Safely publishing your social network data:

The network anonymization problem

Rachel de Jong, Mark van der Loo, **Frank Takes** — Leiden University

28 June 2025, Paris, France

PLANET-NL
Netherlands (NL) Platform for Large-scale Analysis of NETworks

Universiteit van Amsterdam

Universiteit Leiden

NWO

SSHOC
social sciences & humanities open cloud

PDI SSH Platform Digitale Infrastructuur Social Sciences & Humanities

Recommendations for Sharing Network Data and Materials

Neal, Z. P., Almquist, Z. W., Bagrow, J., Clauset, A., Diesner, J., Lazega, E., Lovato, J., Moody, J., Peixoto, T. P., Steinert-Threlkeld, Z., and Teixeira, A. S. (2024). Recommendations for sharing network data and materials. *Network Science*, *12*, 404-417. https://doi.org/10.1017/nws.2024.16

"Researchers should share the network data and materials necessary to reproduce reported results via a publicly accessible repository when an associated manuscript is published...but access to or use of shared data and materials may be restricted if necessary to avoid harm or comply with regulations" (p. 404).

Read the recommendations

Endorse the recommendations

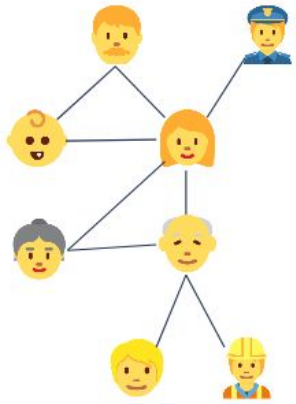
ENDORSED BY

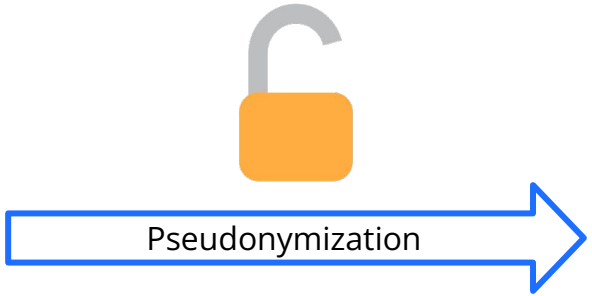
INSNA

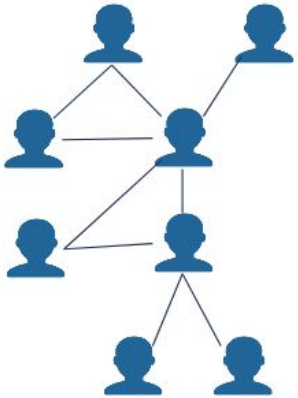
ICPSR

Duke | DUKE NETWORK ANALYSIS CENTER

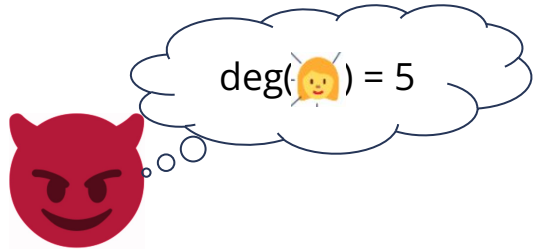
Publishing networks and privacy

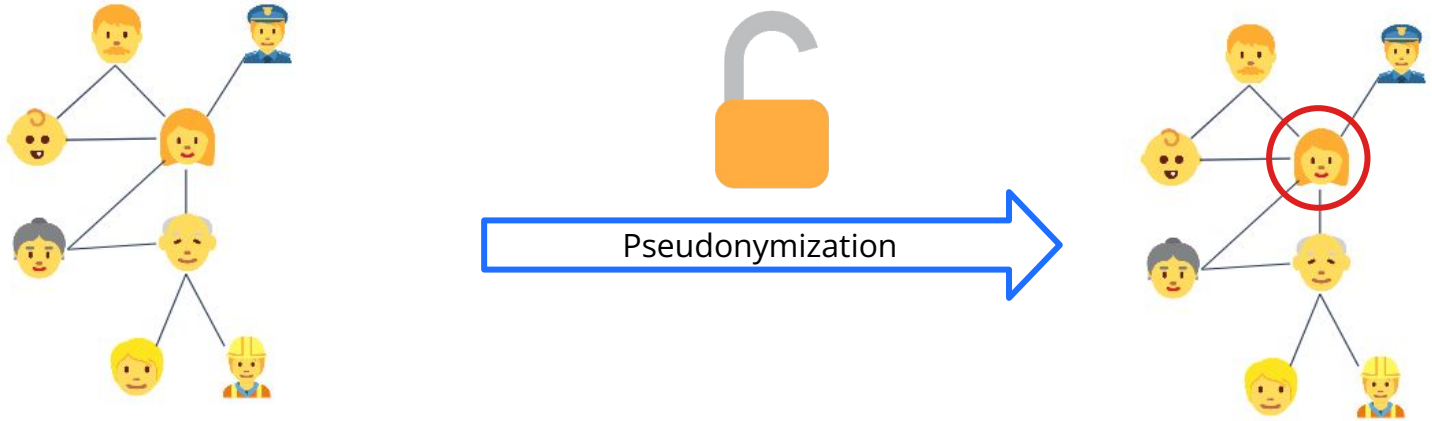






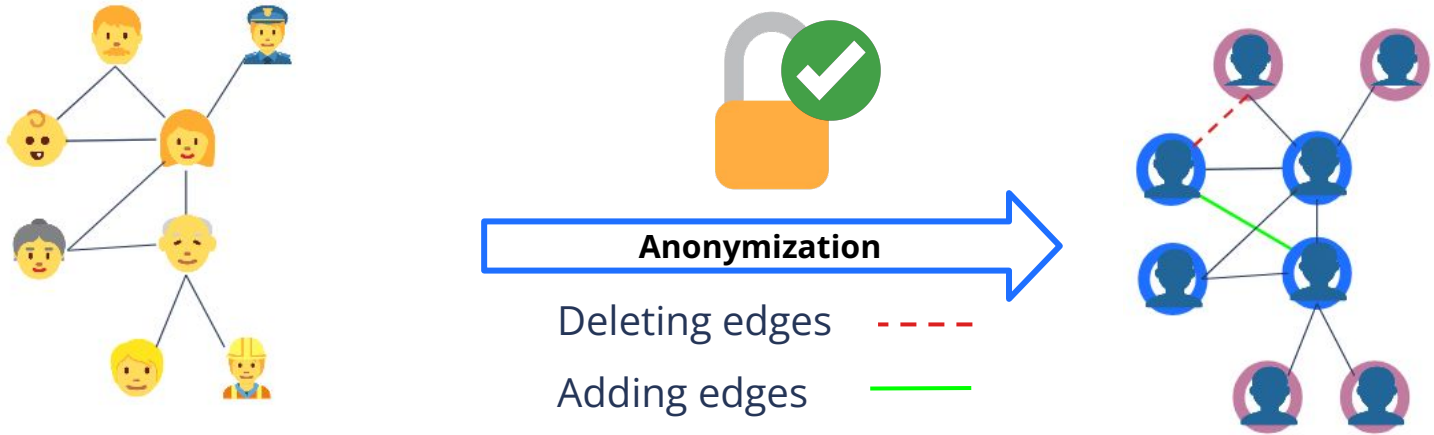
Publishing networks and privacy





Attacker scenario: information in the hands of an adversary

Publishing networks and privacy



This talk

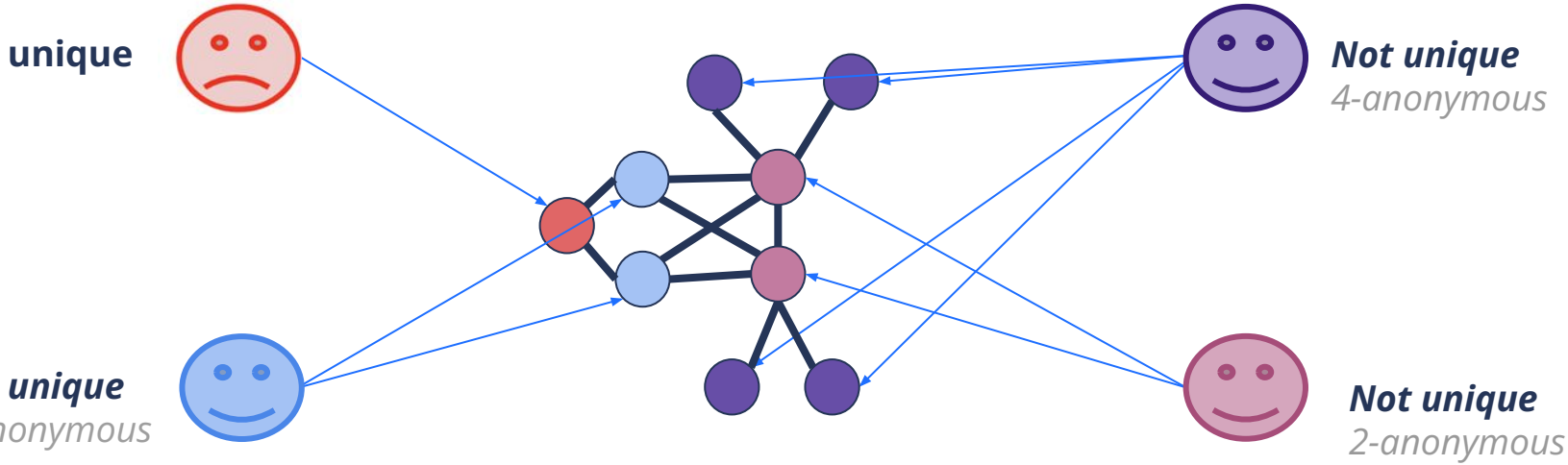
Q1: How to measure anonymity?

Q2: How to anonymize a network?

Q3: What is the utility of the resulting network?

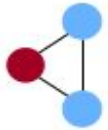
k-Anonymity

- Given a graph *G = (V, E)*,
  - a node is *k*-anonymous if it has *k-1* equivalent nodes
  - a node is **not unique** if it is at least *2*-anonymous
  - For this presentation, assume: not unique → anonymous
- **Network uniqueness** *U(G)* = fraction of unique nodes (1/9 in the graph below)



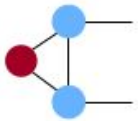
Anonymity measures

- Whether two nodes are equivalent is assessed by an anonymity measure
- Anonymity measures of the **ego network** (distance 1) of the focal node:
  - Degree 2
  - Count (|V|, |E|) (3, 3)
  - Degree distribution {2, 2, 2}
  - Isomorphism 3a4b



runtime

- Anonymity measures looking **slightly further** than focal node ego network:
  - VRQ: degree of neighbors {2, 3, 3}

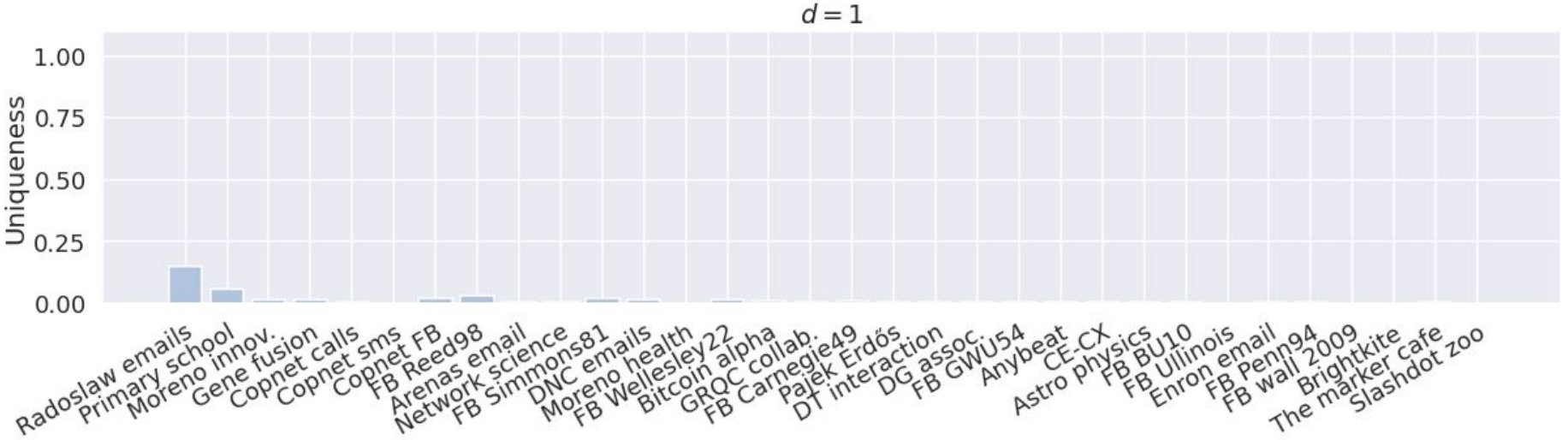


- Anonymity measures looking **further**:
  - All the five measures above at **distance 2**
  - Isomorphism at distance > 2 ca4b5e

R.G. de Jong, M.P.J. van der Loo and F.W. Takes, The effect of distant connections on node anonymity in complex networks, *Scientific Reports* 14: 1156, 2024.

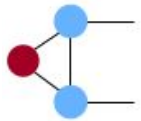
Anonymity measures

Real-world datasets with ~100 up to 3M nodes, and ~100 up to 18M edges



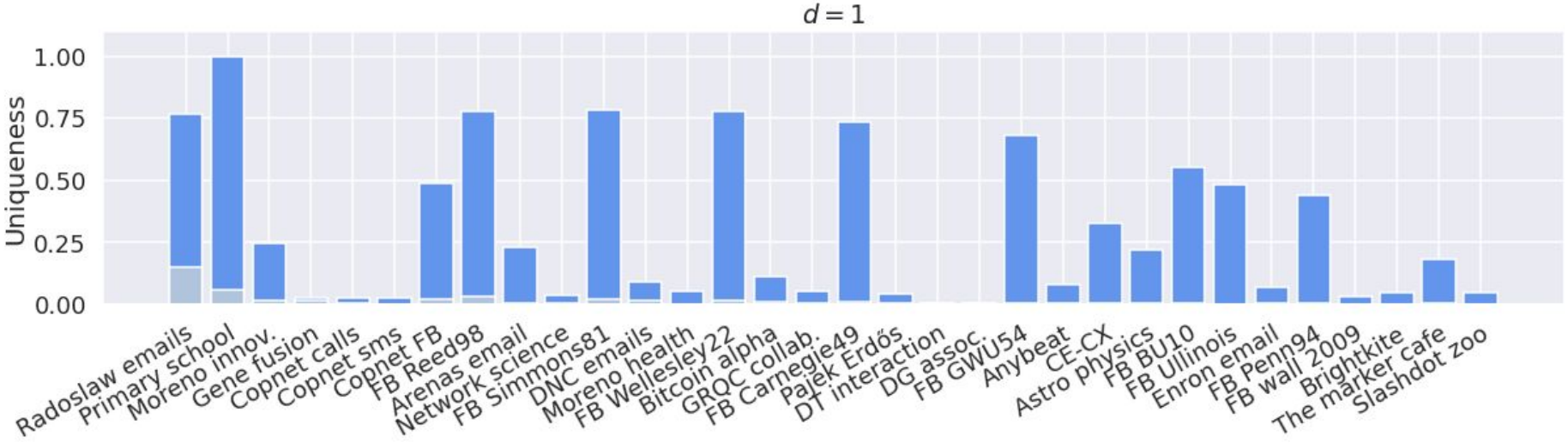
Anonymity measures

Degree 2



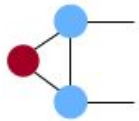
Anonymity measures

Real-world datasets with ~100 up to 3M nodes, and ~100 up to 18M edges



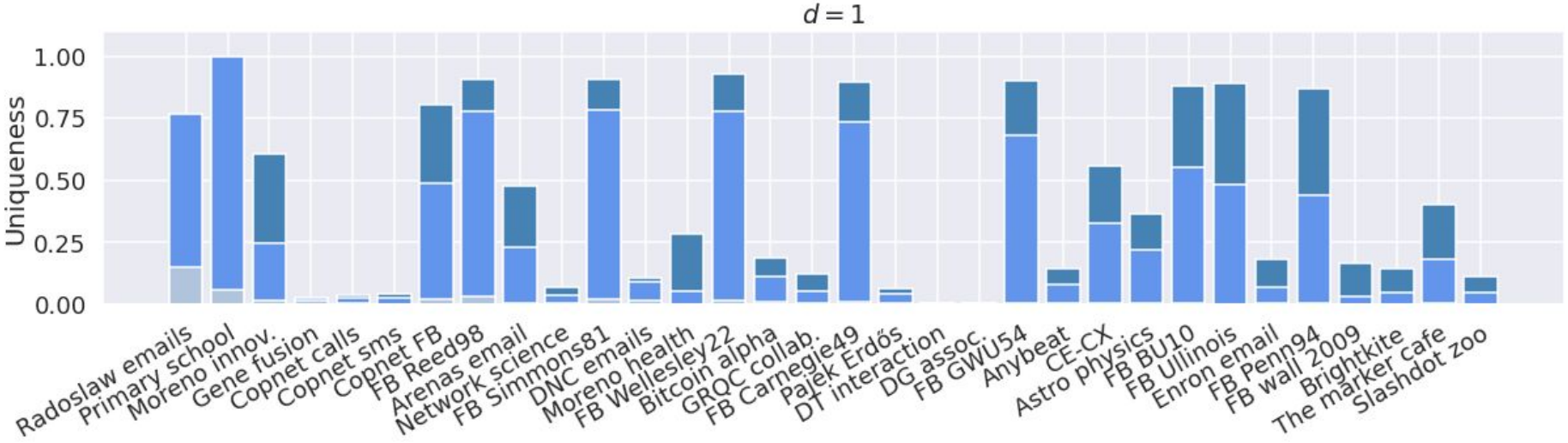
Anonymity measures

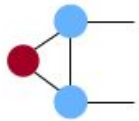
Degree	2
Count (|V|, |E|)	(3, 3)



Anonymity measures

Real-world datasets with ~100 up to 3M nodes, and ~100 up to 18M edges



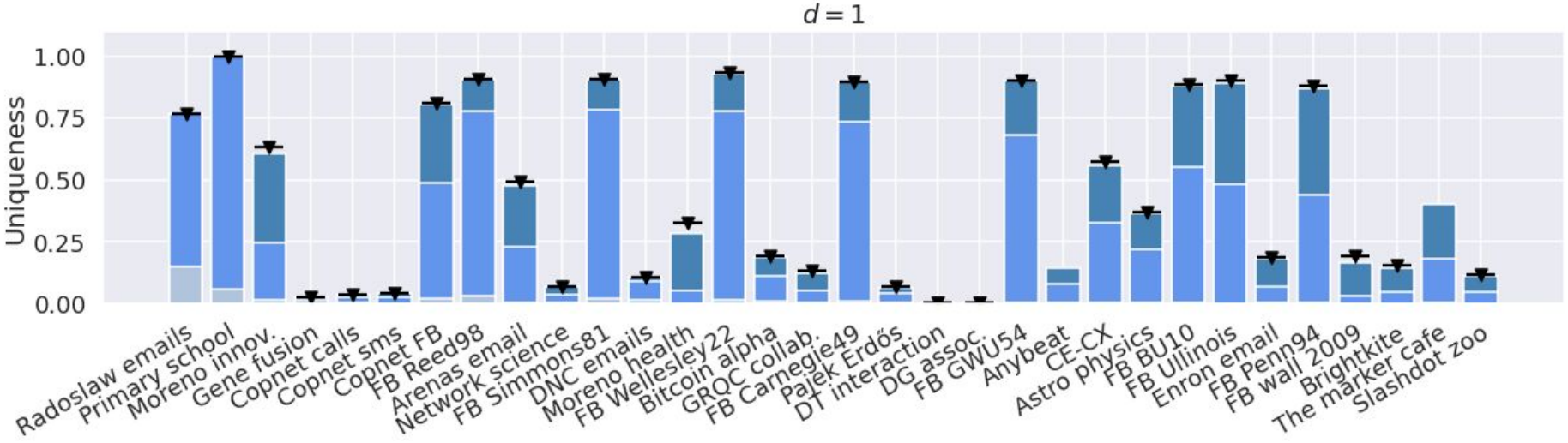


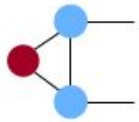
Anonymity measures

Degree	2
Count (|V|, |E|)	(3, 3)
Degree distribution	{2, 2, 2}

Anonymity measures

Real-world datasets with ~100 up to 3M nodes, and ~100 up to 18M edges



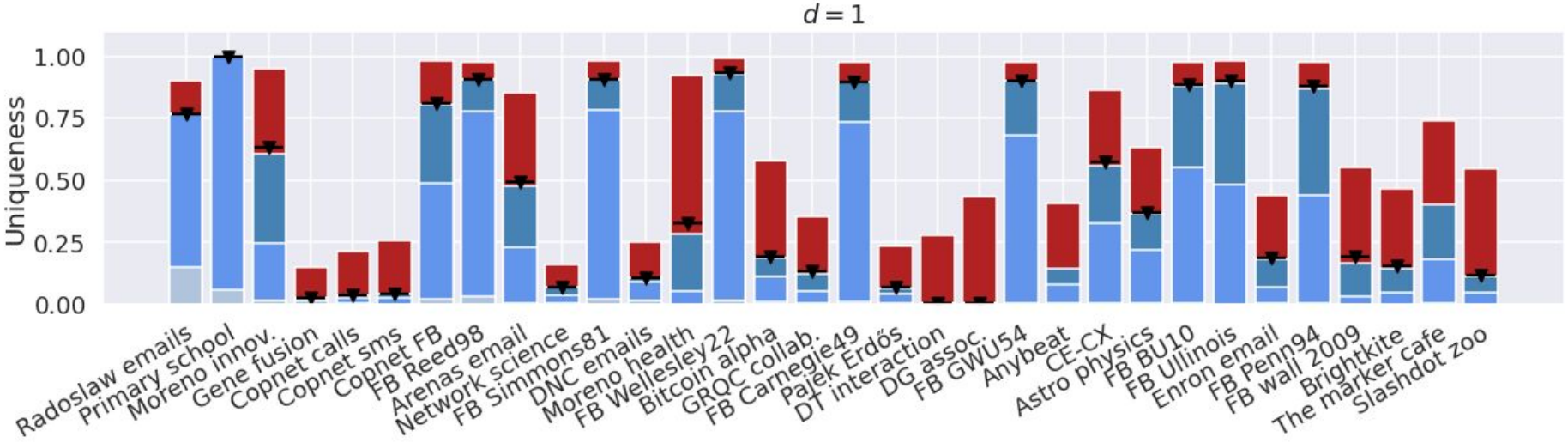


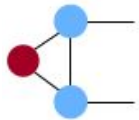
Anonymity measures

Degree	2
Count (|V|, |E|)	(3, 3)
Degree distribution	{2, 2, 2}
Isomorphism	3a4b

Anonymity measures

Real-world datasets with ~100 up to 3M nodes, and ~100 up to 18M edges



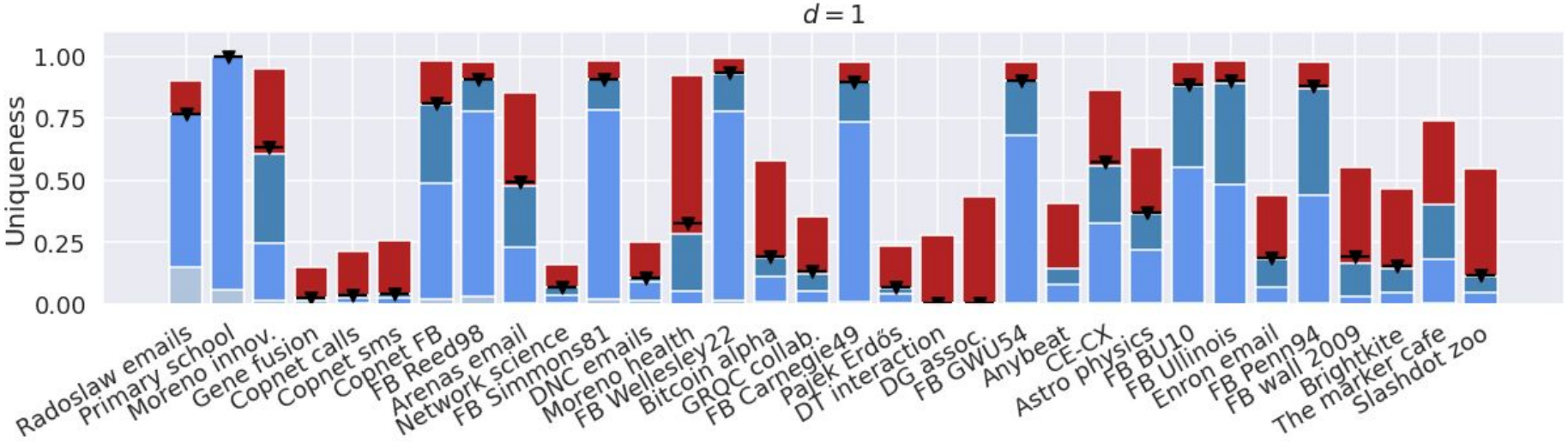


Anonymity measures

Degree	2
Count (|V|, |E|)	(3, 3)
Degree distribution	{2, 2, 2}
Isomorphism	3a4b
VRQ: degree of neighbors	{2, 3, 3}

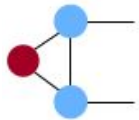
Anonymity measures

Real-world datasets with ~100 up to 3M nodes, and ~100 up to 18M edges



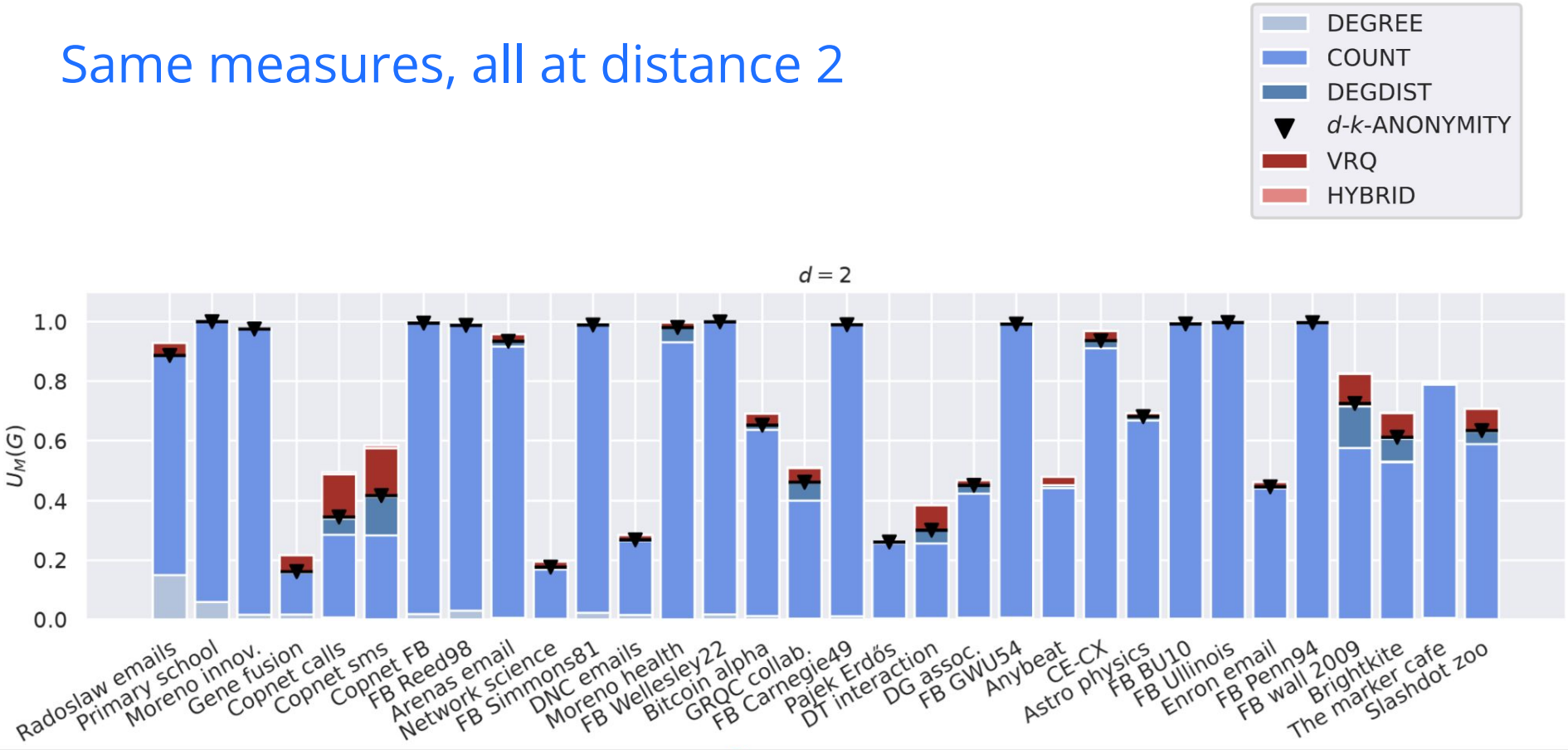
Using imprecise information, yet looking further, is more effective!

R.G. de Jong, M.P.J. van der Loo and F.W. Takes, The effect of distant connections on node anonymity in complex networks, *Scientific Reports* 14: 1156, 2024.



Anonymity measures	
Degree	2
Count (|V|, |E|)	(3, 3)
Degree distribution	{2, 2, 2}
Isomorphism	3a4b
VRQ: degree of neighbors	{2, 3, 3}

Same measures, all at distance 2



This talk

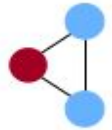
Q1: How to measure anonymity?

Q2: How to anonymize a network?

Q3: What is the utility of the resulting network?

Network anonymization

- In the remainder on the anonymity measure **count** (|V|, |E|) (3, 3)

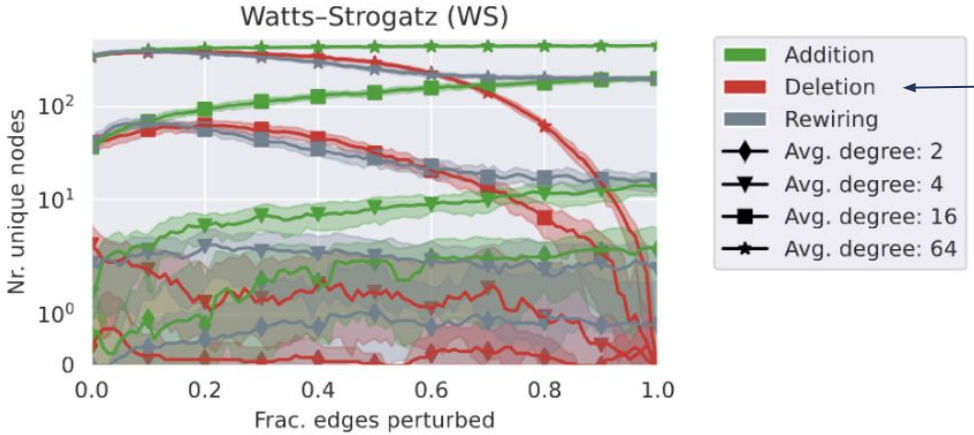


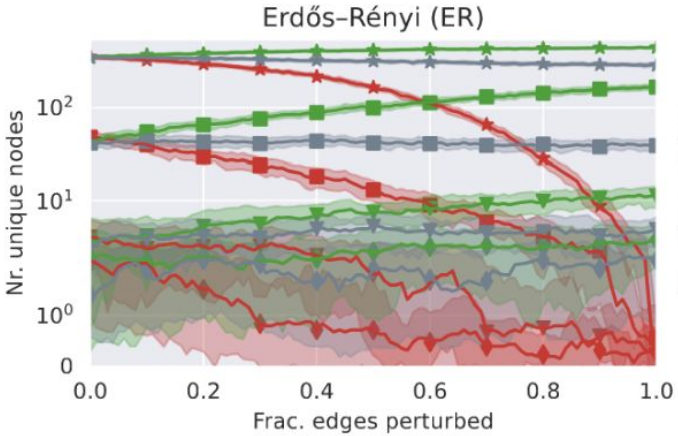
- **Network anonymization problem**: given a network *G = (V, E)*, perturb the network to maximize anonymity *1 - U(G)*

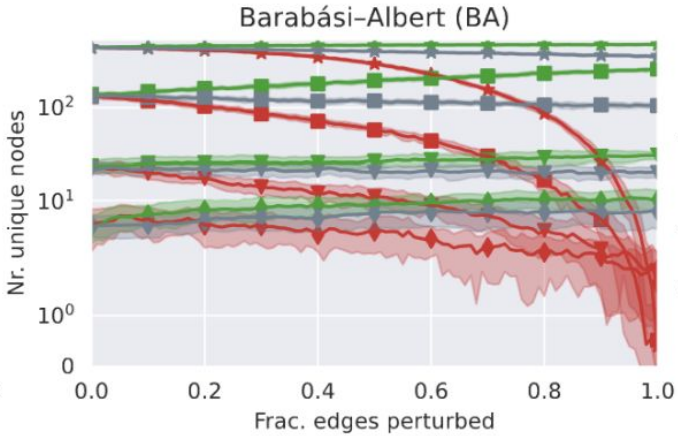
  Alternatively: the uniqueness minimization problem, i.e., minimize *U(G)*

- How to perturb the network?

How to perturb the network?

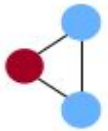






Network anonymization

- In the remainder on the anonymity measure **count** (|V|, |E|) (3, 3)



- **Network anonymization problem**: given a network *G = (V, E)*, perturb the network to maximize anonymity *1 - U(G)*

  Alternatively: the uniqueness minimization problem, i.e., minimize *U(G)*

- Perturbation method: **edge deletion**
  - More sparse graph means fewer possible “states”, i.e., more anonymity
  - Mimics real-world attacker scenario of missing or unobserved links

Anonymization algorithms

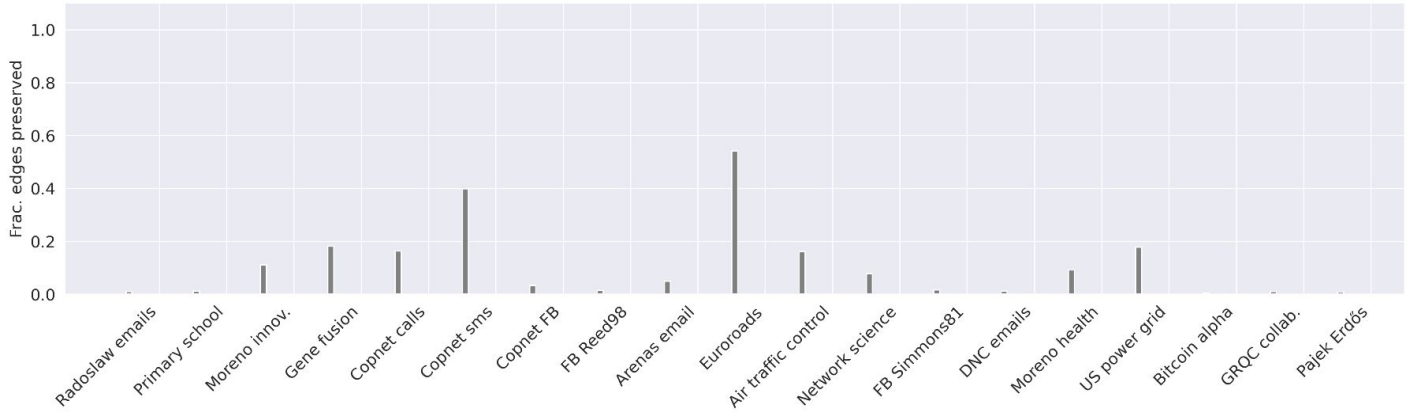
- Random and local optimization/heuristic approaches (this presentation):
  - Random deletion (edge sampling)
  - Structure-based edge selection (e.g., based on degree)
  - Uniqueness-based edge-selection: select edges that **affect the largest number of unique nodes**, in an attempt to make them anonymous

- Global optimization (not in this presentation):
  - Genetic algorithms, see:
    S. Bonello, R.G. de Jong, T.H.W. Baeck, and F.W. Takes, Utility-aware social network anonymization using genetic algorithms, in *Proceedings of the 27th ACM Genetic and Evolutionary Computation Conference (GECCO)*, 2025.
  - Many other ideas that we are working on :-)

- Algorithm performance measured in terms of **fraction of edges preserved**

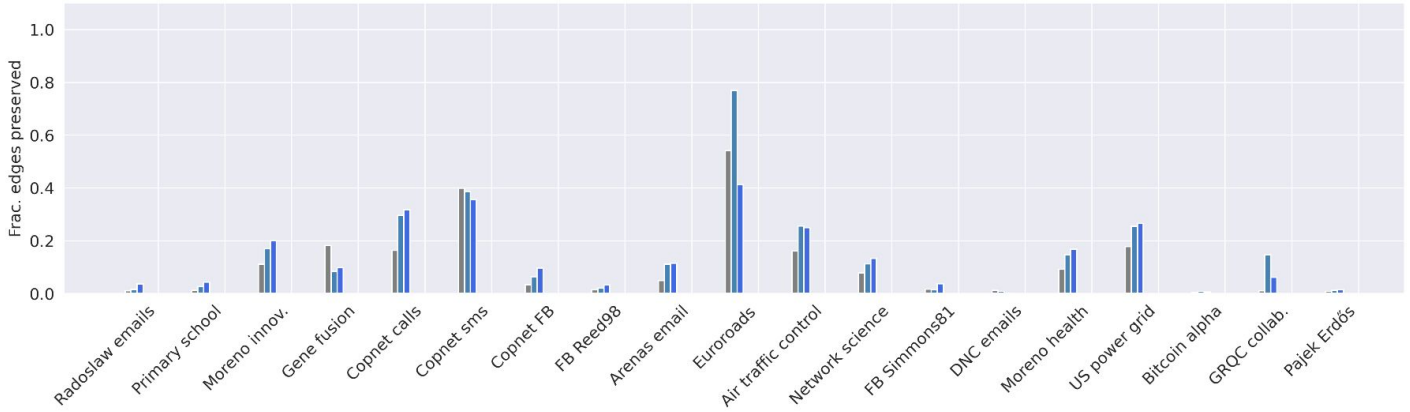
Anonymization algorithms

Random



$|V|$: ~100 - 6,927 $|E|$: ~300 - 11,850

Anonymization

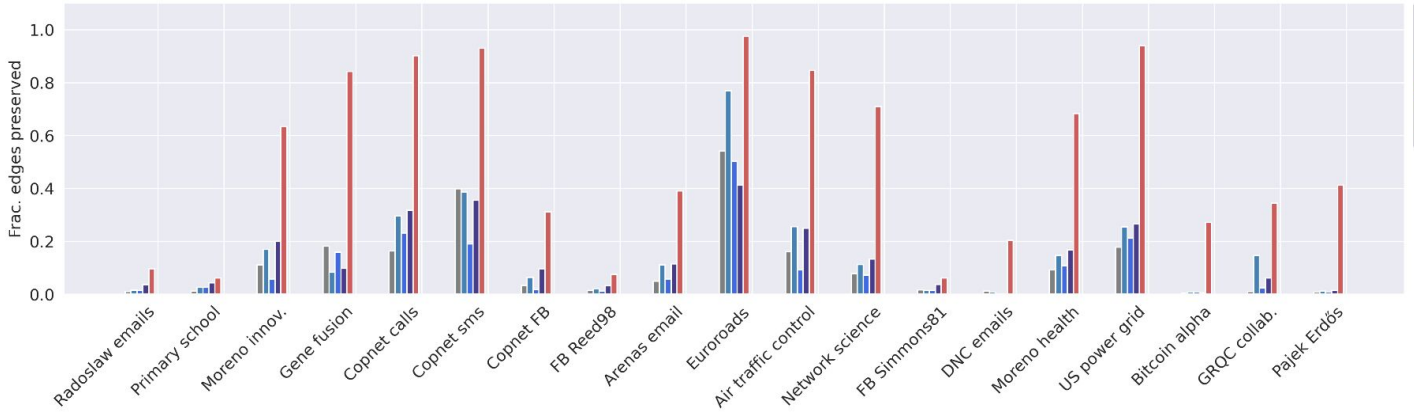


|V|: ~100 - 6,927 |E|: ~300 - 11,850

Random

Structure based

Anonymization



|V|: ~100 - 6,927 |E|: ~300 - 11,850

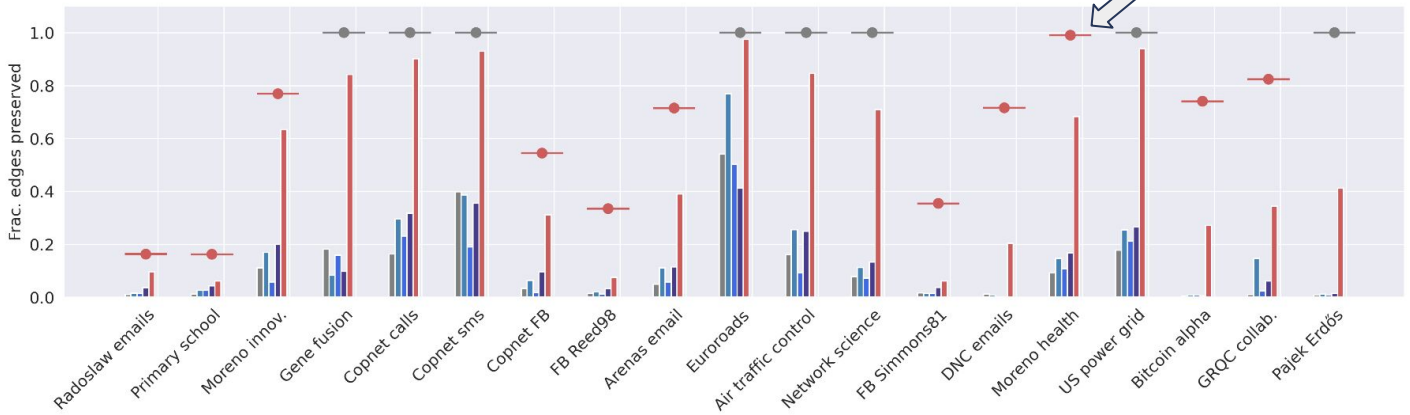
Random

Structure based

Uniqueness based

Anonymization

95% anonymous



Random

Structure based

Uniqueness based

|V|: ~100 - 6,927 |E|: ~300 - 11,850

Partial and budgeted anonymization are variants of the anonymization problem

R.G. de Jong, M.P.J. van der Loo and F.W. Takes, The anonymization problem in social networks, arXiv 2409.16163, 2024.

This talk

Q1: How to measure anonymity?

Q2: How to anonymize a network?

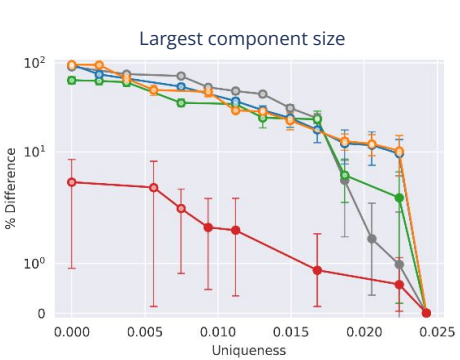
Q3: What is the utility of the resulting network?

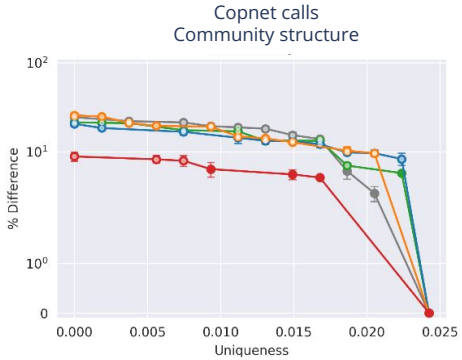
Data utility

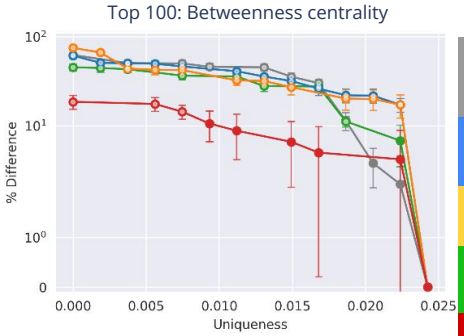
- Data utility in network anonymization: extent to which perturbed network remains usable for common network analysis tasks
  - **Edges retained** — Straightforward measurement of utility
Literature: 5% missing edges is usually fine!
  - **Clustering coefficient** — Percentage difference in average node clustering coefficient
  - **Average distance** — Percentage difference in average distance in giant component
  - **Robustness** — Fraction of nodes in the giant component
  - **Centrality** — Percentage overlap of top-100 nodes based on betweenness
  - **Community structure** — Similarity in terms of NMI of discovered communities

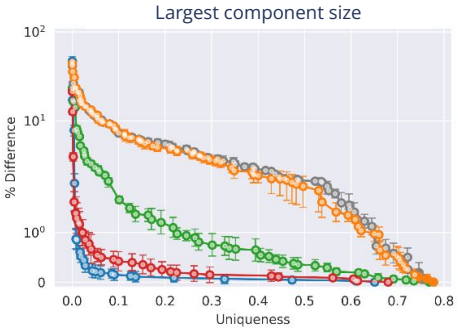
- Privacy/utility trade-off: balance between attained anonymity and data utility

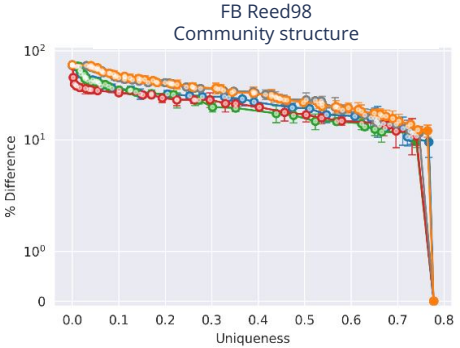
Anonymity vs. utility: **algorithms**

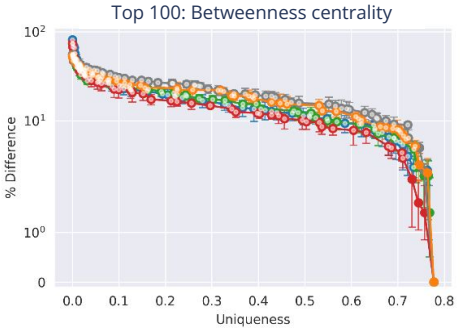












Random

Structure based

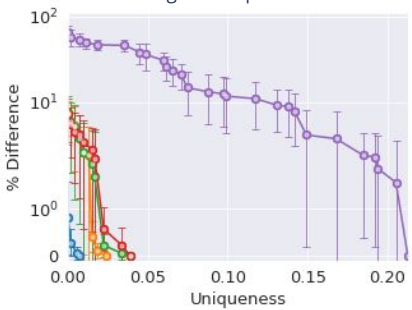
Structure based

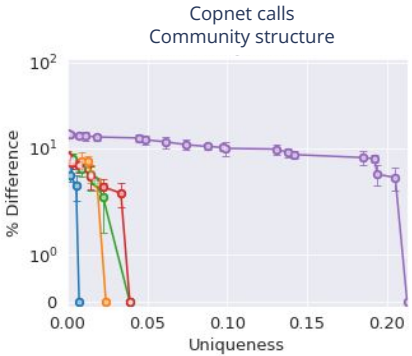
Uniqueness based

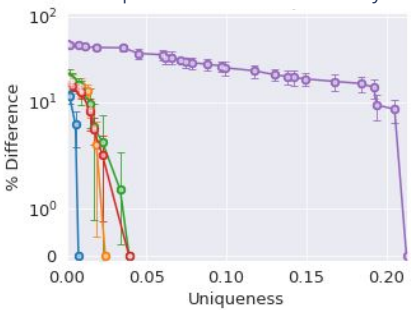
Uniqueness based

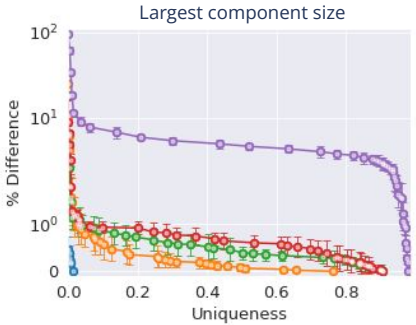
Best algorithm based on unique-ness does not affect trade-off!

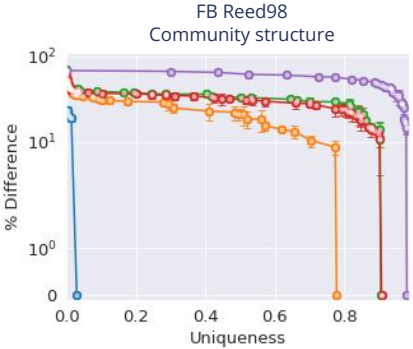
Anonymity vs. utility: **measures**

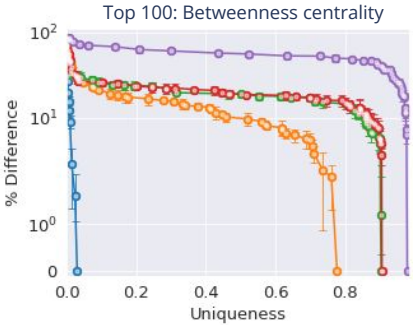












Degrees of neighbors

Isomorphism

Degree distribution

(|V|, |E|)

Degree

More strict measure

→

worse trade-off

Conclusions

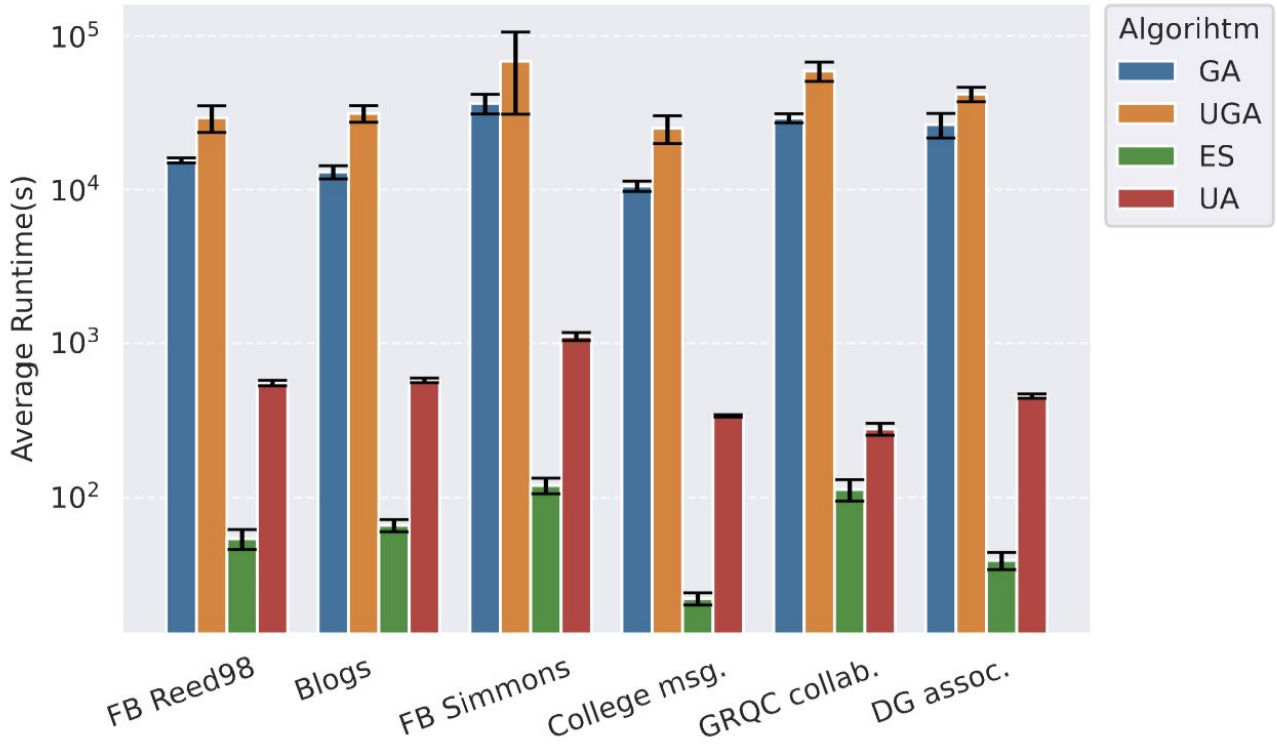
- Network anonymization is an interesting, important problem
- To measure anonymity, look **further**, rather than precisely
- To anonymize a network heuristically, non-random targeting of edges between unique nodes results in
  - higher anonymity than random targeting and
  - equal or higher data utility
- Future work includes utility-aware optimization, applications to population-scale network data, and better algorithms

ANONET package: implement your anonymization algorithms!
github.com/RacheldeJong/ANONET

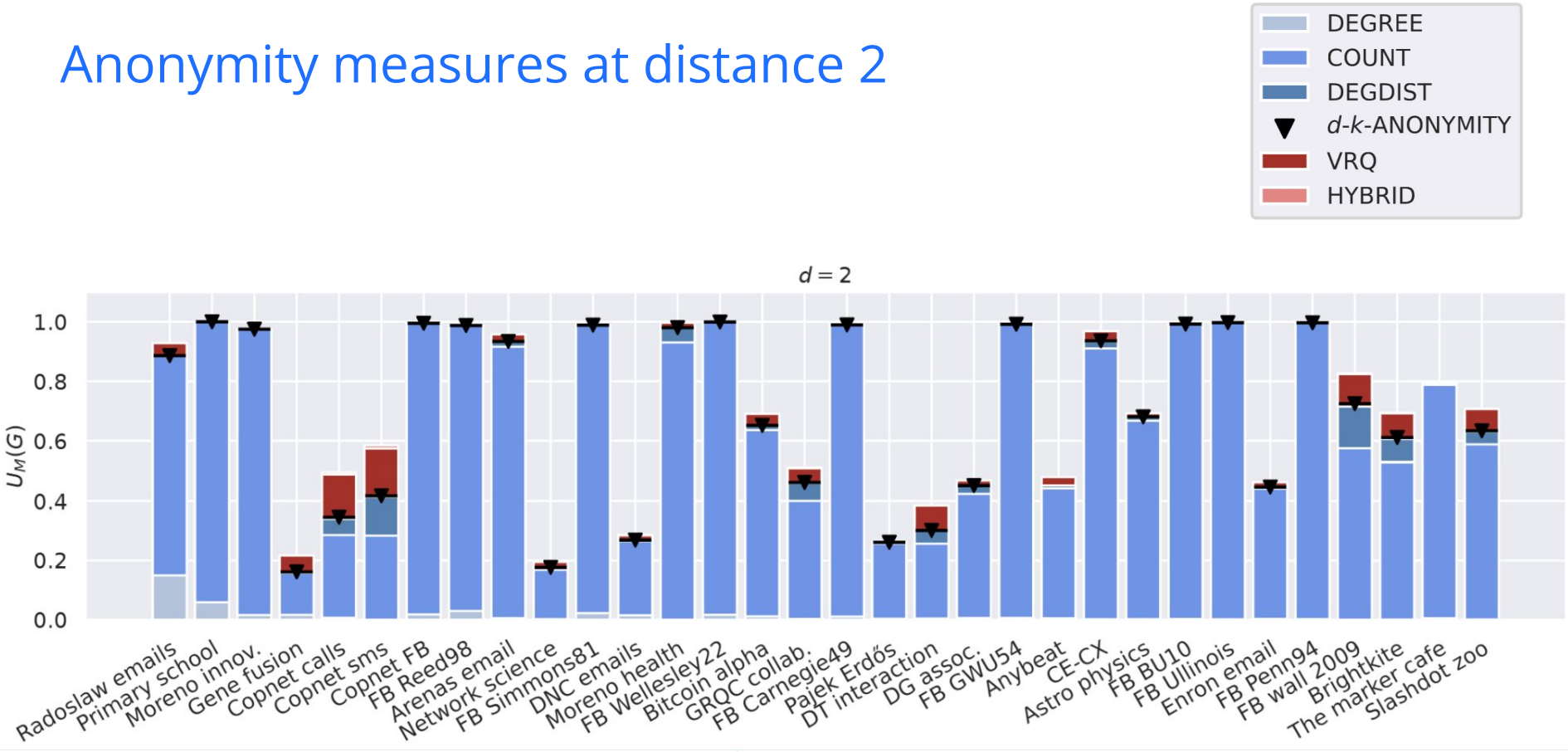
franktakes.nl computationalnetworkscience.org

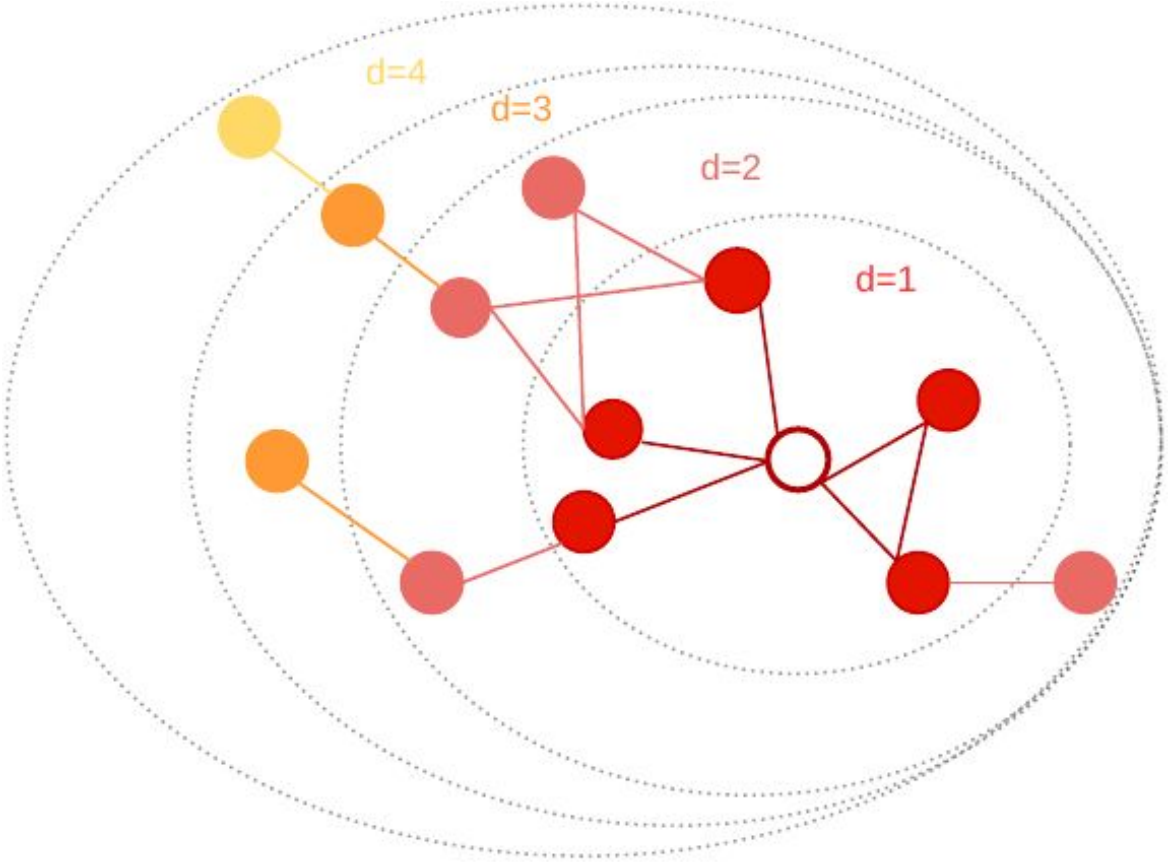
R.G. de Jong, M.P.J. van der Loo and F.W. Takes, A systematic comparison of measures for k-anonymity in networks, arXiv 2407.02290, 2025.
S. Bonello, R.G. de Jong, T.H.W. Baeck, and F.W. Takes, Utility-aware social network anonymization using genetic algorithms, in *Proceedings of the 27th ACM Genetic and Evolutionary Computation Conference (GECCO)*, 2025.
R.G. de Jong, M.P.J. van der Loo and F.W. Takes, The anonymization problem in social networks, arXiv 2409.16163, 2024.
R.G. de Jong, M.P.J. van der Loo and F.W. Takes, The effect of distant connections on node anonymity in complex networks, *Scientific Reports* 14: 1156, 2024.
R.G. de Jong, M.P.J. van der Loo and F.W. Takes, Algorithms for Efficiently Computing Structural Anonymity in Complex Networks, *ACM Journal of Experimental Algorithmics* 28: 1-22, 2023.

Supplementary slides

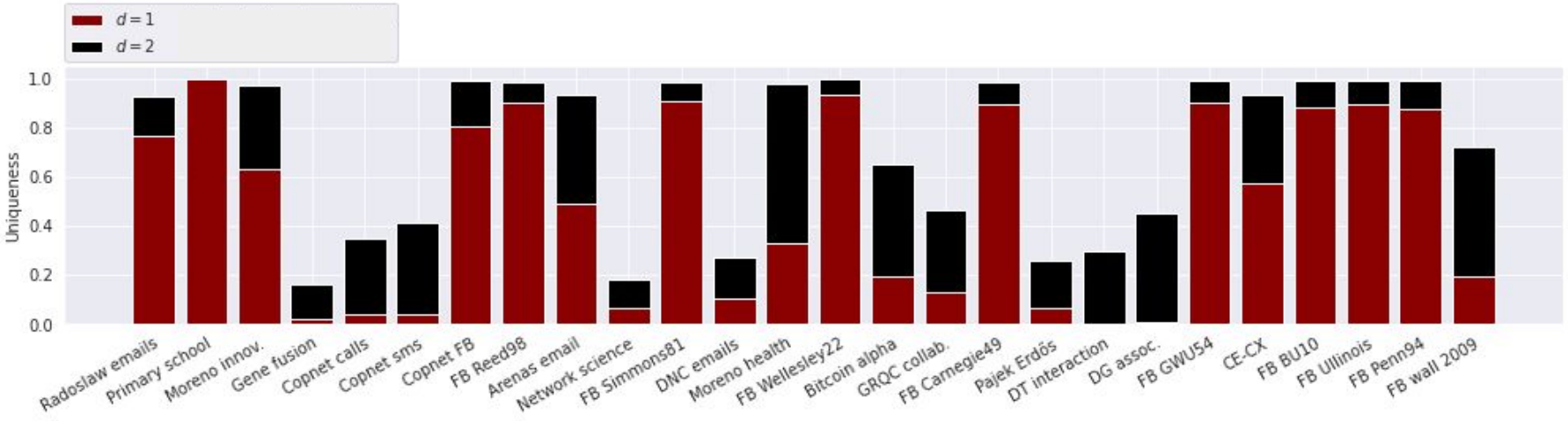

Average Runtime(s)
10^5
10^4
10^3
10^2
Algorihtm
GA
UGA
ES
UA
FB Reed98
Blogs
FB Simmons
College msg.
GRQC collab.
DG assoc.

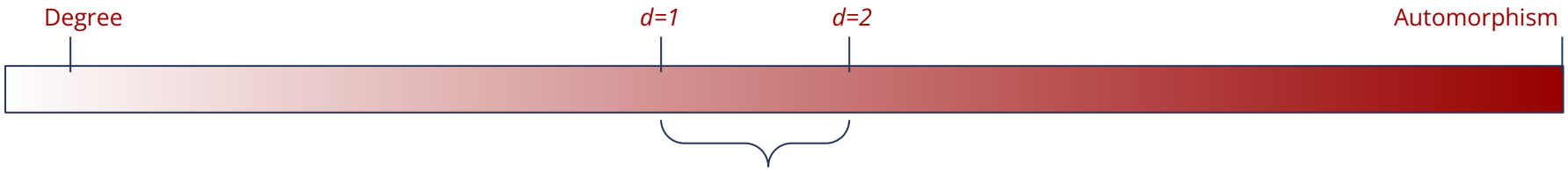
Anonymity measures at distance 2



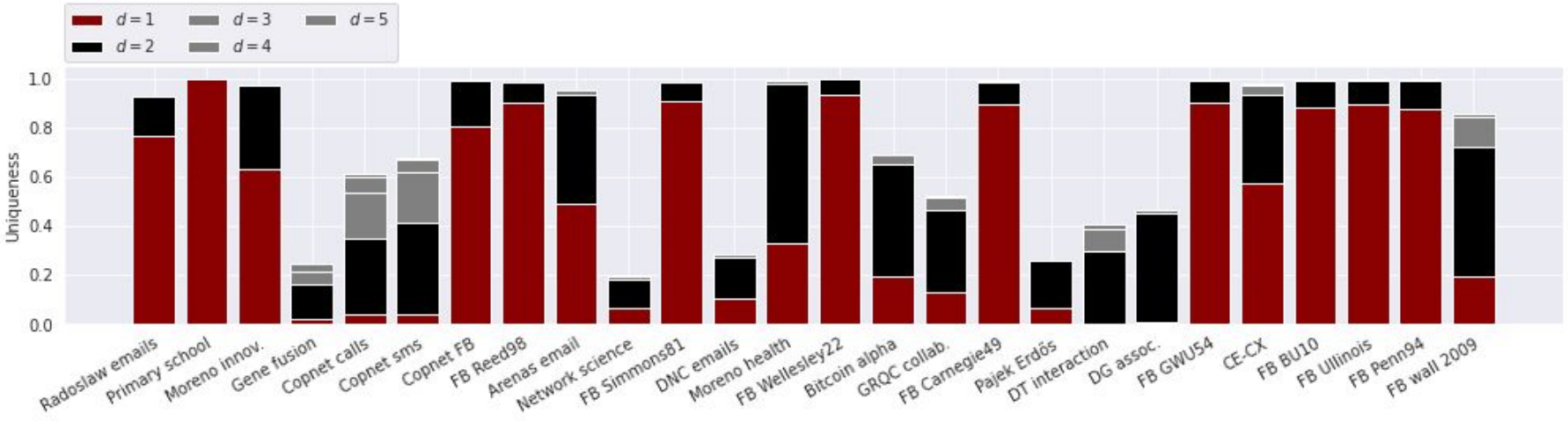

d=4
d=3
d=2
d=1

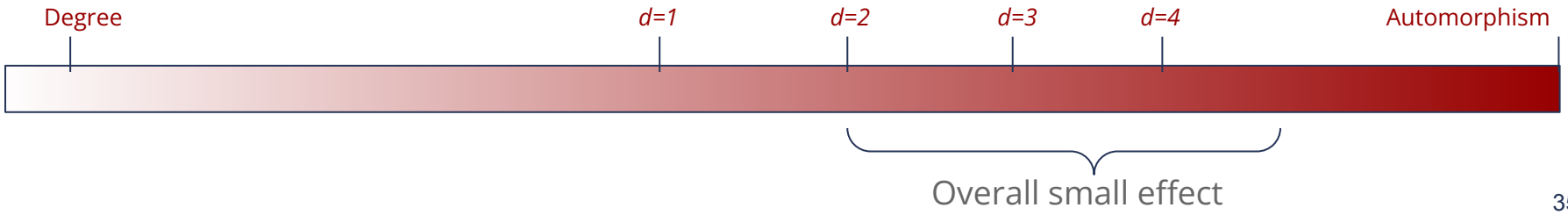
Isomorphism as anonymity measure



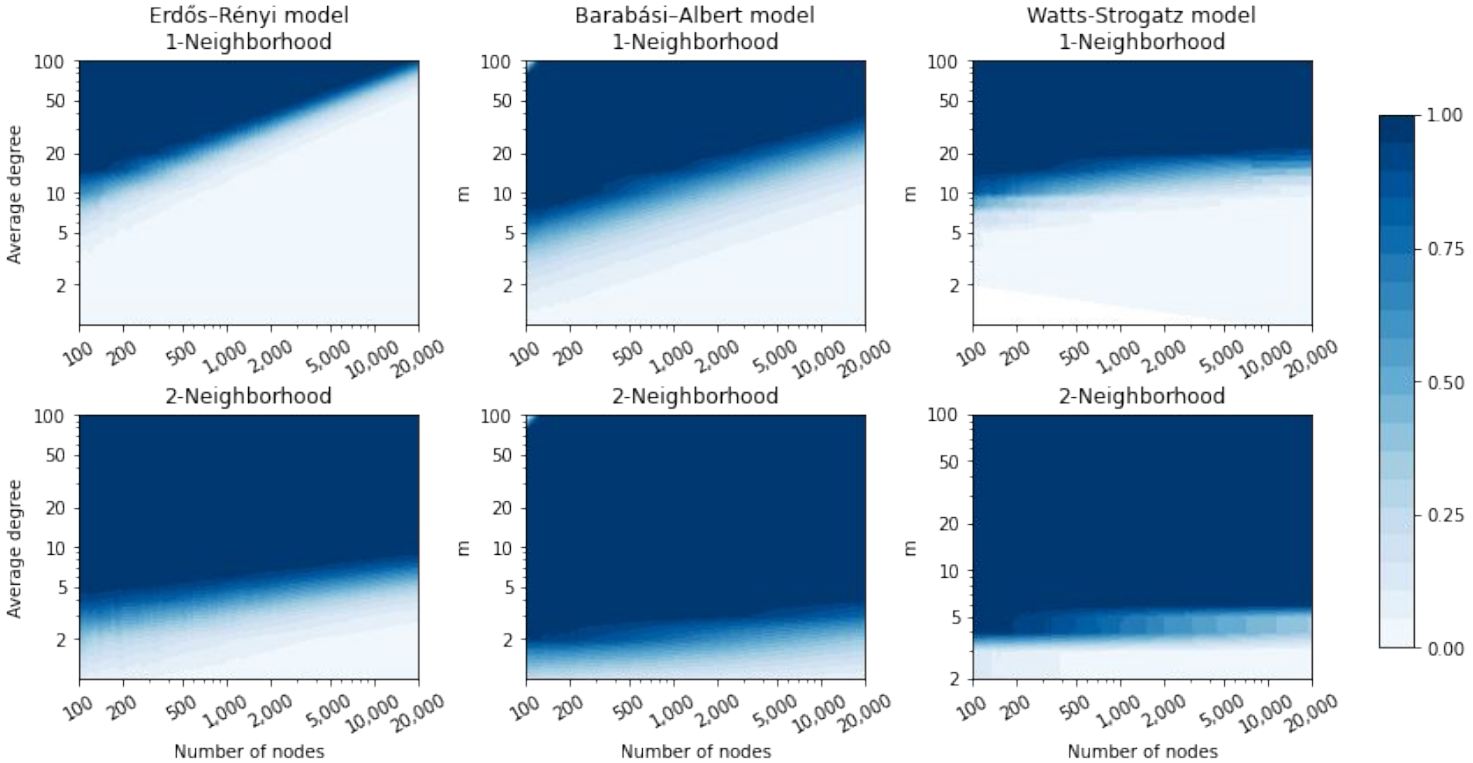


Isomorphism as anonymity measure



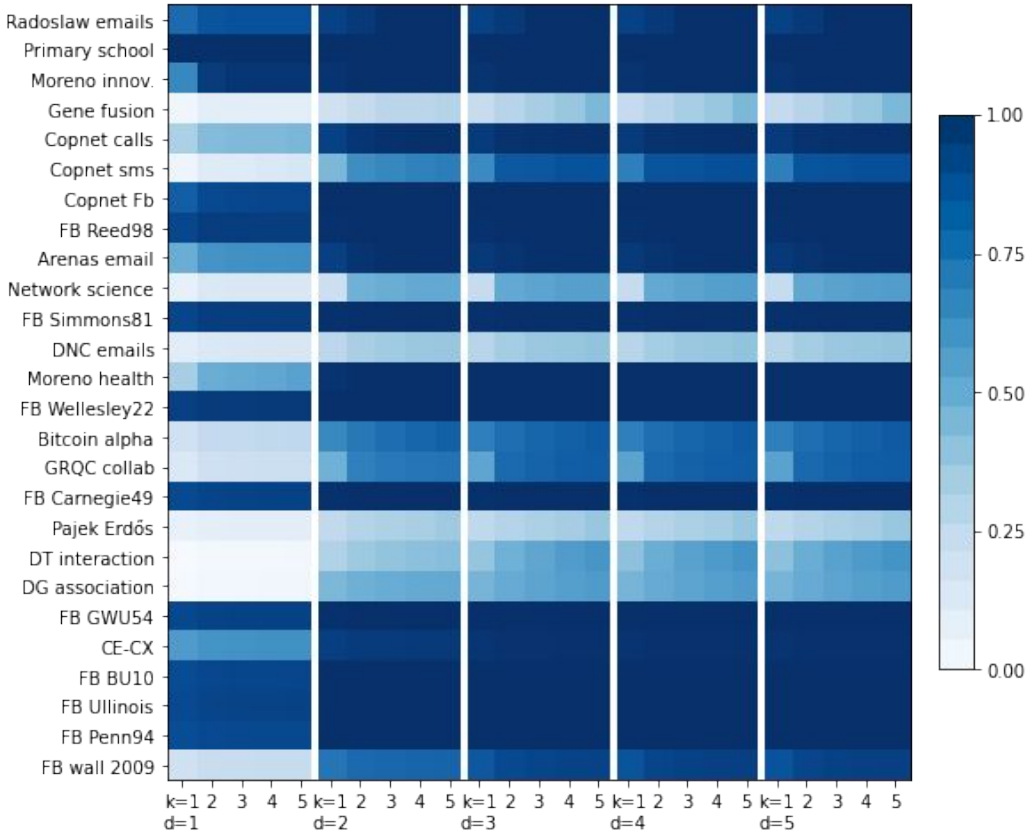


d-*k*-Anonymity

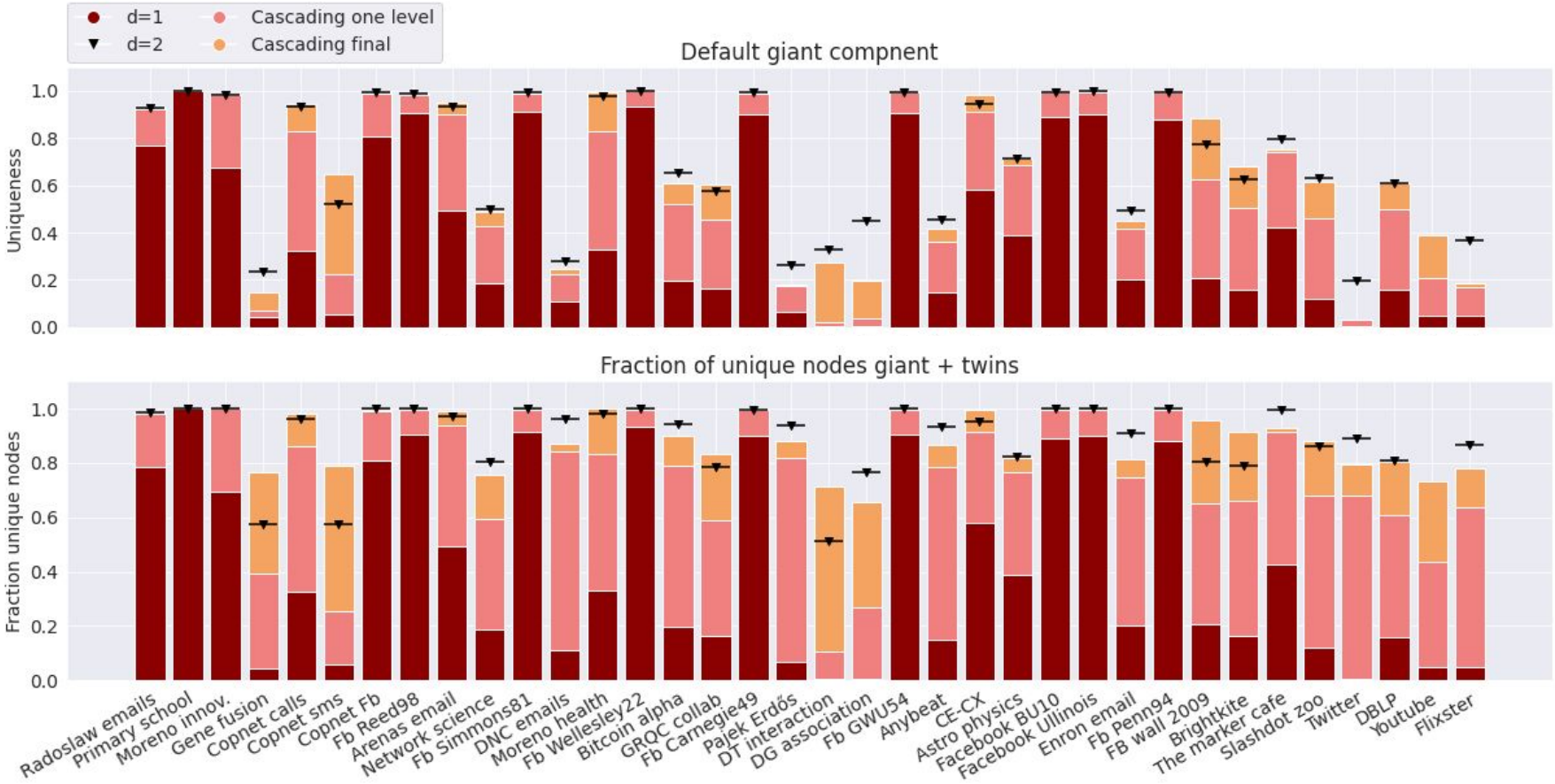


$$\frac{\#\text{unique nodes}}{\#\text{nodes}}$$

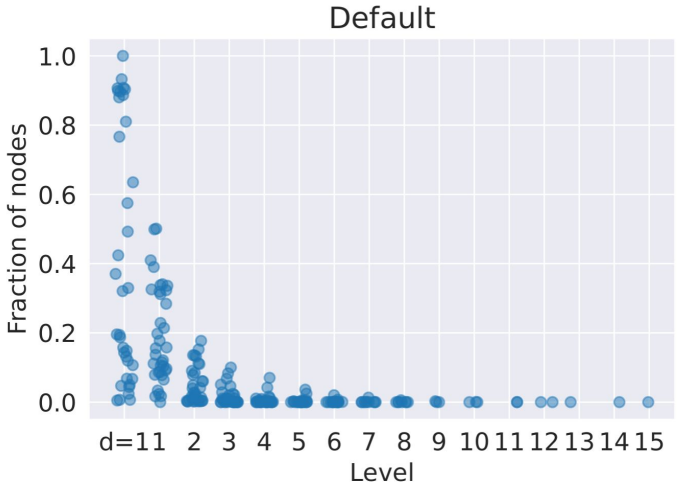
d-*k*-Anonymity

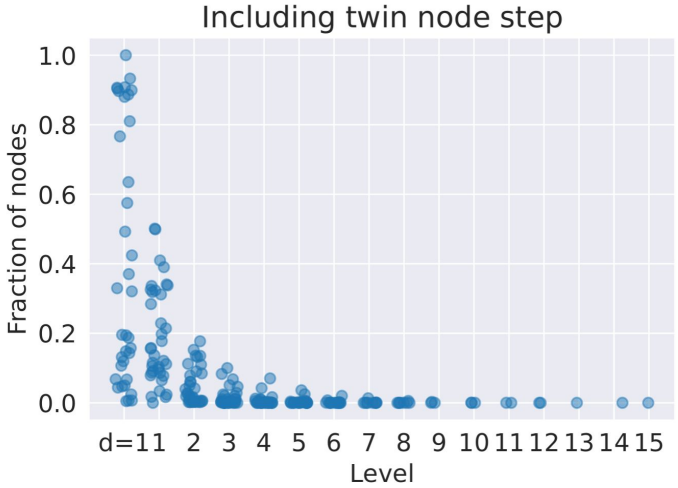


- *d=1* → *d=5*
  - *d=1* is not enough
  - After *d=2* often not much change
- *k=1* → *k=5*
  - Average: +0.06
  - For some networks larger effect
- Focus on:
  - *2*-neighborhood
  - Uniqueness (k=1)

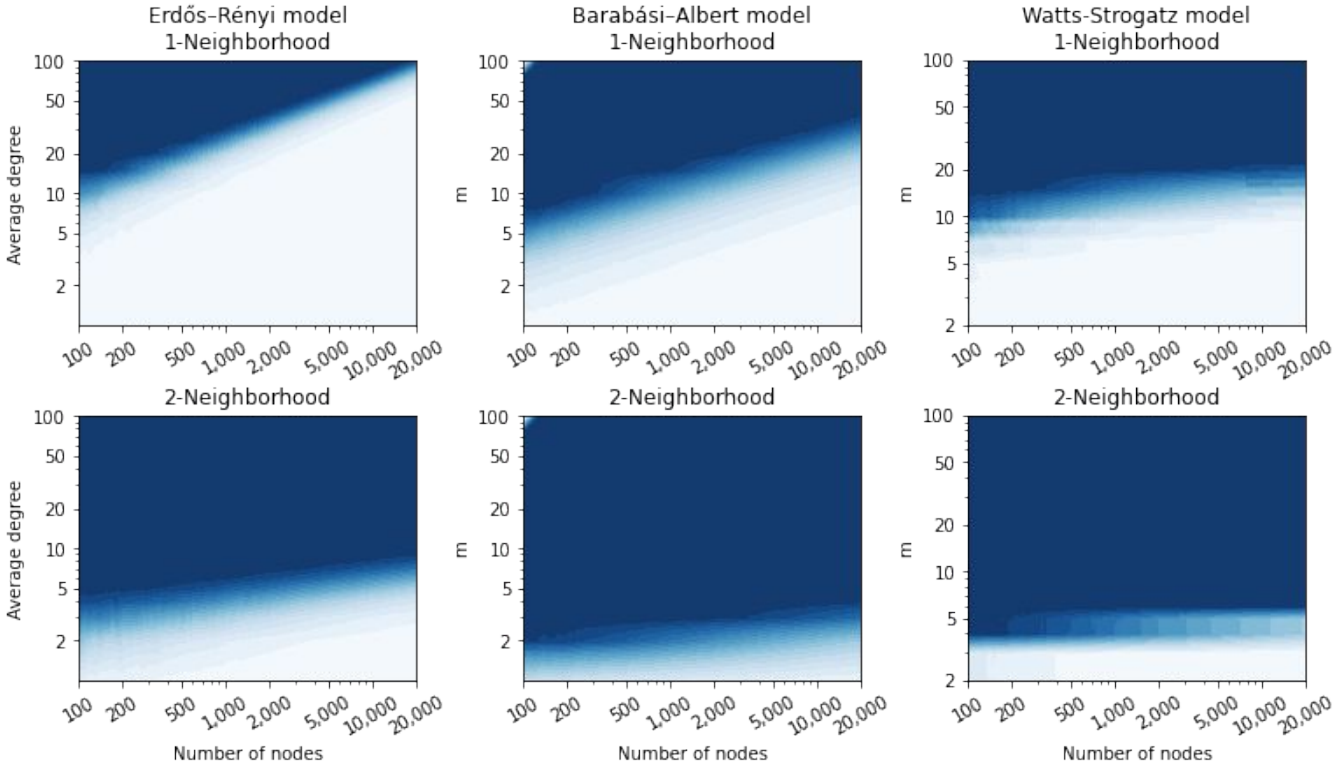

d=1
d=2
Cascading one level
Cascading final
Default giant compnent
Uniqueness
1.0
0.8
0.6
0.4
0.2
0.0
Fraction of unique nodes giant + twins
Fraction unique nodes
1.0
0.8
0.6
0.4
0.2
0.0
Radoslaw emails
Primary school
Moreno innov.
Gene fusion
Copnet calls
Copnet sms
Copnet Fb
Fb Reed98
Arenas email
Network science
Fb Simmons81
DNC emails
Moreno health
Fb Wellesley22
Bitcoin alpha
GRQC collab
Fb Carnegie49
Pajek Erdős
DT interaction
DG association
Fb GWU54
Anybeat
CE-CX
Astro physics
Facebook BU10
Facebook Ullinois
Enron email
Fb Penn94
FB wall 2009
Brightkite
The marker cafe
Slashdot zoo
Twitter
DBLP
Youtube
Flixster

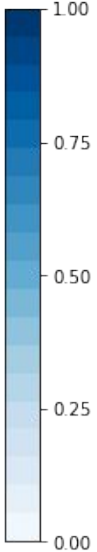
Cascading effect per level





d-*k*-Anonymity on graph models





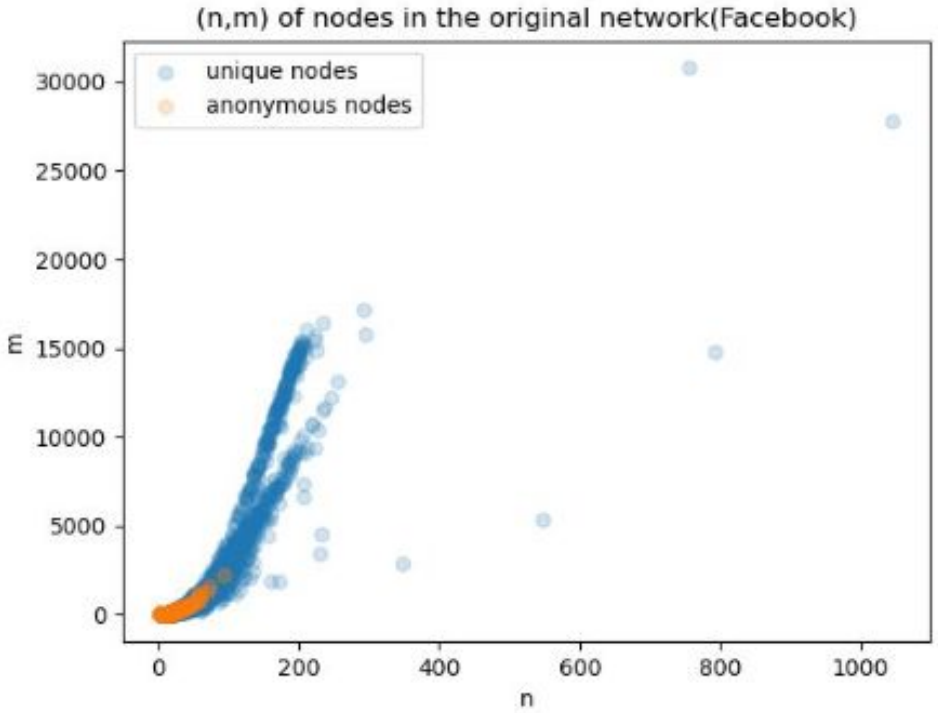
All unique 😦

$$\frac{\#\text{unique nodes}}{\#\text{nodes}}$$

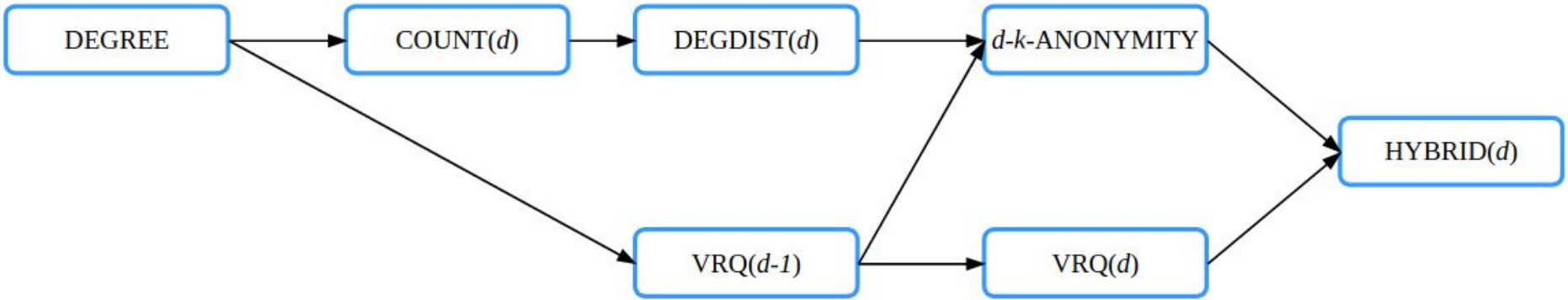
All anonymous 😀

Almost no anonymity for d=2. Size has small / no effect

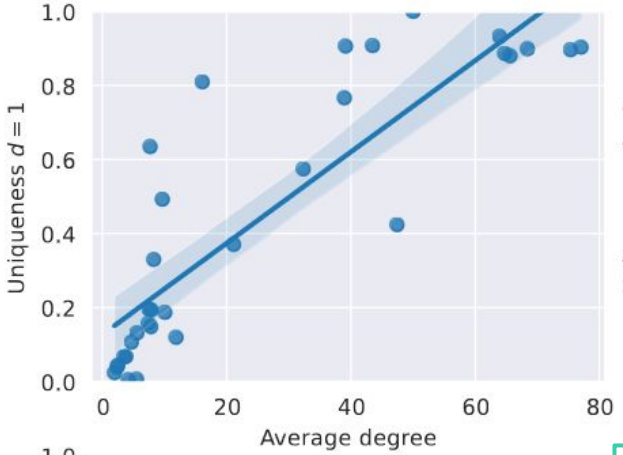
Impossible to anonymize?

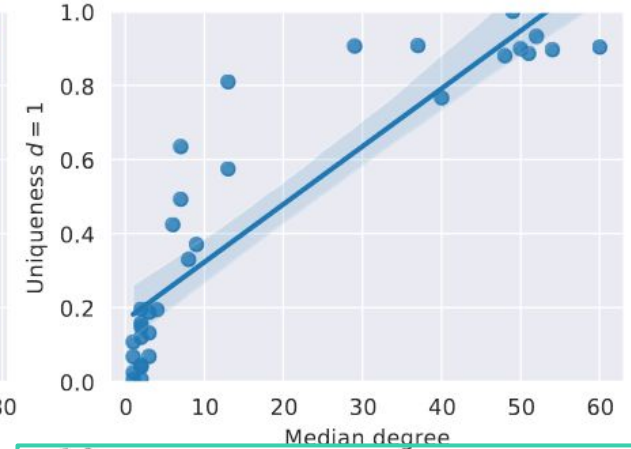


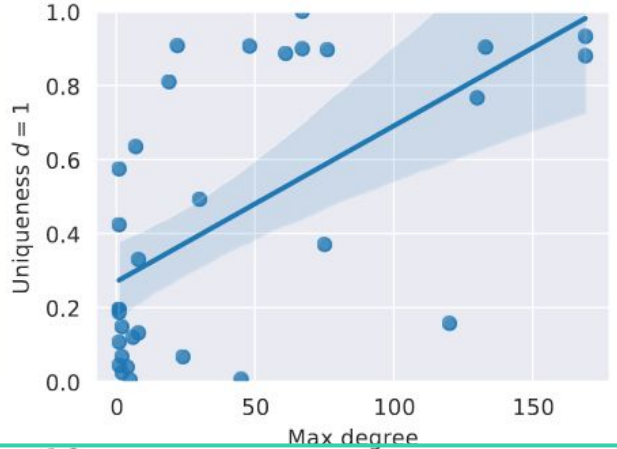
Measures: strictness

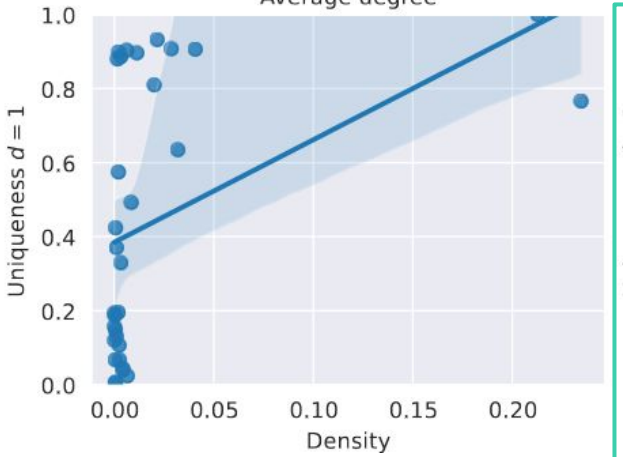


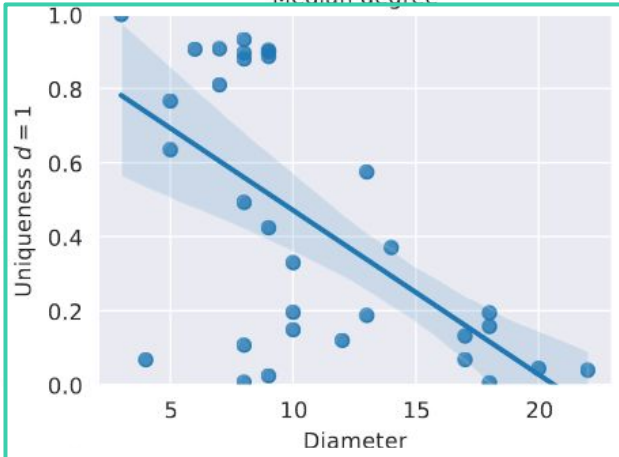
More edges → less anonymity

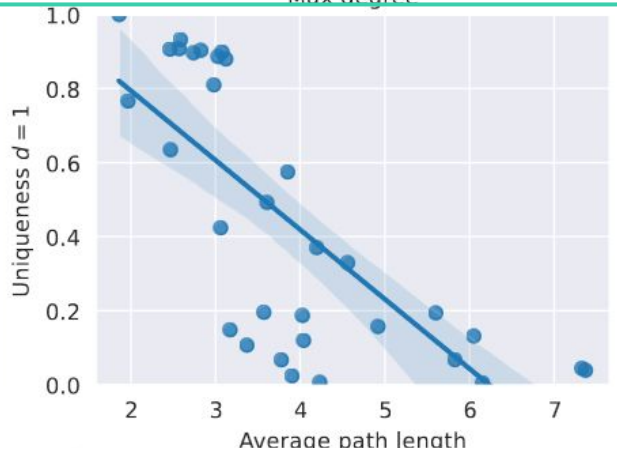


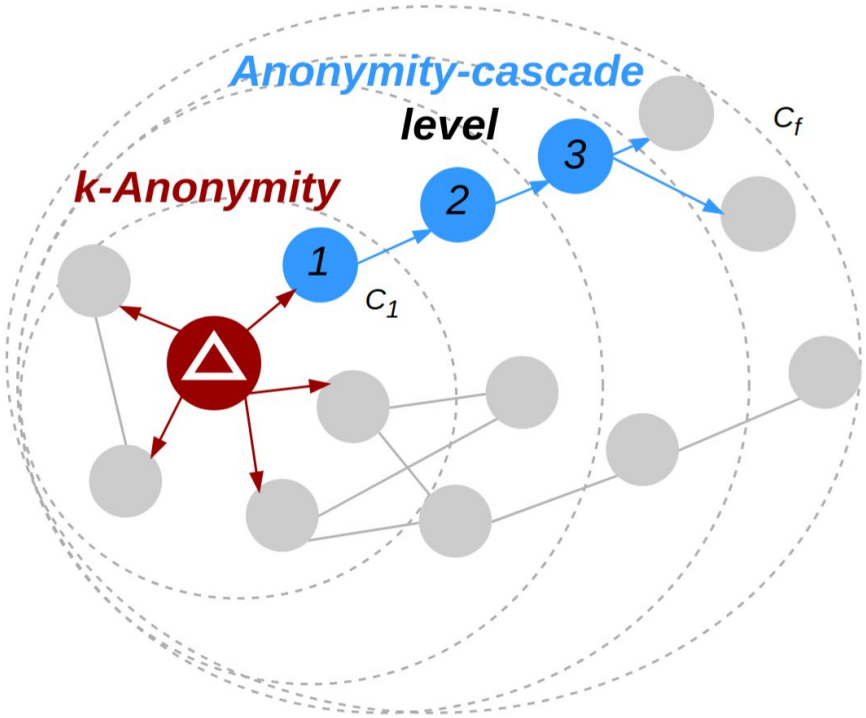










More "spread" network → more anonymity

Anonymity-cascade