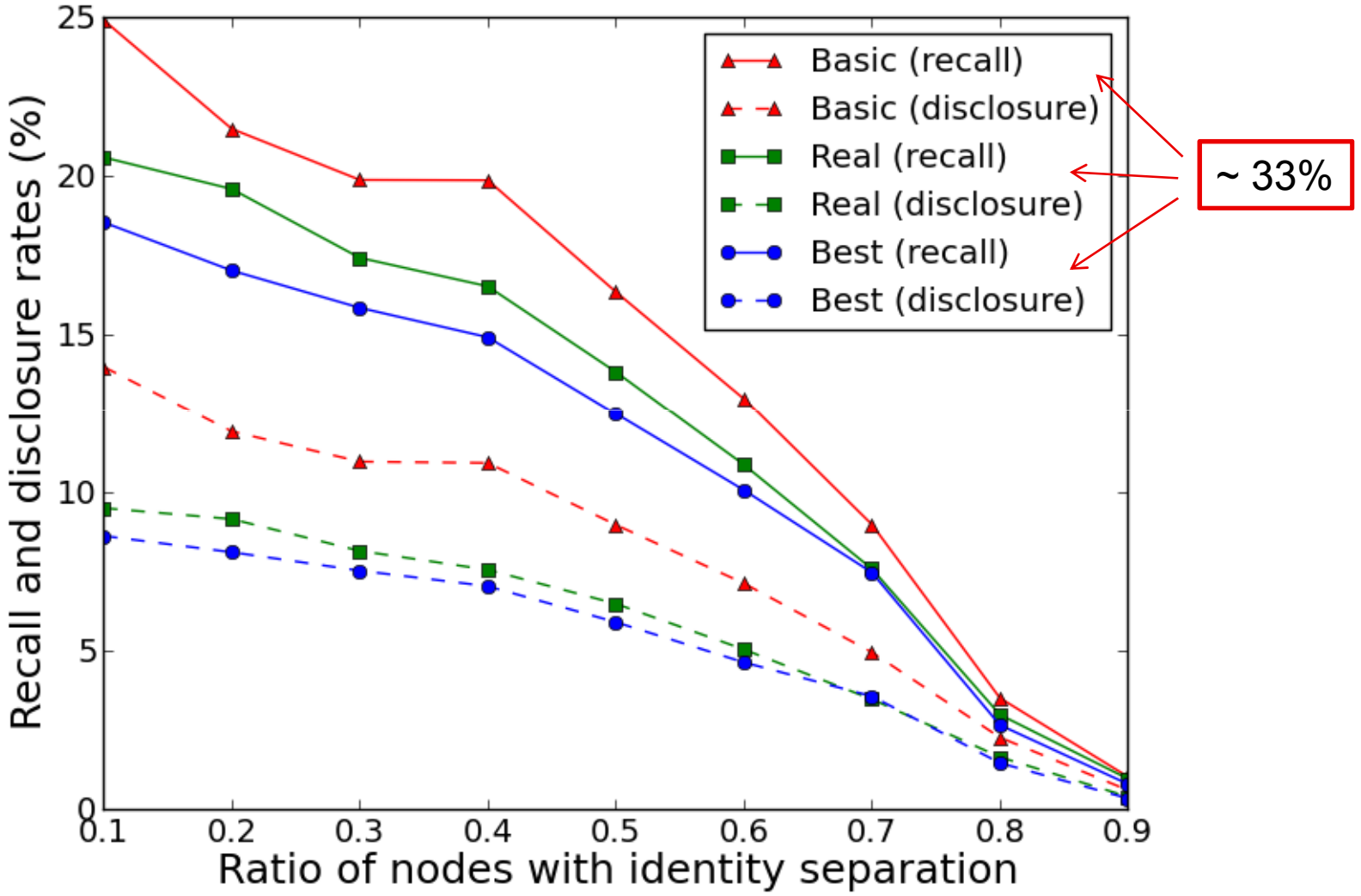


What if only few users care?



Best model, Y=5, random deletion

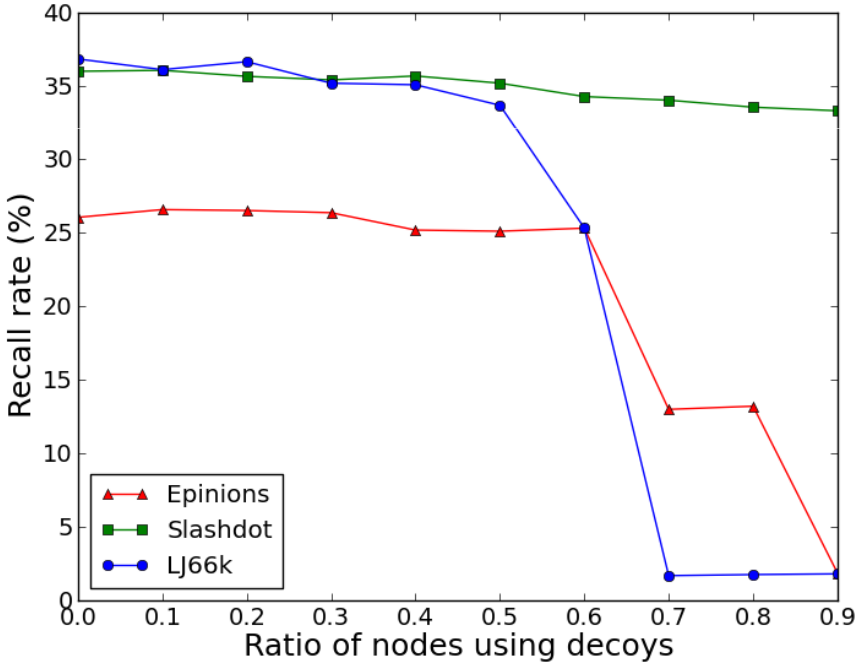
Multiple models present in parallel

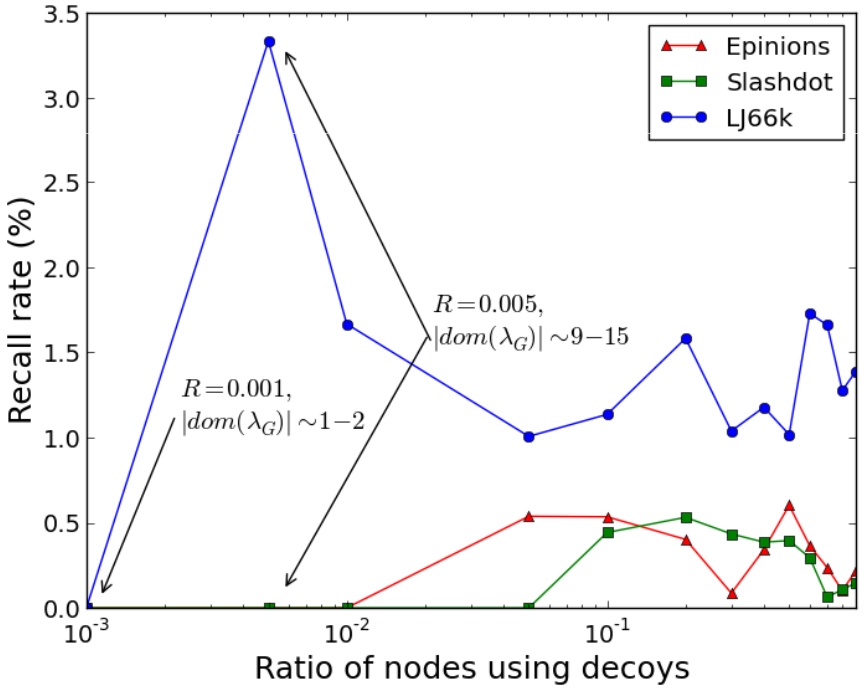


With Y=2 in the LiveJournal network.

Using decoy identities

- Goal: to control what the adversary can discover
  - Decoy identity: a public profile with most connections (90%)
  - Hidden identity: having a few connections (20% with 10% overlap)





Future work

- Is it possible to tackle the attack with only a few nodes involved?
- Using identity separation in both networks?
- Enhancing the decoy method

Questions?

THANK YOU FOR YOUR ATTENTION!

Department of Networked Systems and Services

Gábor György Gulyás
assistant research fellow
Dept. Networked Systems and Services (BME)
gulyasg@hit.bme.hu