Protecting Privacy Against Structural De-anonymization Attacks in Social Networks

Public defense of

Gábor György Gulyás

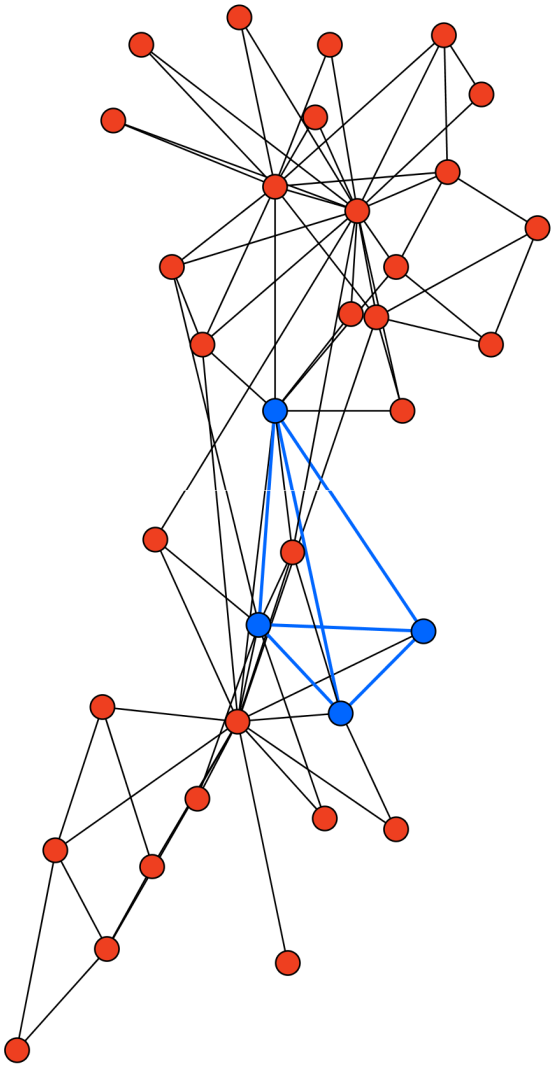
Supervisor: *Sándor Imre, DSc*
Reviewers:
Gergely Biczók, PhD (BME-TMIT)
Julien Freudiger, PhD (PARC, USA)

MOTIVATION

Structural privacy in social networks

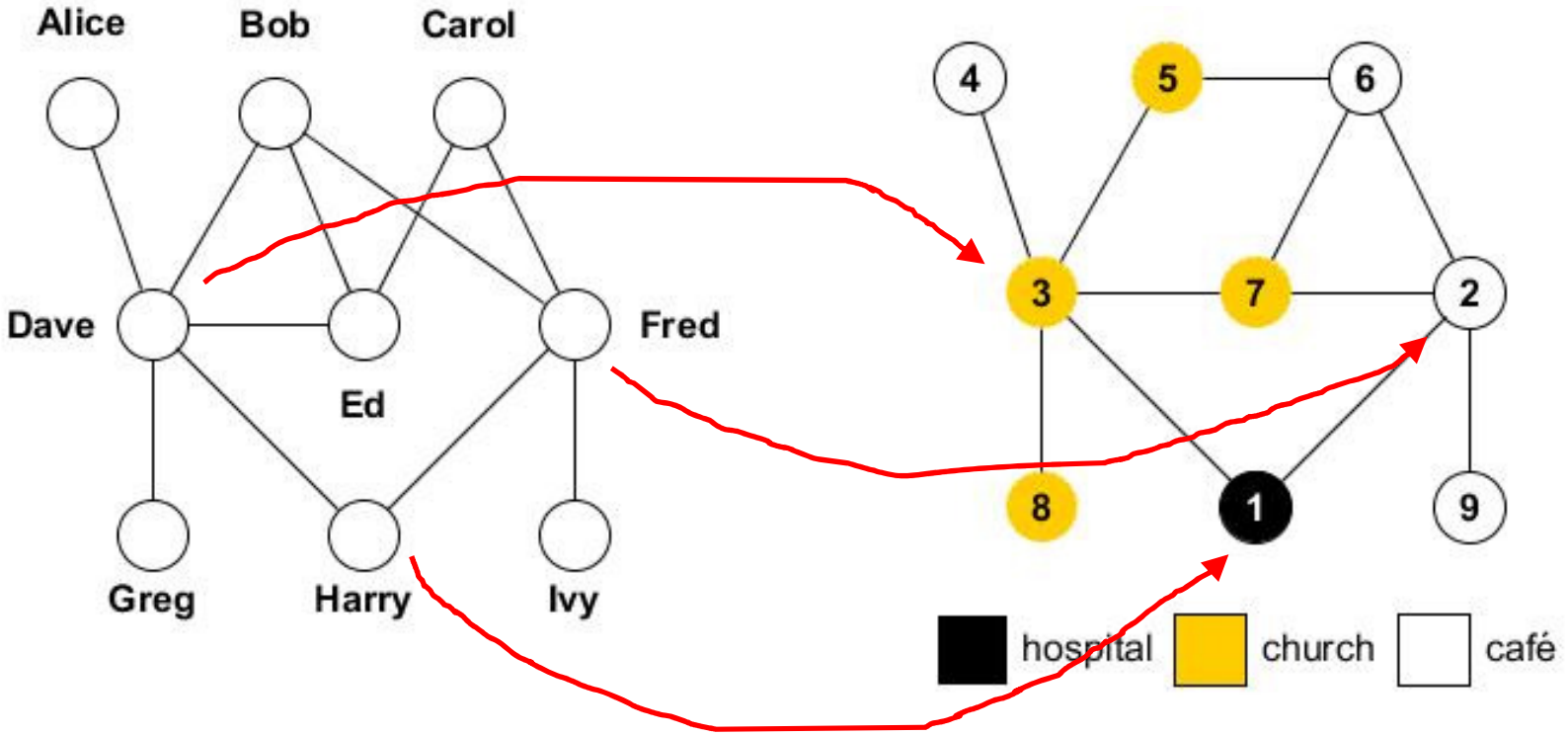
- One of most challenging privacy issues: how structural information can be hidden

- Can be leaked [anonymously] easily
  - TP businesses
  - TP research partners
  - Implicit social networks
    - E-mail networks
    - Call information
    - ...



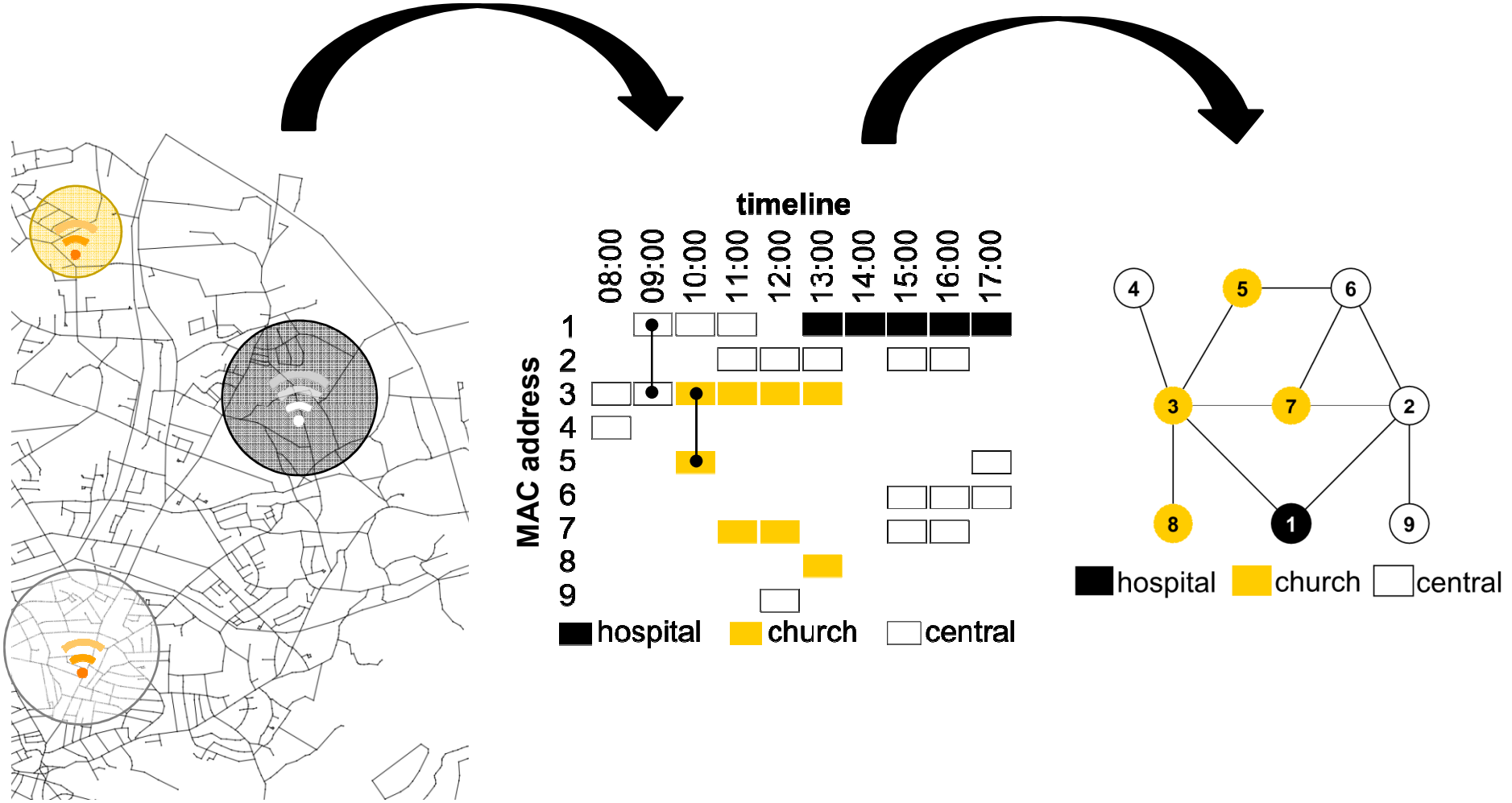
Structural privacy in social networks (2)

Phases:
1. Initialization (seeding)
2. Propagation



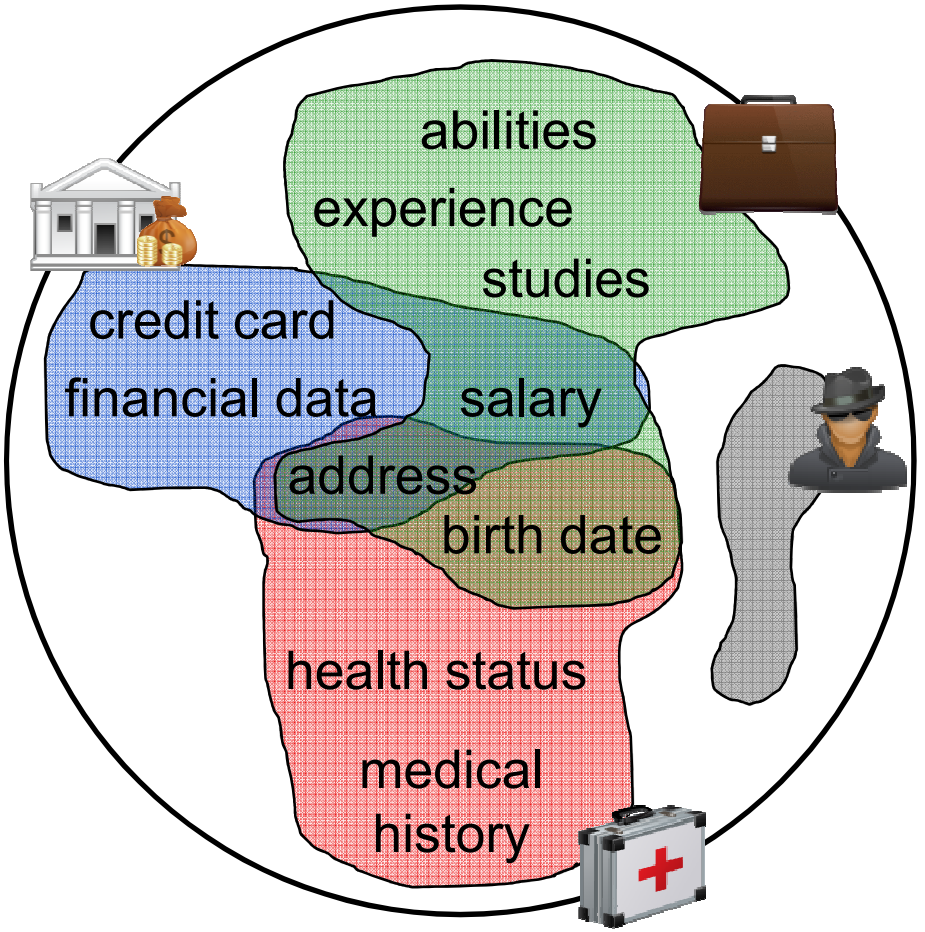
Goals:
- Attacker: achieve high recall
- Users: minimize network or individual information disclosure

Structural privacy in social networks (3)



How to protect user privacy?

- Proposed solution:
  - Identity separation
- Desired properties?
  - Work with existing services, gradually adoptable
  - User control: client side solutions
    - Without consent of any service providers
    - All actions on client side



Global and partial identities of John Doe

NEW RESULTS

Problem sets of the dissertation

Analysis of re-identification attacks

- Measuring anonymity
- Analysis of the initialization of attacks

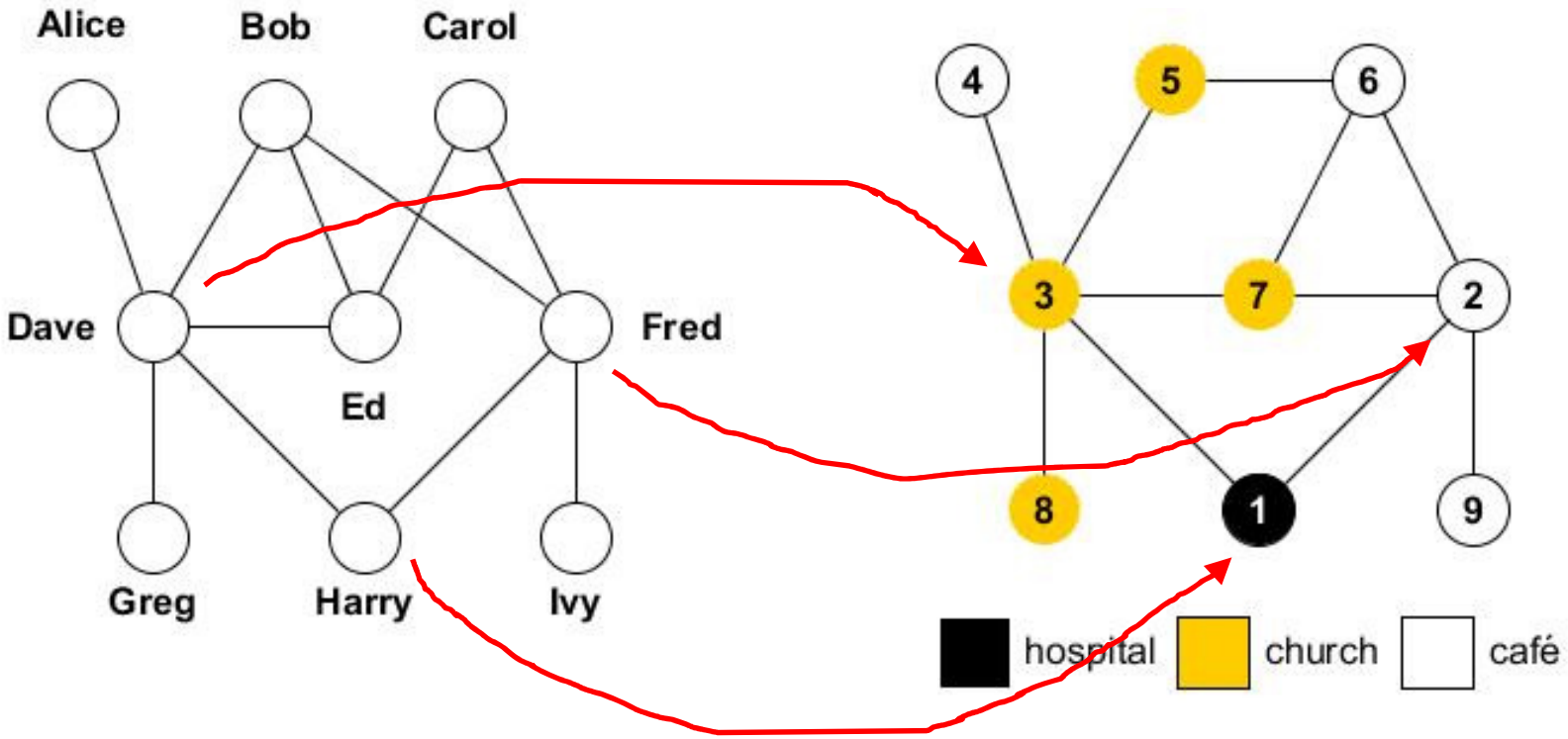
Evaluation of identity separation

- Against attack initialization
- Evaluation of strategies to stop re-identification
  - Individual adoption
  - Cooperation

Individual strategies with identity separation

- Small adoption rates and use of decoys
- Reversebility
- K-anonymity
- Game-theoretic approach

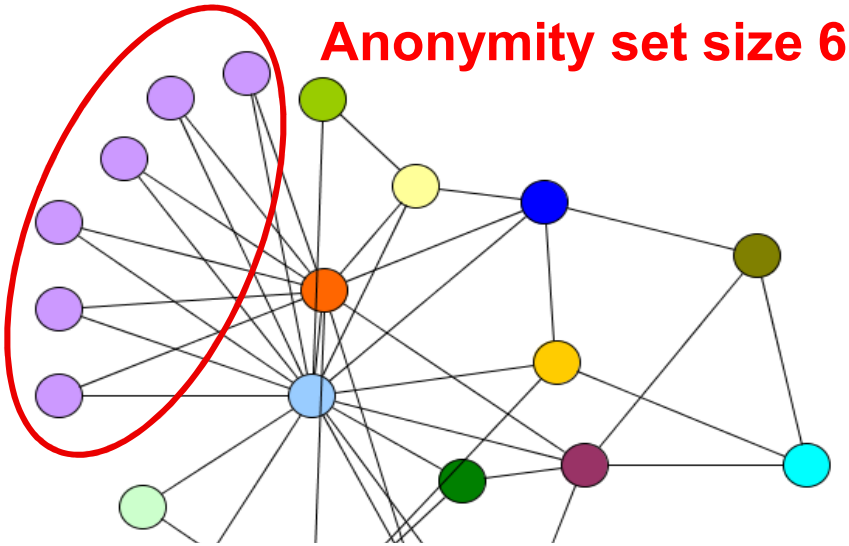
Problem Set 1: *analysis of re-identification attacks*



a. Measuring anonymity?
   - No global anonymity sets

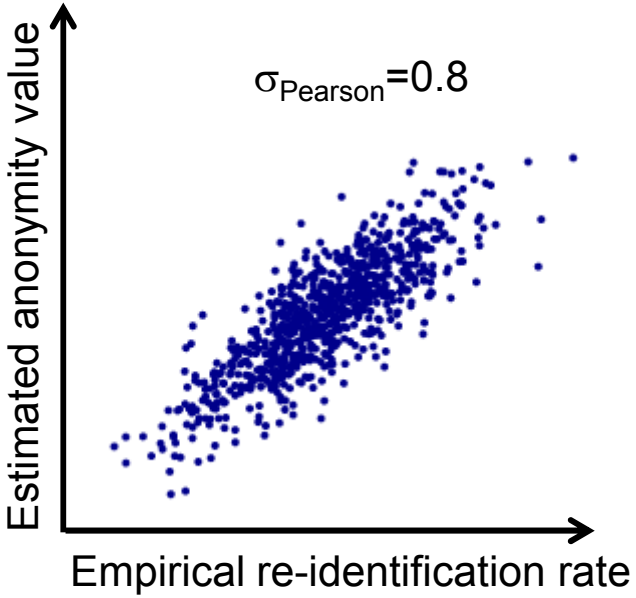
b. 2-phase attacks:
   - seeding (initialization)
   - propagation

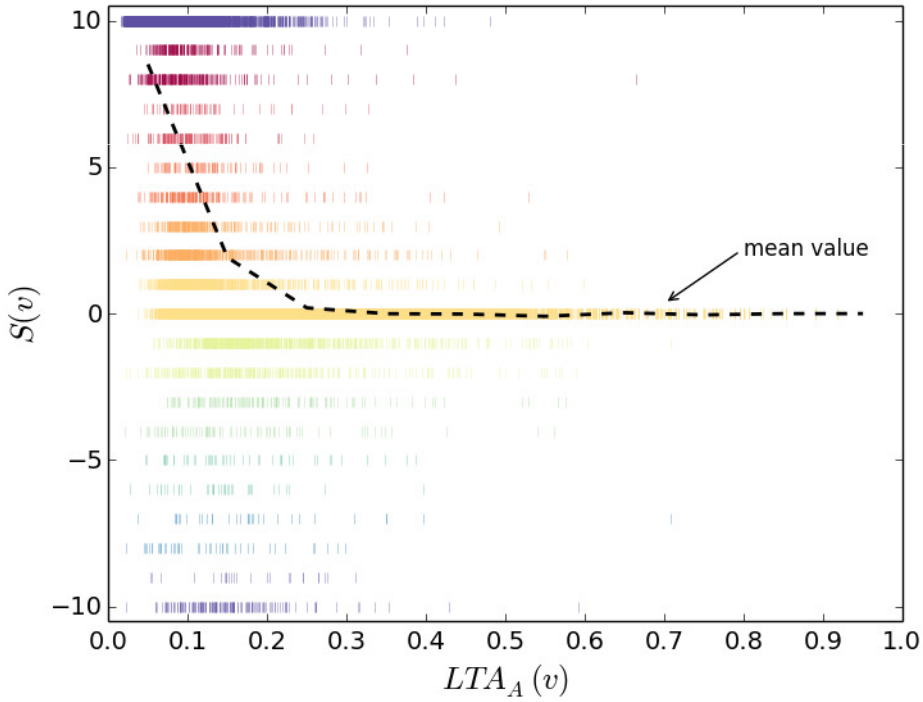
[T1.1-2] Can we measure anonymity?



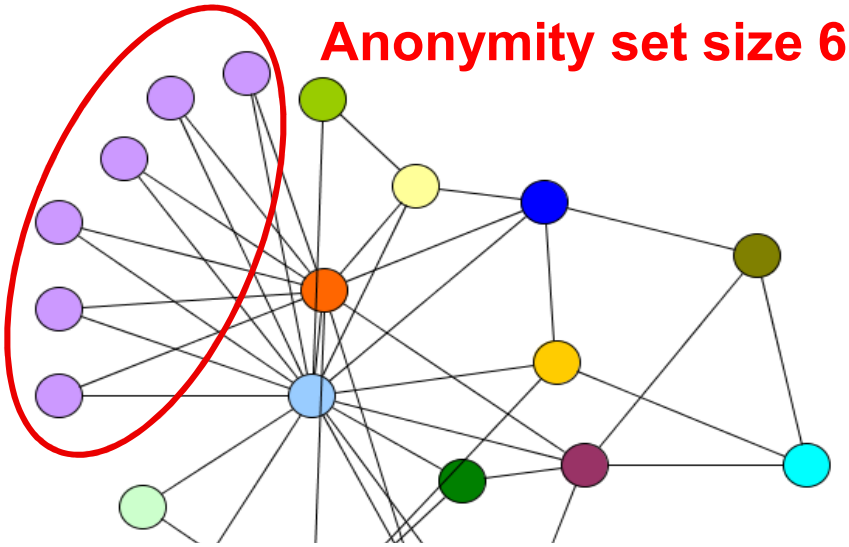
$$\mathrm{LTA}_{deg}(v_i) = \deg(v_i).$$

$$\mathrm{LTA}_{A}(v_i) = \sum_{\forall v_k \in V_i^2} \frac{CosSim(v_i, v_k)}{|V_i^2|},$$



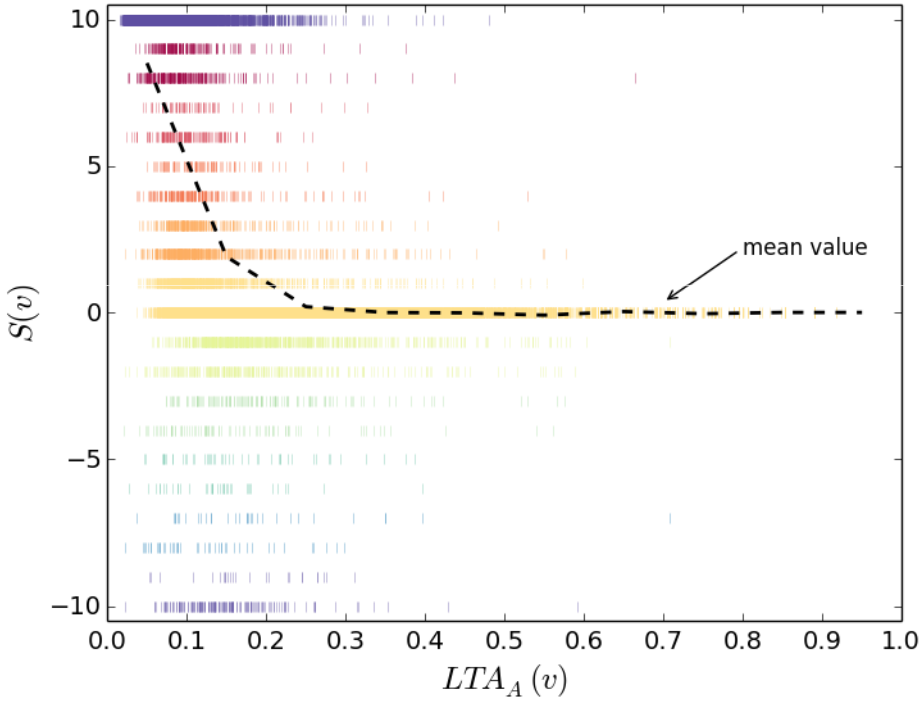


[T1.1-2] Can we measure anonymity?

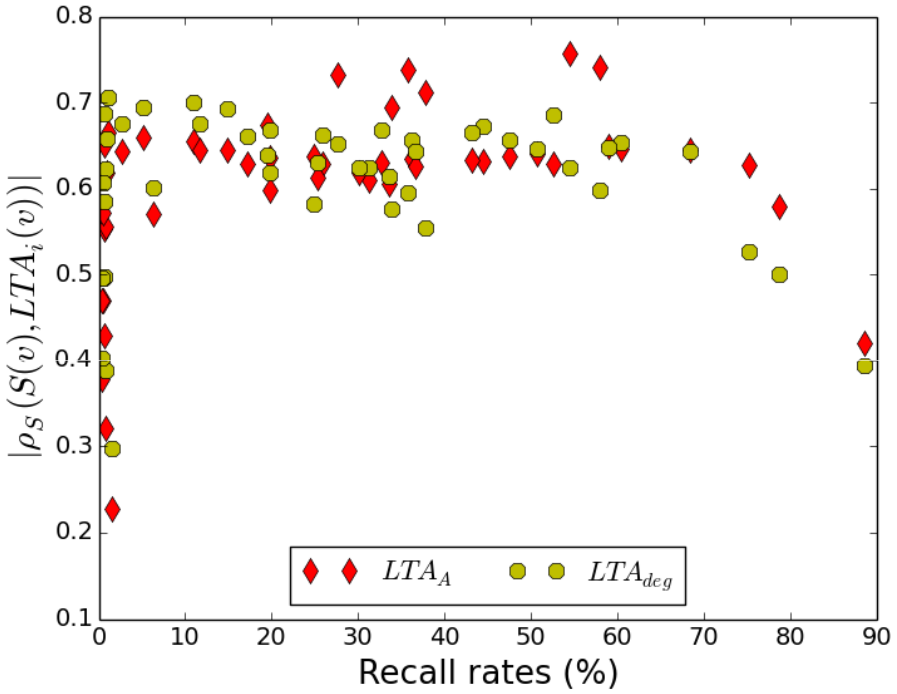


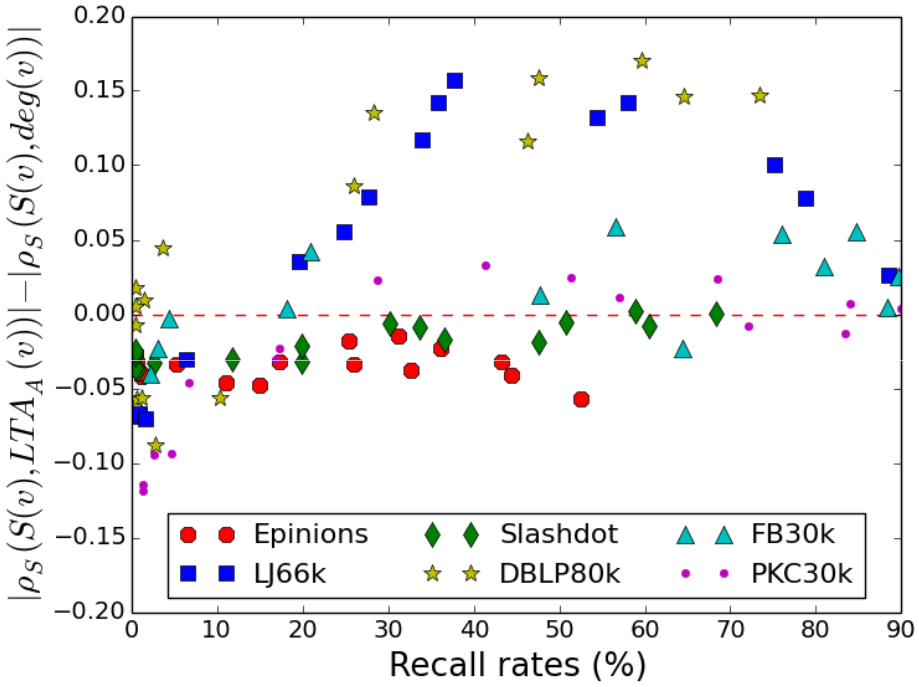
$$\mathrm{LTA}_{deg}(v_i) = \deg(v_i).$$

$$\mathrm{LTA}_{A}(v_i) = \sum_{\forall v_k \in V_i^2} \frac{CosSim(v_i, v_k)}{|V_i^2|},$$



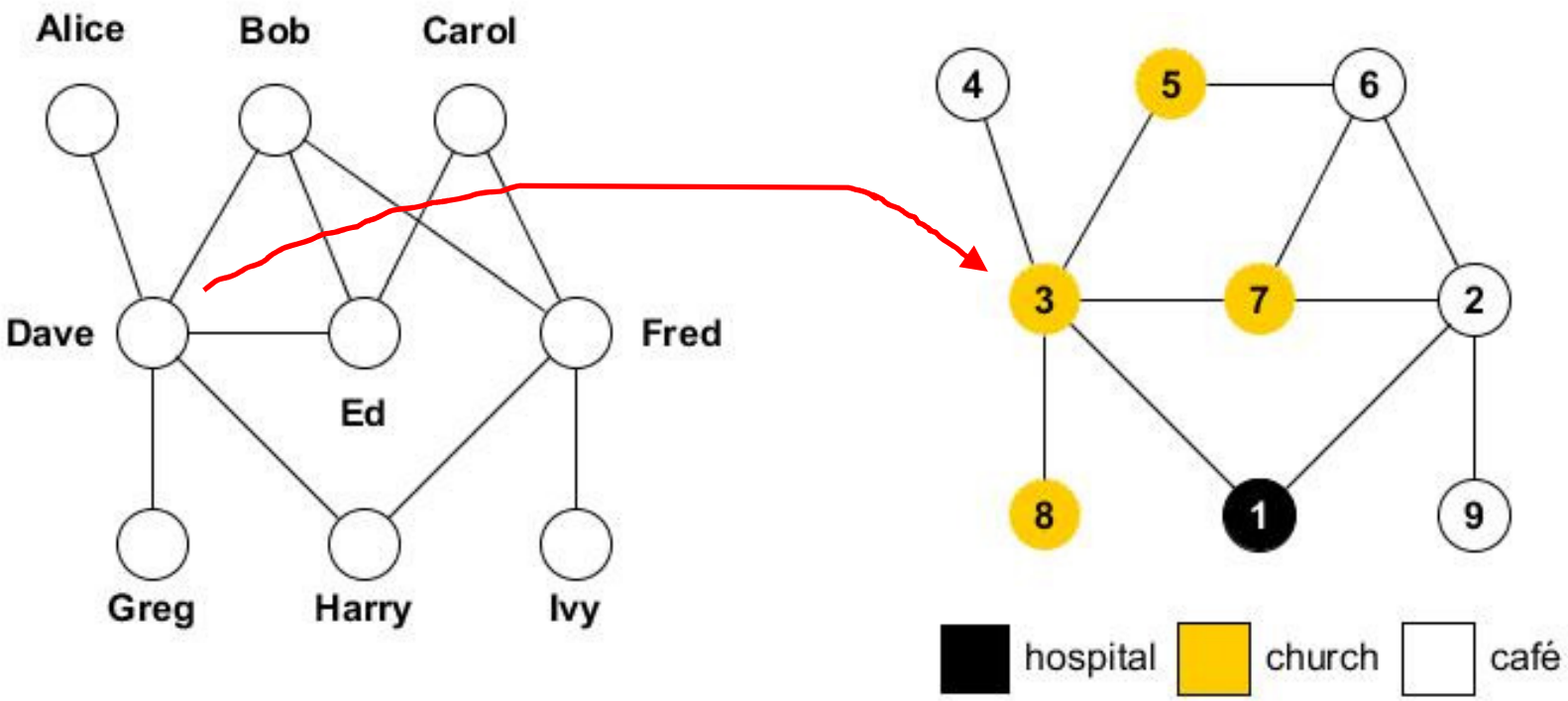
[T1.1-2] Can we measure anonymity? (2)





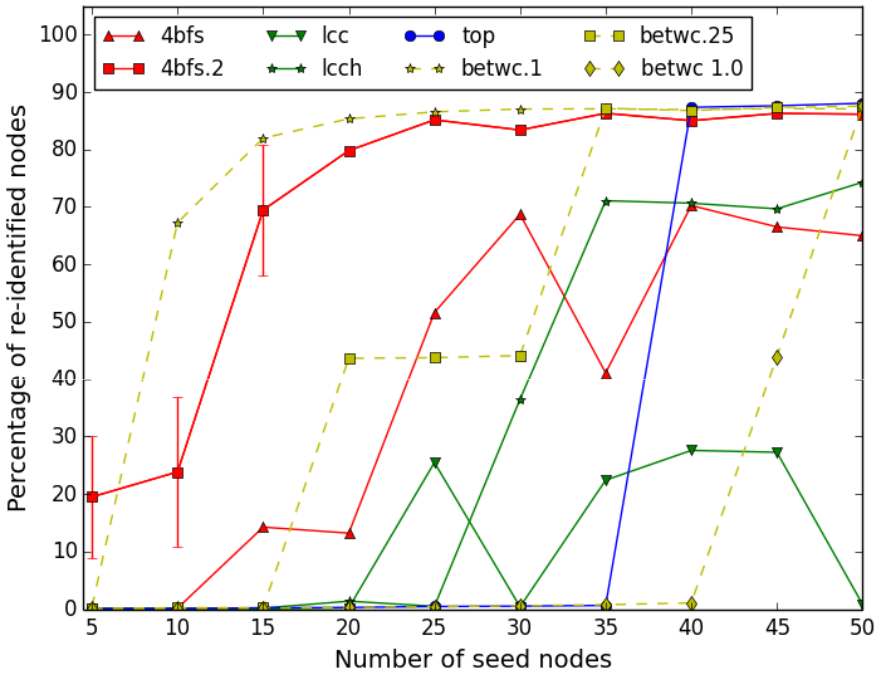
Applications, e.g.:
- individual / provider risk assessment
- as measures of importance

[T1.3] How initialization effects overall results?



[T1.3] How initialization effects overall results? (2)

- For the state-of-the-art algorithm, I characterized the importance of initialization.
- In particular:
  - I showed how the maximum number of re-identified nodes can depend on the seeding method and its parameters
  - I showed how the minimum number of seed nodes depends on network properties and the seeding method
  - I characterized seed stability and showed that even an extremely low number of seed nodes can also lead to large-scale propagation



Seeding method
- should be carefully chosen
- part of the attacker model

Related publications to Problem Set 1

[J2] B. Simon, G. G. Gulyás, and S. Imre, "Analysis of grasshopper, a novel social network de-anonymization algorithm," *Periodica Polytechnica Electrical Engineering and Computer Science*, January 2015. (accepted for publication).

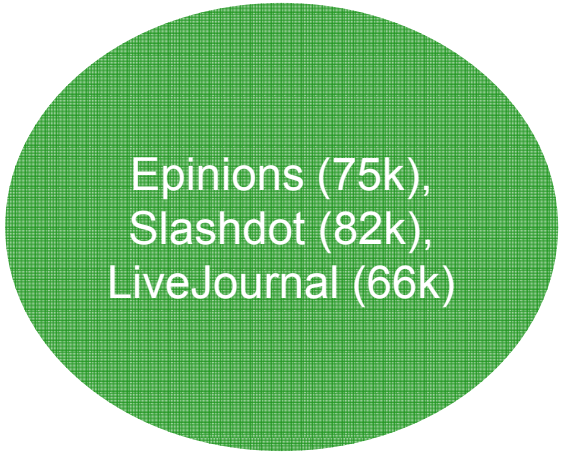
[J3] G. G. Gulyás and S. Imre, "Using identity separation against de-anonymization of social networks," *Transactions on Data Privacy*, January 2015. (accepted for publication).

[C1] G. G. Gulyás and S. Imre, "Measuring importance of seeding for structural de-anonymization attacks in social networks," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, 2014.

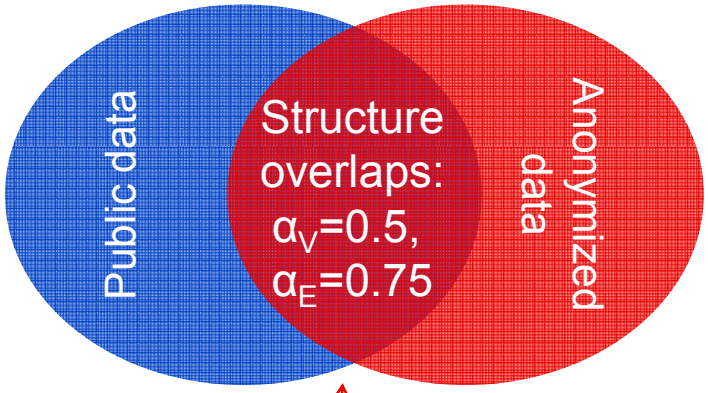
[C3] G. G. Gulyás and S. Imre, "Measuring local topological anonymity in social networks," in *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pp. 563–570, 2012.

Problem Set 2: *evaluation of identity separation*

Step 1: anonymized network



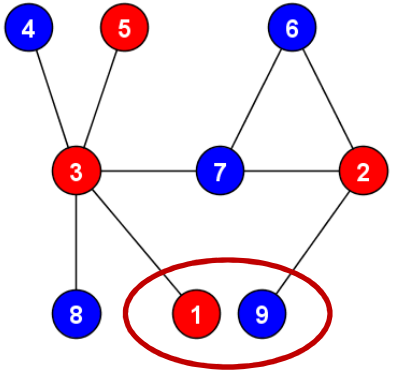
Step 2: perturbation



ground truth

Multiple models
- Number of identities (*Y*)
- Deleting edges, duplicating edges

Step 3: simulating identity separation

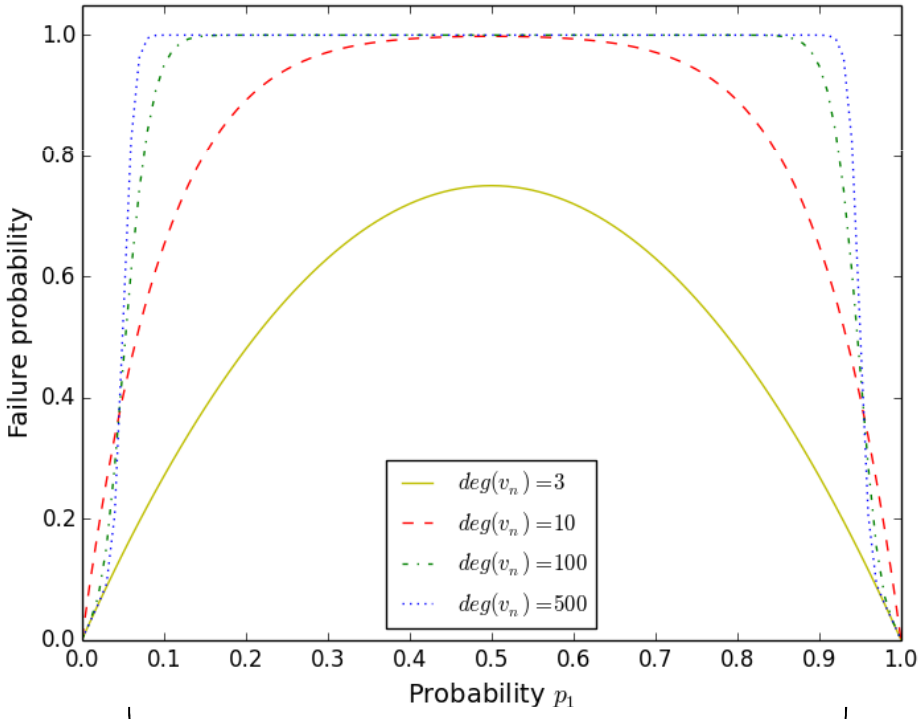


[T2.1] Can we estimate failure probability of seeding?

By providing the general formula of failure probability, I elaborated the lower estimate of failure probability for clique-based, and top node based seeding:

$$P^{B}_{clique}(\text{"failure"}|Y=y) =$$

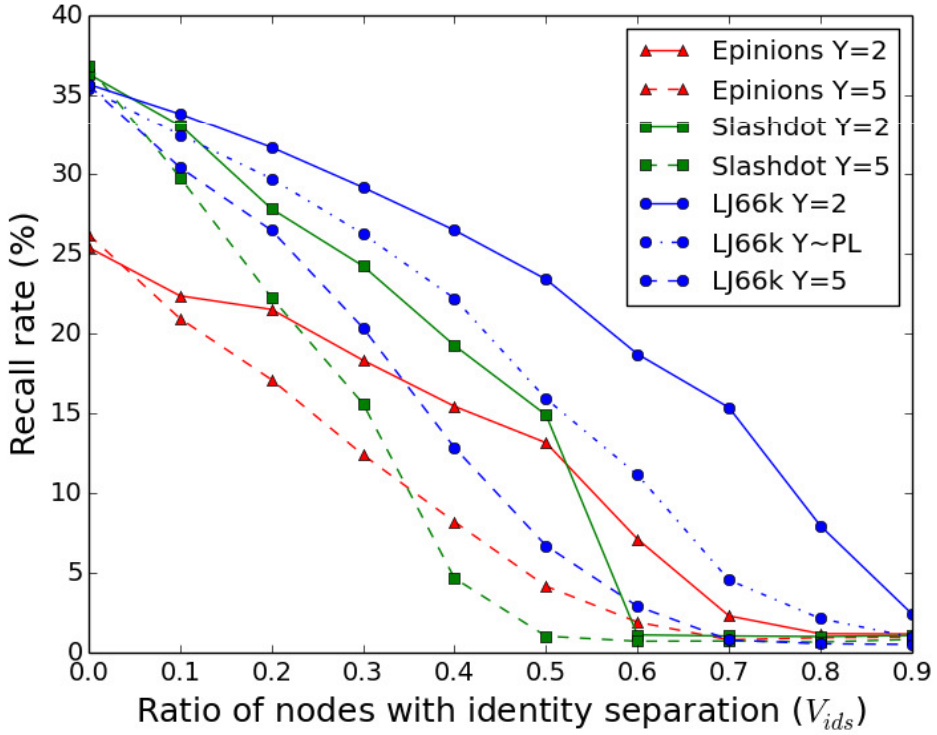
$$1+\sum_{\forall i\in[0,\dots,y]} p_i^{k-1}\cdot\left(\sum_{x''_1+\dots+x''_y=n-k+1}\left(\frac{(n-k+1)!}{x''_1!\cdot\ldots\cdot x''_y!}\cdot p_1^{x''_1}\cdot\ldots\cdot p_y^{x''_y}\cdot e(k-1,x''_i)\right)-1\right) \quad (3)$$

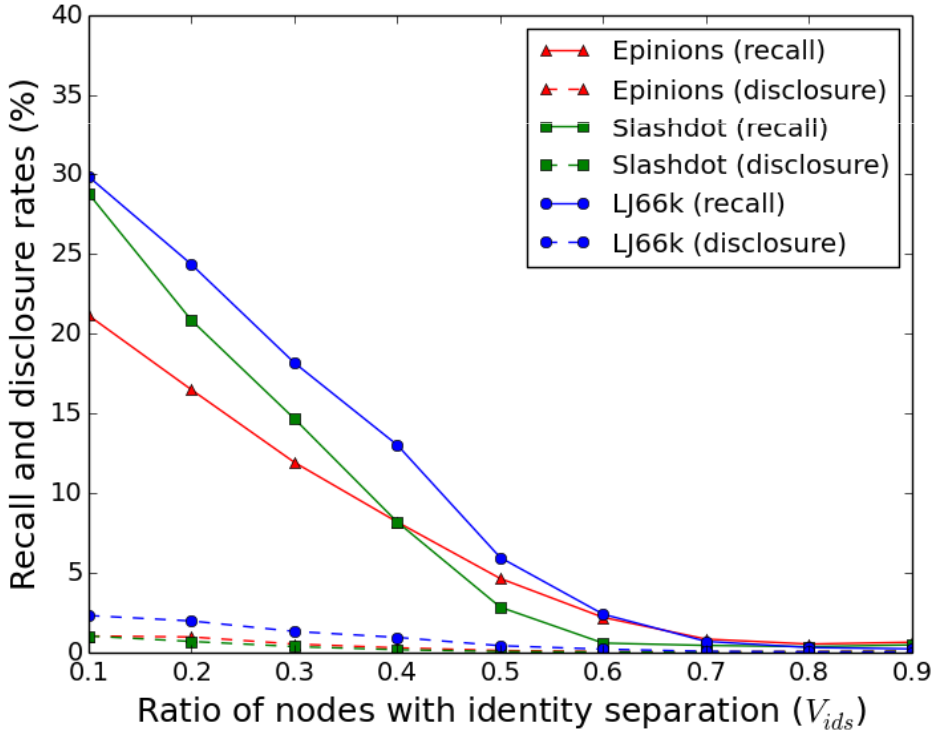


Feasible strategies!

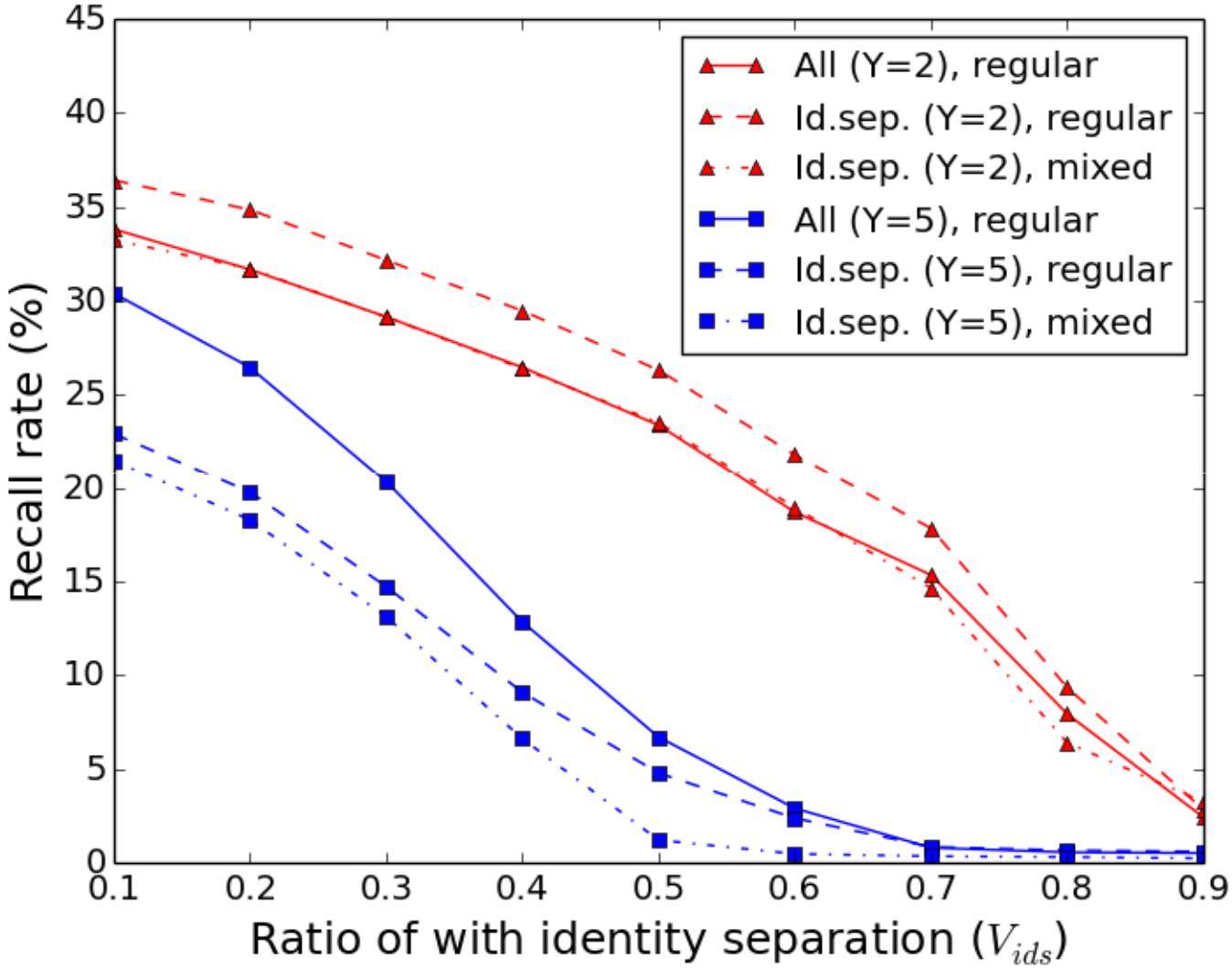
[T2.2-3] Can identity separation provide protection for the network?

- I evaluated several settings of identity separation with structural de-anonymization attacks, e.g.:
  - Only using many separated identities is not enough (fig. on the left)
  - Plus deleting some edges introduces individual privacy only (fig. on the right)



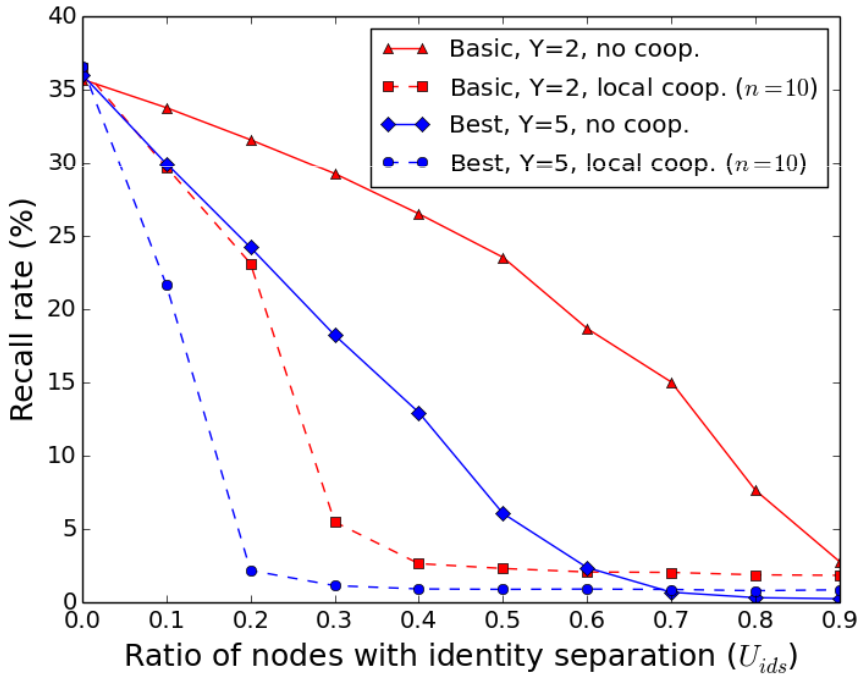


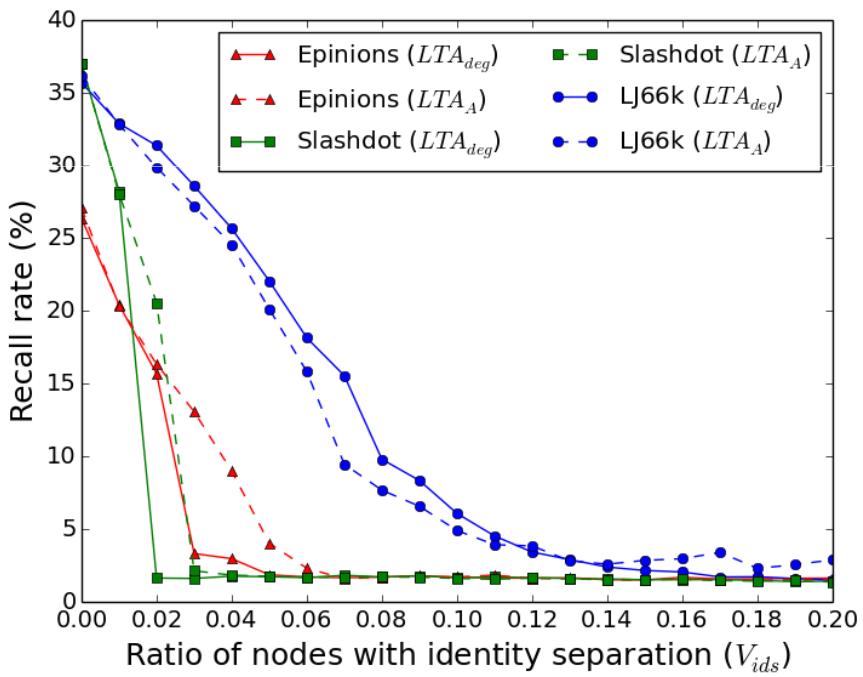
[T2.2-3] Can identity separation provide protection for the network? (2)



[T2.4-5] Can cooperation protect the network?

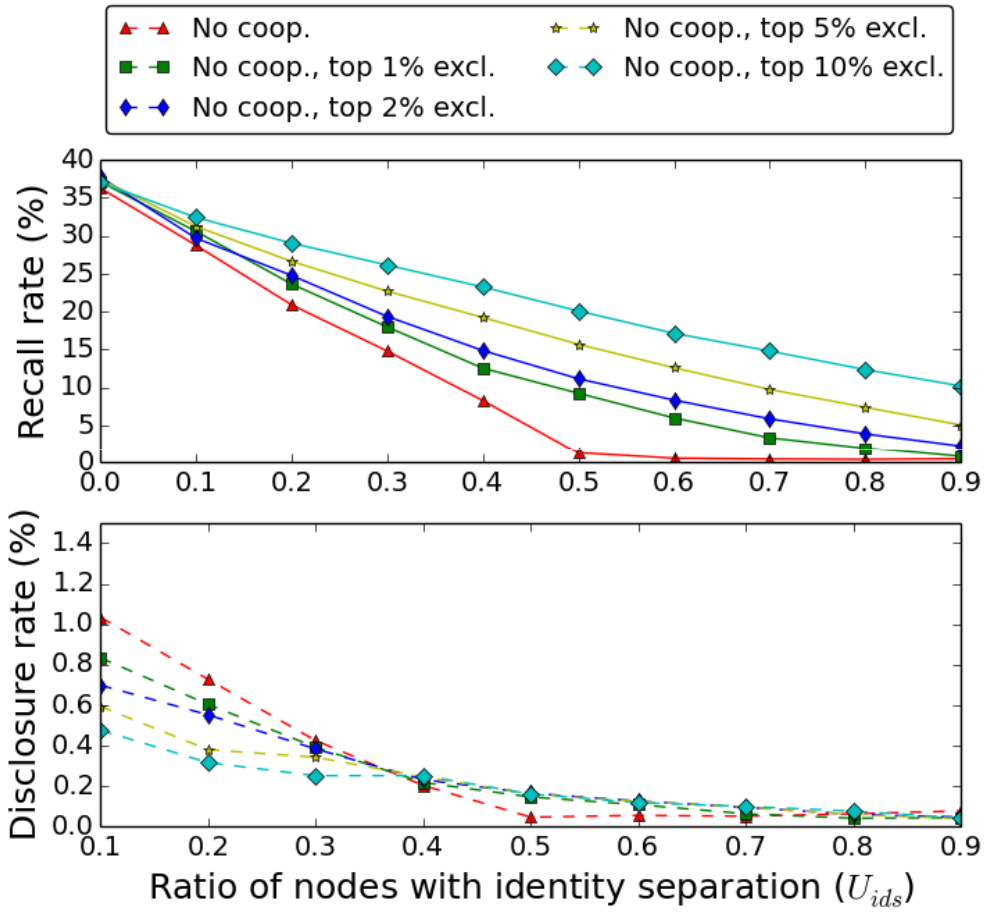
- Evaluation of different cooperation models:
  - Locally organized cooperation
  - Global cooperation: important nodes (low anonymity) should act





[T2.6] What happens if top degree nodes do not cooperate?

- Without their support, the performance of protection of network privacy degrades rapidly.



Related publications to Problem Set 2

[J1] G. G. Gulyás and S. Imre, "Hiding information against structural re-identification," *Telecommunication Systems*, September 2014. (under review).

[J2] B. Simon, G. G. Gulyás, and S. Imre, "Analysis of grasshopper, a novel social network de-anonymization algorithm," *Periodica Polytechnica Electrical Engineering and Computer Science*, January 2015. (accepted for publication).

[J3] G. G. Gulyás and S. Imre, "Using identity separation against de-anonymization of social networks," *Transactions on Data Privacy*, January 2015. (accepted for publication).

[J4] G. G. Gulyás and S. Imre, "Analysis of identity separation against a passive clique-based de-anonymization attack," *Infocommunications Journal*, vol. 4, pp. 11–20, December 2011.

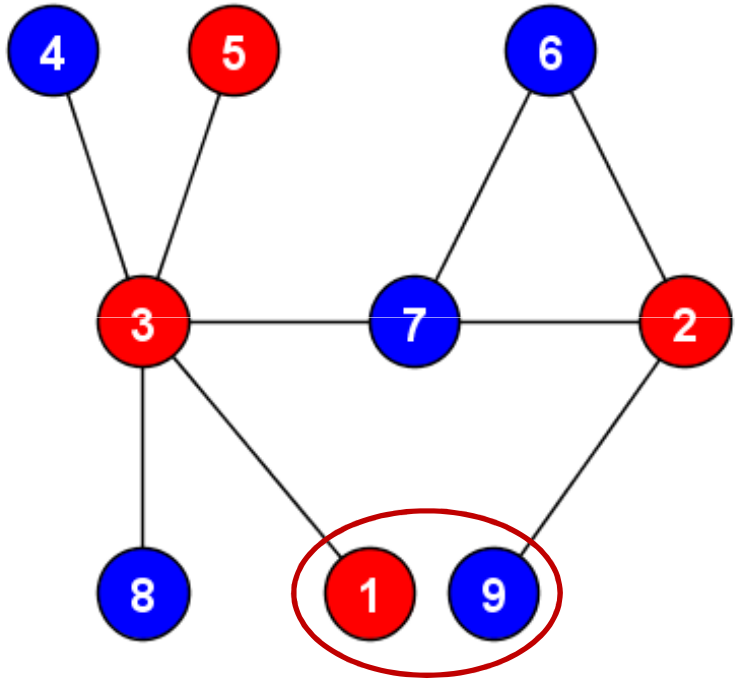
[C2] G. G. Gulyás and S. Imre, "Hiding information in social networks from de-anonymization attacks by using identity separation," in *Communications and Multimedia Security* (B. Decker, J. Dittmann, C. Kraetzer, and C. Vielhauer, eds.), vol. 8099 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2013.

[C7] G. G. Gulyás, R. Schulcz, and S. Imre, "Modeling role-based privacy in social networking services," in *Third International Conference on Emerging Security Information, Systems and Technologies (SECURWARE 2009*, pp. 173–178, June 2009.

[C8] G. G. Gulyás, "Design of an anonymous instant messaging service," in *Proceedings of PET Convention 2009.1* (S. Köpsell and K. Loesing, eds.), pp. 34–40, Fakultät Informatik, TU Dresden, March 2009.

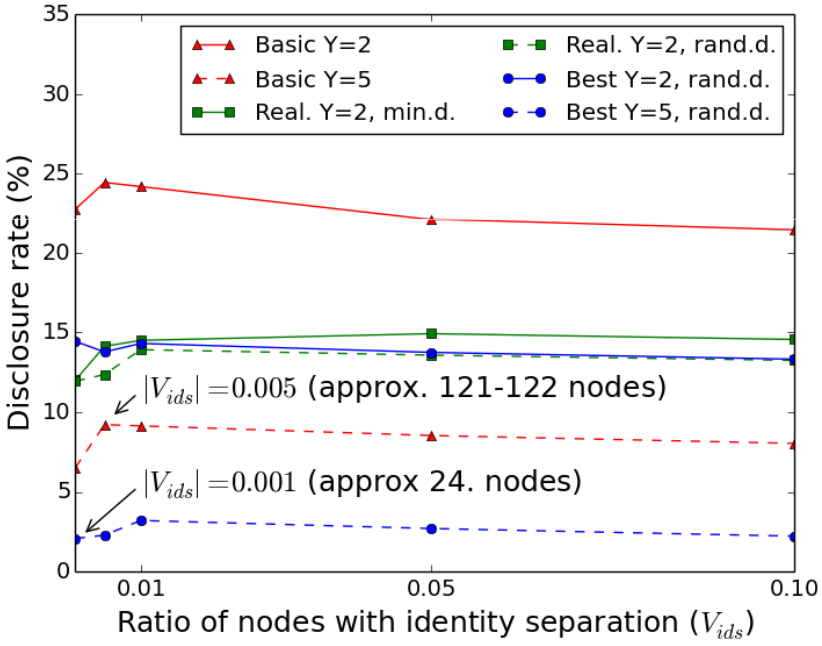
Problem Set 3: *individual strategies with identity separation*

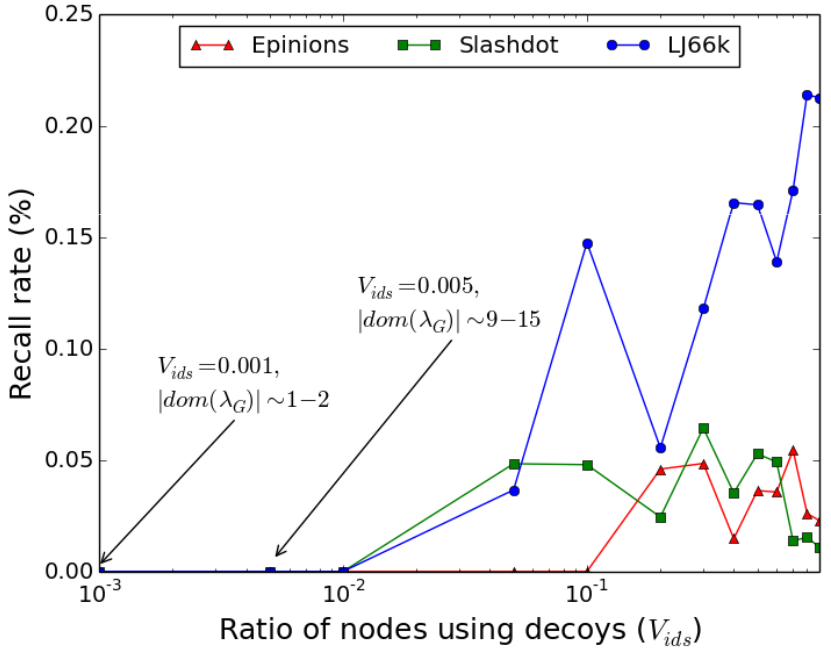
- Are there any advised individual strategies?
  - decoying: the naive approach
  - (reversibility of identity separation)
  - adopting k-anonymity to the context
  - y-identity: a new model



[T3.1] How effective a naive individual strategy can be?

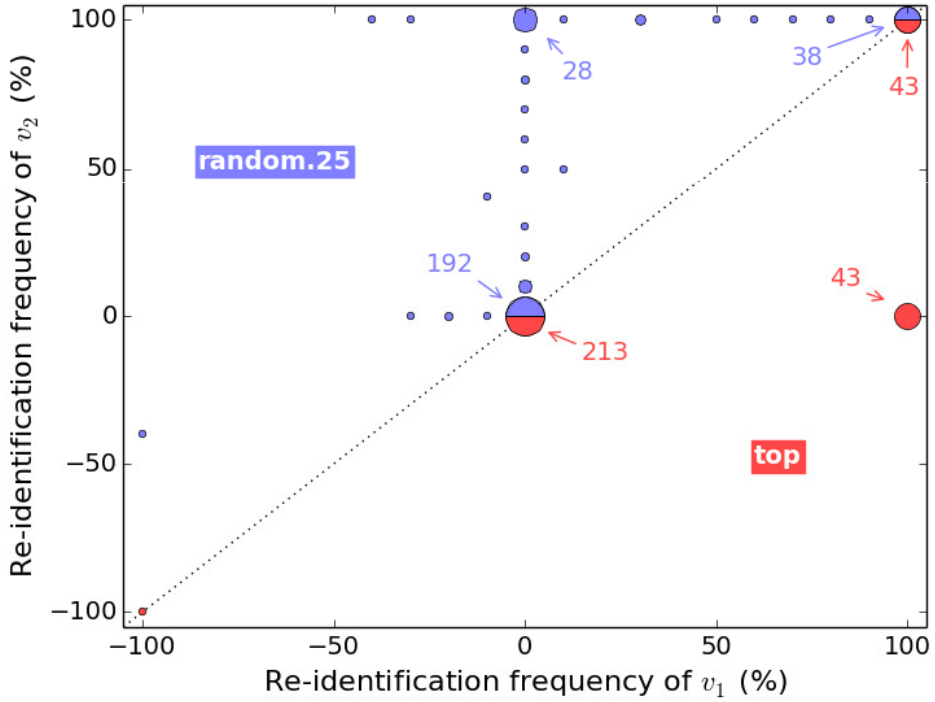
- Handful of users are still good.
- Targeted information hiding: using decoy identities.





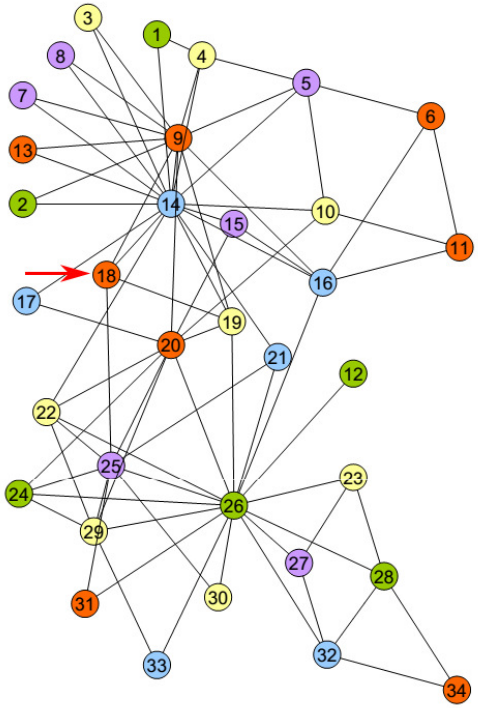
[T3.2] Can we reverse identity separation?

- I provided a method for calculating the lower bound of the probability of the discovery of partial identities with a simple modification of the state-of-the-art attack.

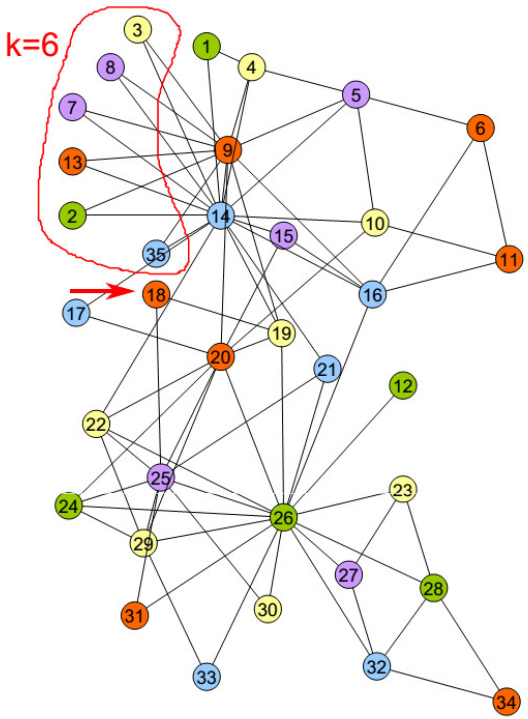


[T3.3] Can a k-anonymity variant enhance privacy?

Original network



k-anonymity example



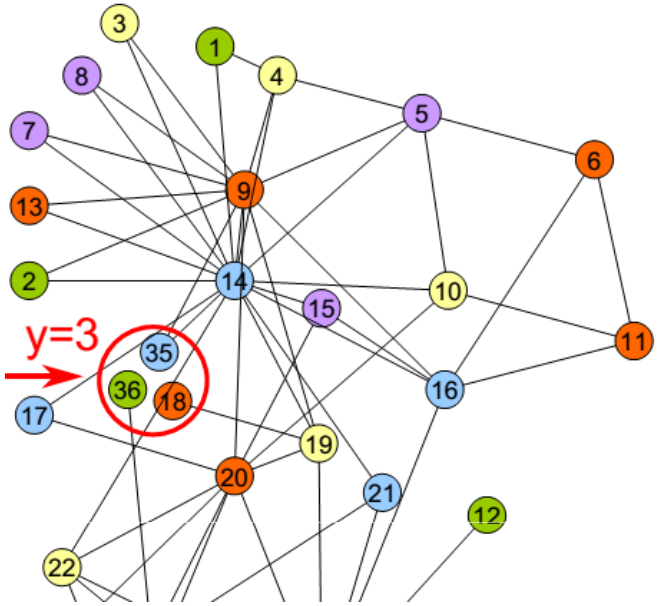
Algorithm 1: $(k, 2)$-anonymity with edge modification. It takes as input: the graph structure G, a node v_i selected for identity separation, c denoting the number of connections to anonymize, and parameter k of k-anonymity.

```
1: procedure K-AnonymizeNode(G, v_i, c, k)
2:     Calculate V_i, V_i^2
3:     c' ← c, V_k ← {}, E_k ← {}
4:     while c' ≥ 1 and |V_k| = 0 do
5:         κ ← {}                                  ▷ Groups having c' common neighbors with v_i
6:         for all v_j ∈ V_i^2 do
7:             V_i∩j ← V_i ∩ G.nbrs(v_j)
8:             if |V_j| = c and |V_i∩j| = c' then
9:                 κ[V_i∩j] ← κ[V_i∩j] ∪ {v_j}
10:            end if
11:        end for
12:        for all κ[V_i∩j] if |κ[V_i∩j]| ≥ k − 1 do
13:            if c = c' then                       ▷ k-anonymity without modification
14:                V_k ← κ[V_i∩j]
15:                break
16:            end if
17:            ψ ← {}                               ▷ Get new neighbors related to the k-group
18:            for all v_j ∈ κ[V_i∩j] do
19:                V_j\i ← G.nbrs(v_j) \ V_i \ κ[V_i∩j] \ {v_i}
20:                for all v_l ∈ V_j\i do
21:                    ψ[v_l] ← G.nbrs(v_l) ∩ κ[V_i∩j]
22:                end for
23:            end for
24:            η ← {}                               ▷ Filter applicable groups and neighbors
25:            for all ψ[v_l] do
26:                for all γ ⊆ ψ[v_l] if |γ| = k − 1 do
27:                    η[γ] ← η[γ] ∪ {v_l}
28:                end for
29:            end for
30:            if ∃η[γ] that |η[γ]| ≥ c − c' then
31:                pick η[γ] where |η[γ]| ≥ c − c'
32:                V_k ← γ
33:                E_k ← η[γ]
34:                break
35:            end if
36:        end for
37:        c' = c' − 1
38:    end while
39:    return V_k, E_k                              ▷ Existing and new neighbors for k-anonymity
40: end procedure
```

[T3.4] Is there a theoretically best approach?

- The y-identity model

- Definition as a game:
  - Player set $\mathcal{P}$, Strategy set $\mathcal{S}$, utility set $\mathcal{U}$
  - Strong attackers know they revealed $\mathcal{S}$
  - Weak attackers reveal only $\mathcal{S}' \subseteq \mathcal{S}$ (with uncertainty)



y-identity model

- For the given attacker model
  - Proven best strategies for each attacker type
  - There is proven a feasible strategy against an unknown attackers

Related publications to Problem Set 3

[J1] G. G. Gulyás and S. Imre, "Hiding information against structural re-identification," *Telecommunication Systems*, September 2014. (under review).

[J3] G. G. Gulyás and S. Imre, "Using identity separation against de-anonymization of social networks," *Transactions on Data Privacy*, January 2015. (accepted for publication).

[C2] G. G. Gulyás and S. Imre, "Hiding information in social networks from de-anonymization attacks by using identity separation," in *Communications and Multimedia Security* (B. Decker, J. Dittmann, C. Kraetzer, and C. Vielhauer, eds.), vol. 8099 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2013.

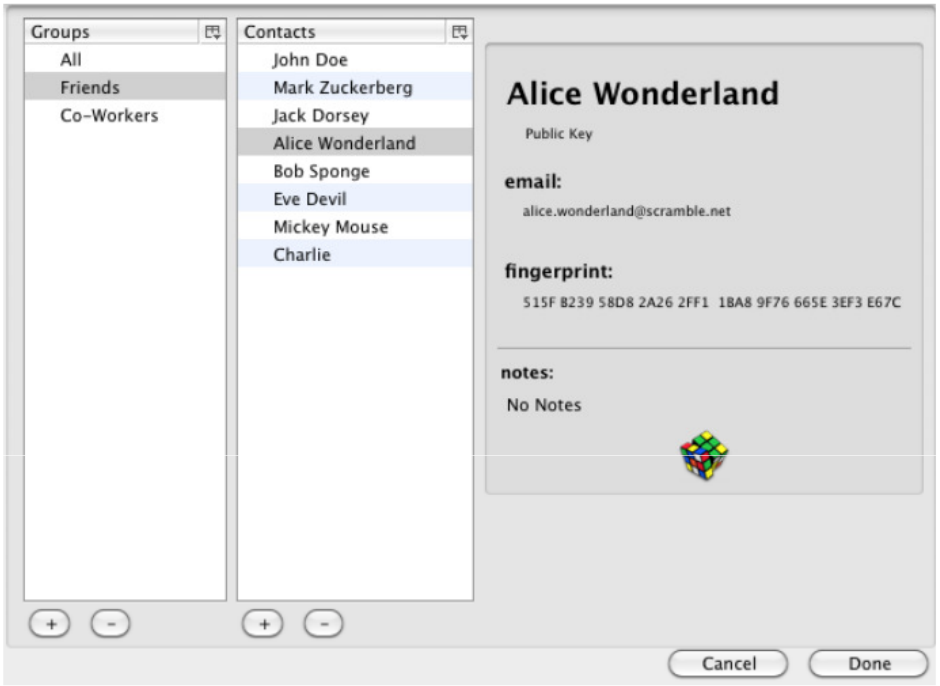
[C7] G. G. Gulyás, R. Schulcz, and S. Imre, "Modeling role-based privacy in social networking services," in *Third International Conference on Emerging Security Information, Systems and Technologies (SECURWARE 2009*, pp. 173–178, June 2009.

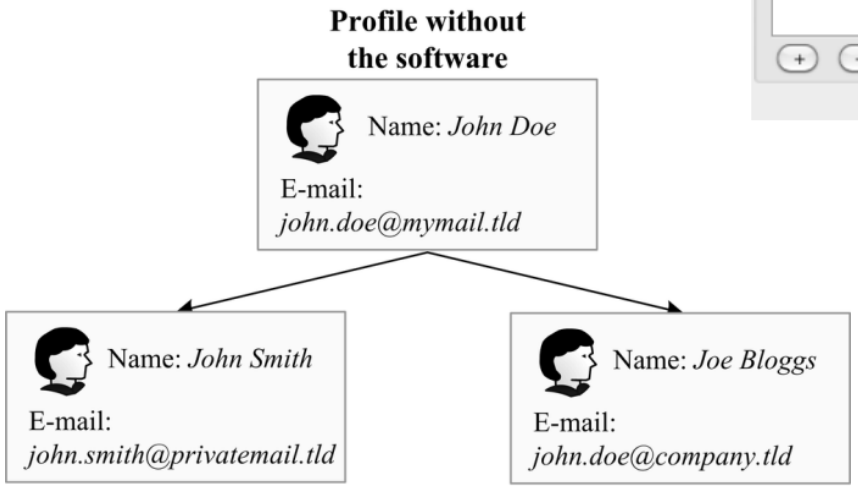
[C8] G. G. Gulyás, "Design of an anonymous instant messaging service," in *Proceedings of PET Convention 2009.1* (S. Köpsell and K. Loesing, eds.), pp. 34–40, Fakultät Informatik, TU Dresden, March 2009.

Applications of Results

- Conclusion
  - Global protection is hard
  - Individually it works
- Client side implementation?
  - Strategies similarly as proposed

F. Beato, et al. "Scramble! your social network data.", 2011.





Publications statistics

Type	Hungaran	National	Internat.	Count	Score
Journal					
Peer-reviewed		2	1	**3**	**12**
Non peer-reviewed	1			**1**	**½**
Book					
Book chapter	4	1	2	**7**	**21**
Conference papers			9	**9**	**18**
Patents					
Others	2			**2**	**5/6**
Sum	**7**	**4**	**11**	**22**	**52.3**

- 3 peer-reviewed international journal papers (all in Scopus)
  - +1 currently under review in a WoS journal (only minor revision)
- 42+4 independent citations

Acknowledgement goes to...

- Sándor Imre
- Levente Buttyán
- Iván Székely
- Colleagues for inspiring discussions: Tamás Holczer, Márk Félegyházi, Máté Horváth, András Telcs
- Co-authors
- Members of the CrySyS & MCL Labs
- Reviewers of my dissertation: Julien Freudiger and Gergely Biczók
- My Wife, and ...

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

REVIEWER QUESTIONS

Julien Freudiger #1

- *What is the best way to bring the multiple identity management to market? Should we assist users into managing multiple identities, or should social networks adopt such practices?*

- Possible targets: ‘natural’ users of identity separation and privacy conscious folks
- Client/server side: depends on the service
  - Client side: web identity management, MAC change (+social netw.)
  - Server side: social networks
- Some apps/features already show that the idea could work, e.g., hmmmapp.com or circles in Google+

Gergely Biczók #1

- *The candidate has tapped into an intriguing line of research with his work on simple identity partitioning games in Chapter 6. Which family of (naturally more complex) game-theoretical models could be utilized to better capture the connected nature of the social networks under attack?*

- Possible extensions are mentioned in my dissertation, e.g., another type of weak attacker who can assess the possible unknown strategy set of the user
- Connections could be according to the model in [1] to evaluate how the adoption of the identity separation. This model focuses on the cost-benefit ratio.

[1] Ohtsuki, Hisashi, et al. "A simple rule for the evolution of cooperation on graphs and social networks." Nature 441.7092 (2006): 502-505.