Universiteit Leiden
The Netherlands

Population-Scale Networks in Official Statistics

Use Cases and Privacy Implications

Mark van der Loo (mpj.vanderloo@cbs.nl)

Österreichische Statistiktage 2025, JKU Linz 2-4 September

Network Science

Network Science

The study of phenomena emerging from relations between entities.

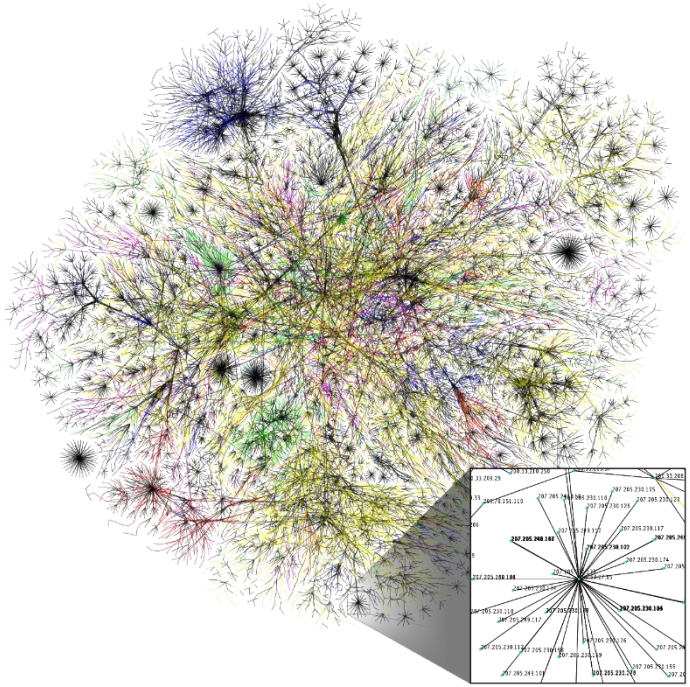
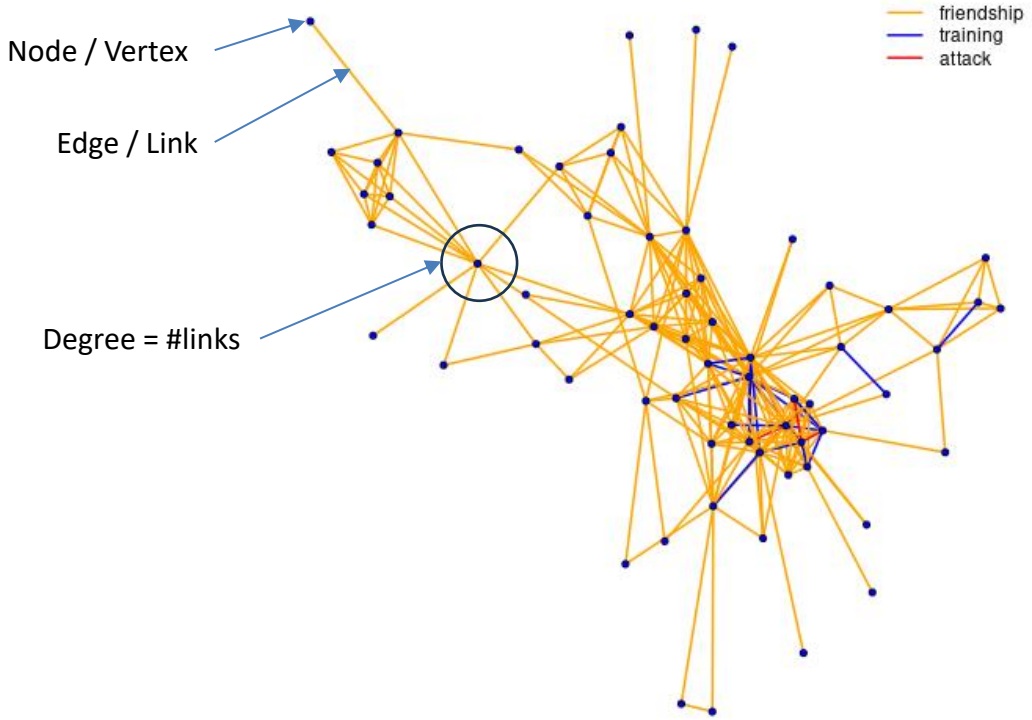


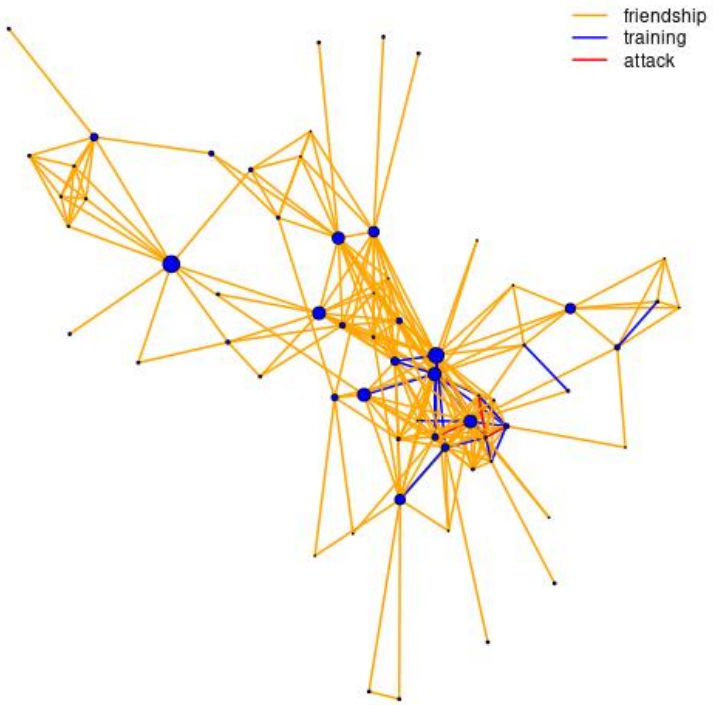
Image: https://en.wikipedia.org/wiki/Internet

Example: the Moreno Terrorist Network



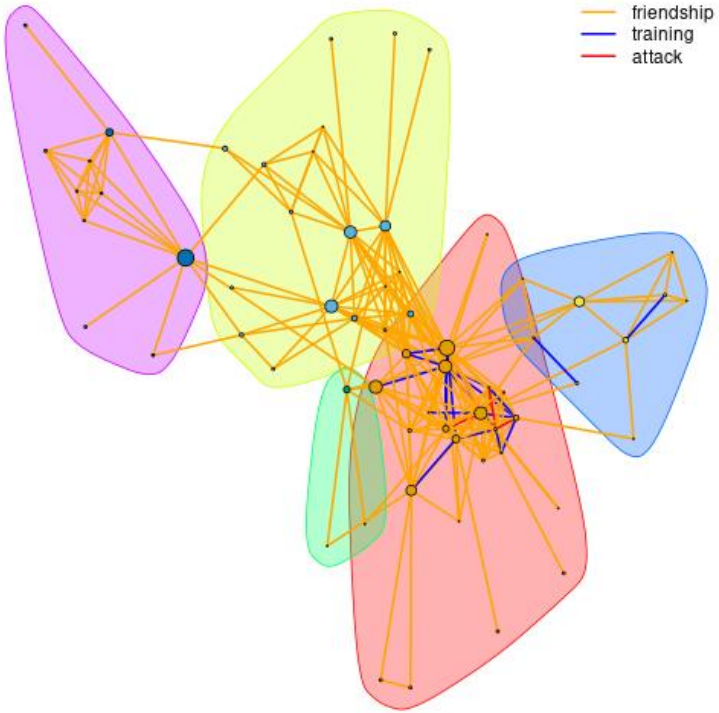
https://konect.uni-koblenz.de/networks/moreno_train

Importance of nodes: centrality



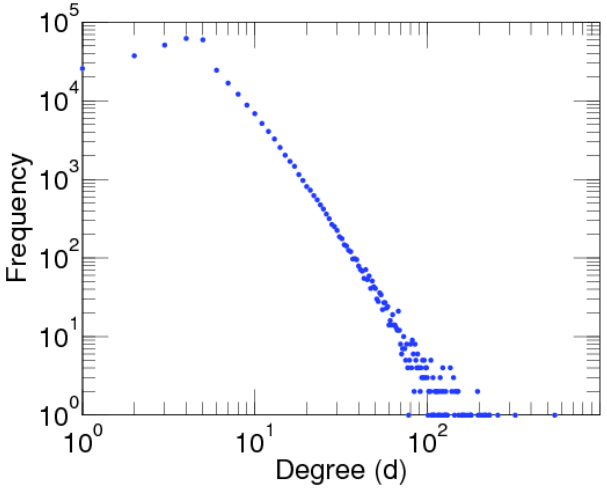
Betweenness centrality

Clustering of nodes

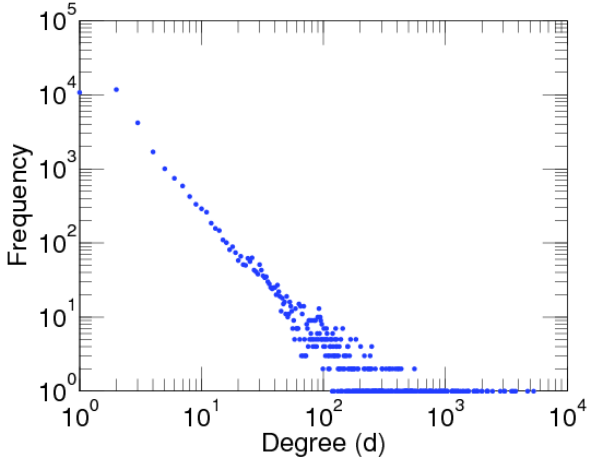


Louvain Clustering

Network models and universal behaviour



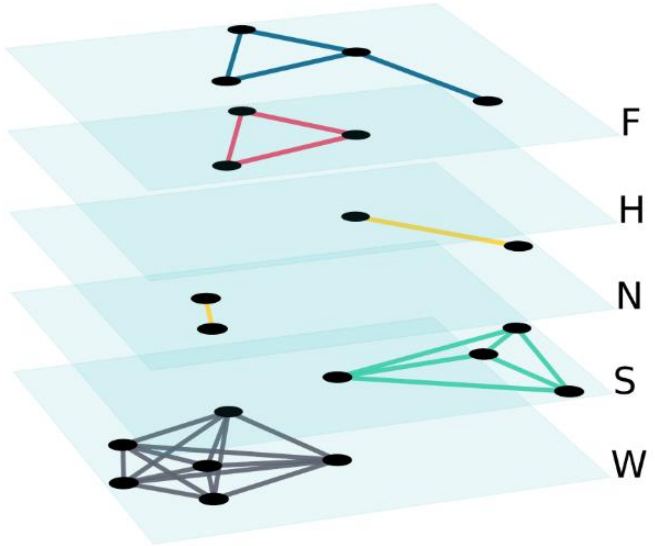
Amazon product co-buying



Internet server connections

Population Scale Social Network

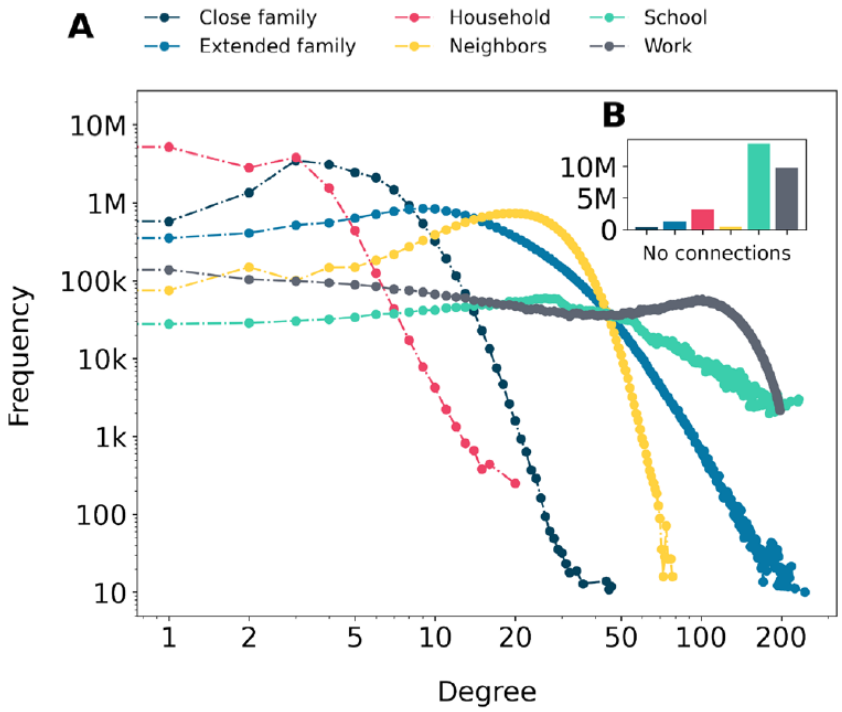
A Population-Scale Social Network Based on Administrative Data



Family
Household
Neighbours
School
Work

Van der Laan *et al* (2023) European Sociological Review **39** 147-160 https://doi.org/10.1093/esr/jcac026

Some properties

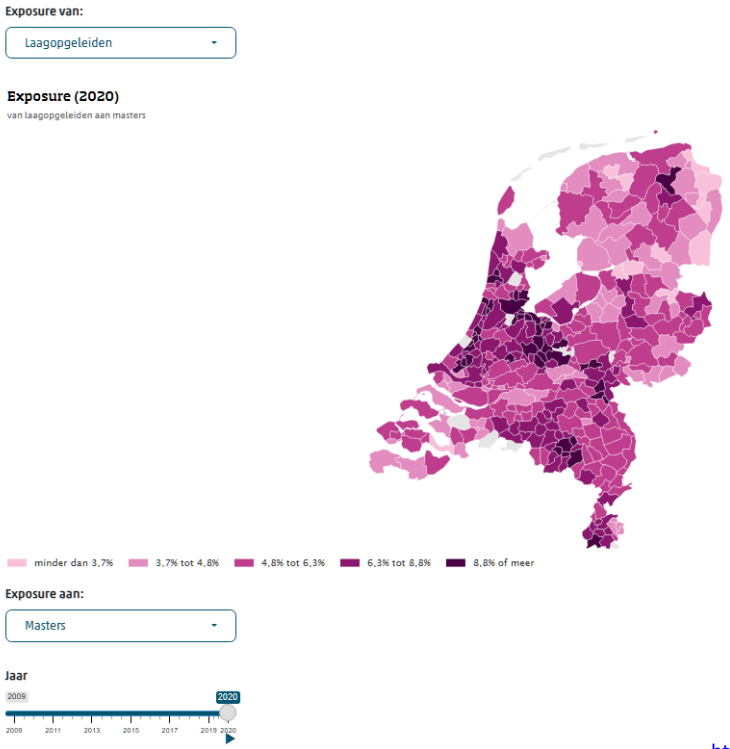


17 Mn nodes
1.4 Bn links

Network as *contact opportunity structure* (social capital)

Bokanyi *et al* (2023) Scientific Reports **13** 9209 https://doi.org/10.1038/s41598-023-36324-9

Educational Seggregation



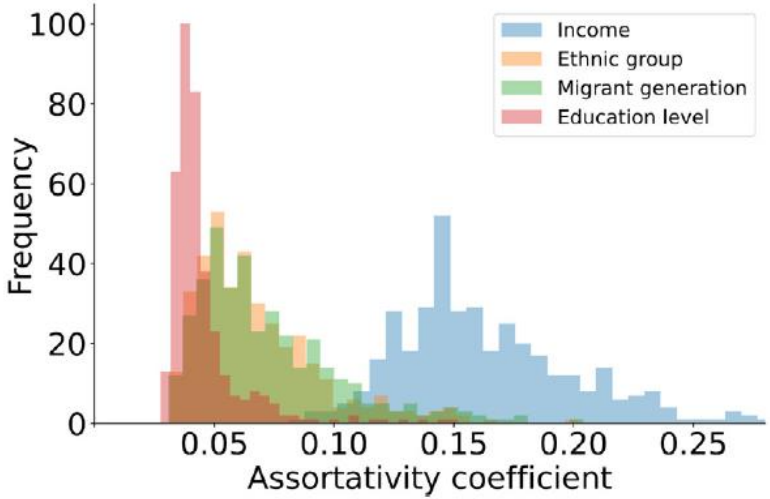
Exposure of educational level A to B:
Percentage of B in the network surroundings of A

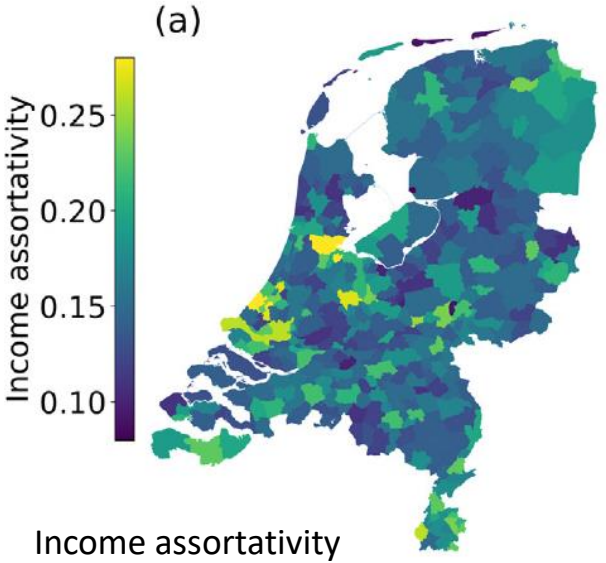
https://dashboards.cbs.nl/v4/opl_segregatie/

Van der Laan *et al* (2023) European Sociological Review **39** 147-160 https://doi.org/10.1093/esr/jcac026

Socio-economic seggregation

Assortativity:
The tendency of similar nodes to be interlinked.

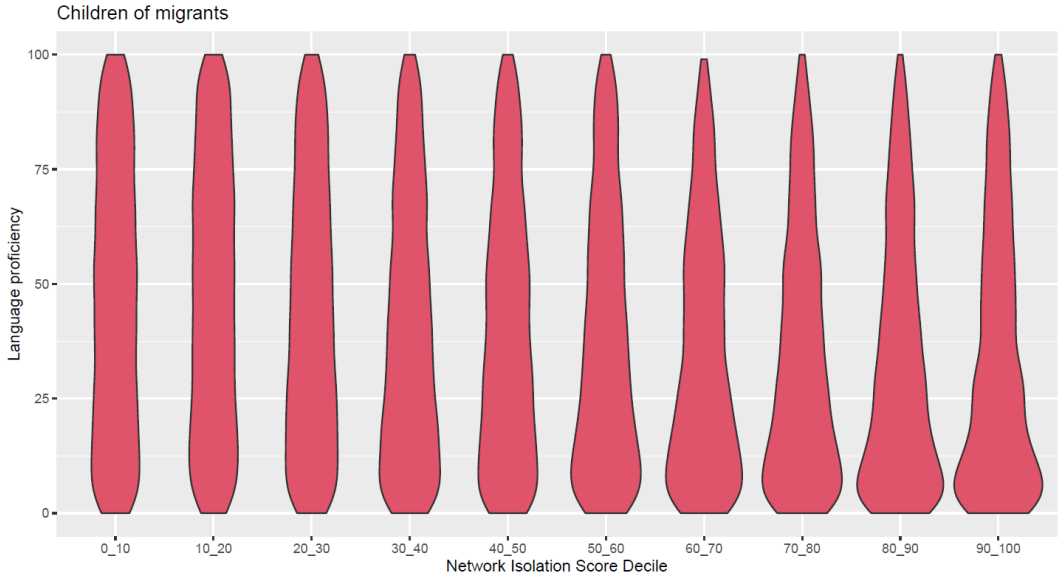




Income assortativity

Kazmina et al (2024) Social Networks **78** 279-291 https://doi.org/10.1016/j.socnet.2024.02.005

Dutch Language Skills of Migrant's Children



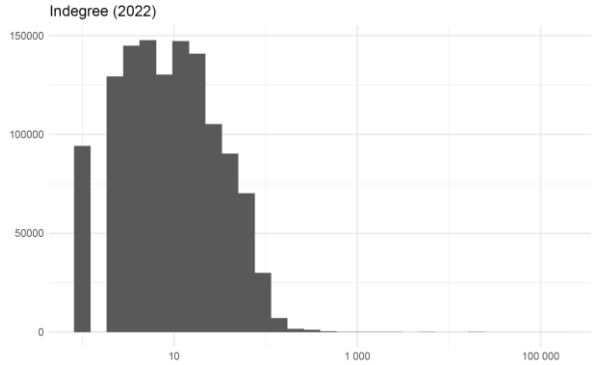
Network Isolation: weighted fraction of persons with same migration background in network surroundings of child.

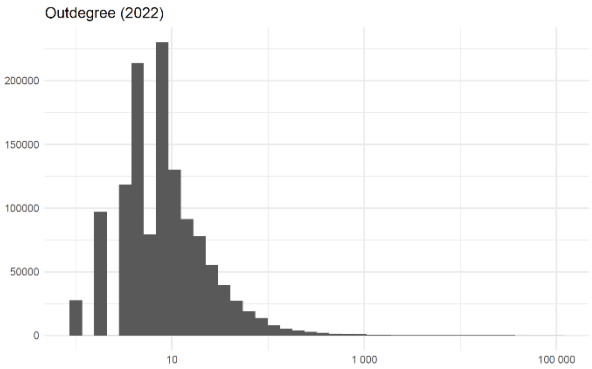
Language proficiency distribution vs Network Isolation

Van Klingeren *et al.* (to be submitted) The role of network isolation on language proficiency of children of migrants
Van der Laan *et al.* (2025) The Effect of the Network of Children of Migrants on their Language Proficiency. IC2S2 Nordkopping (SE)

Population Scale Business Network

The Dutch Production Network





Estimation of trading network

10^6 Businesses

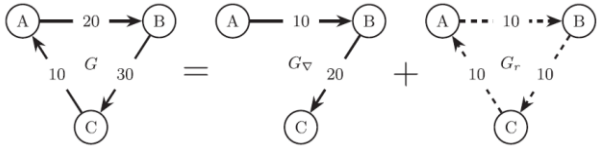
$1.6 \cdot 10^7$ Transactions

523 Commodity groups

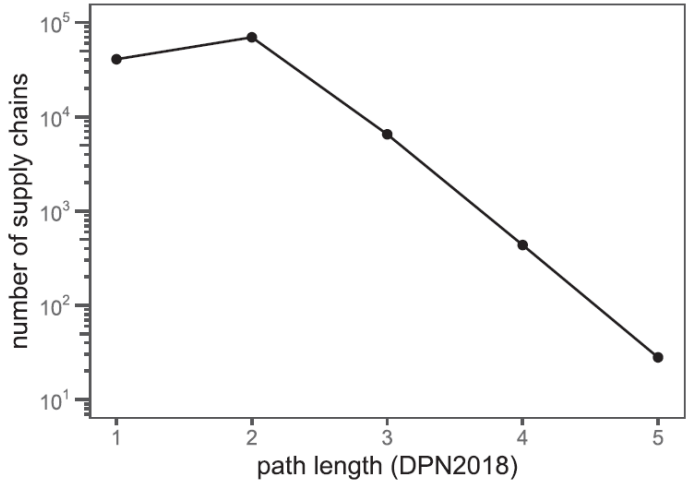
Links are estimated

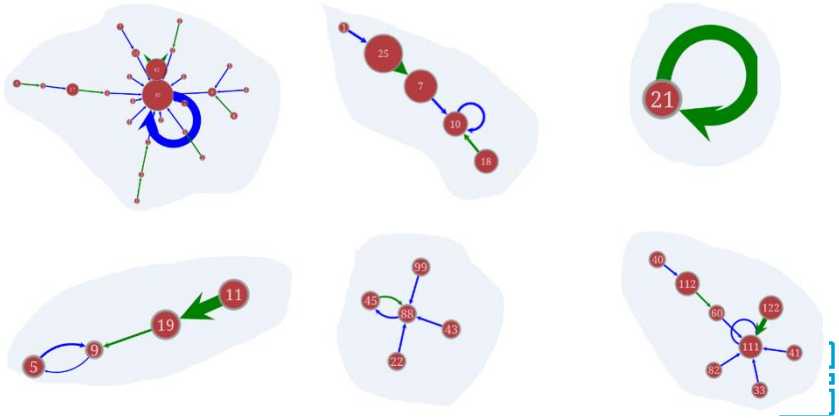
Buiten *et al.* (2021) Technical Report http://dx.doi.org/10.13140/RG.2.2.16685.77286

Detection of Production Chains



Restricted Gradient Graph Algorithm
Splits a graph into cycles and paths.

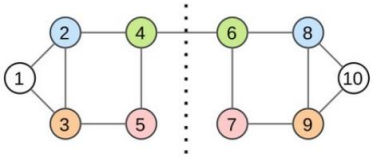




De Jonge (2025) Chaos **25(5)** 053158 https://doi.org/10.1063/5.0270180
Image by Martijn Gösgens (CWI, Amsterdam)

Privacy and Anonymity in Social Networks

Anonymity in Complex Networks



ANO-NET

Universiteit Leiden

POPNET

Centraal Bureau voor de Statistiek

Rachel de Jong (1,2)

Mark van der Loo(2,1)

Frank Takes (1)

(1) Computational Network Science Group at the Leiden Institute of Advanced Computer Science
(2) Statistics Netherlands

Research questions

(1) How does **network structure** affect risk of disclosure?

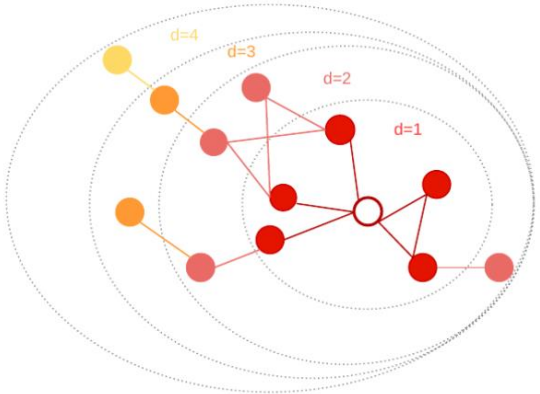
- **Identity**
- Attributes

(2) How can we mitigate disclosure risks?

- Utility

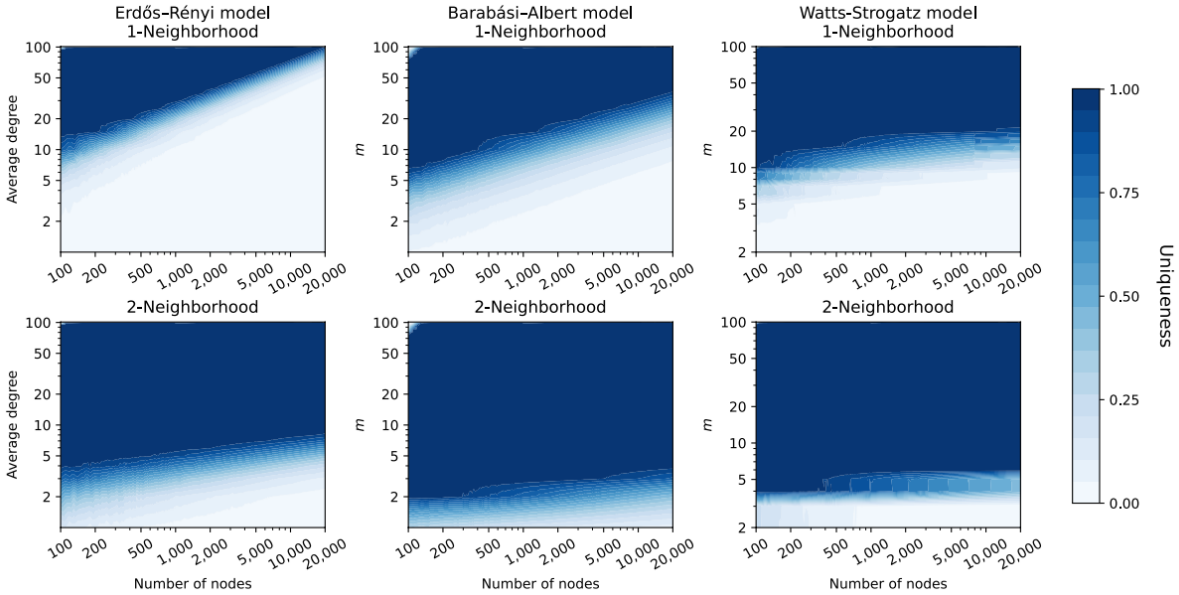
Graded *k*-anonymity: *d-k* anonymity

- Attacker knows *full* network structure surrounding a node, up to and including distance *d*
- *A node is k-anonymous if there are k-1 nodes with the same neighbourhood.*



Van der Loo (2022) Technical Report: https://www.cbs.nl/en-gb/background/2022/17/topological-anonymity-in-networks

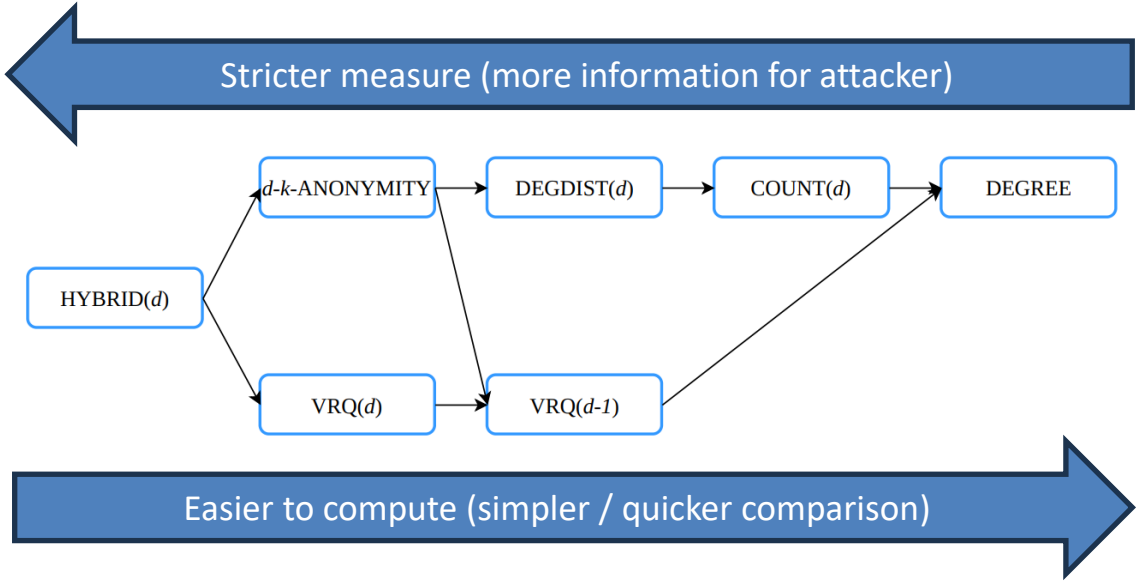
d-k anonymity



Most nodes reidentified:
If surrounding structure upto d ≤ 2 is known exactly

De Jong *et al.* (2023) ACM J of Experimental Algorithmics **28** 1-22 https://doi.org/10.1145/3604908
De Jong *et al.* (2024) Scientific Reports **14** 1156 https://doi.org/10.1038/s41598-023-50617-z

Approximating *d-k* anonymity



A→B: A is stricter than B
Stricter means that the attacker is assumed to have more knowledge.

Incomplete information over larger distance poses more risk than complete information close by.

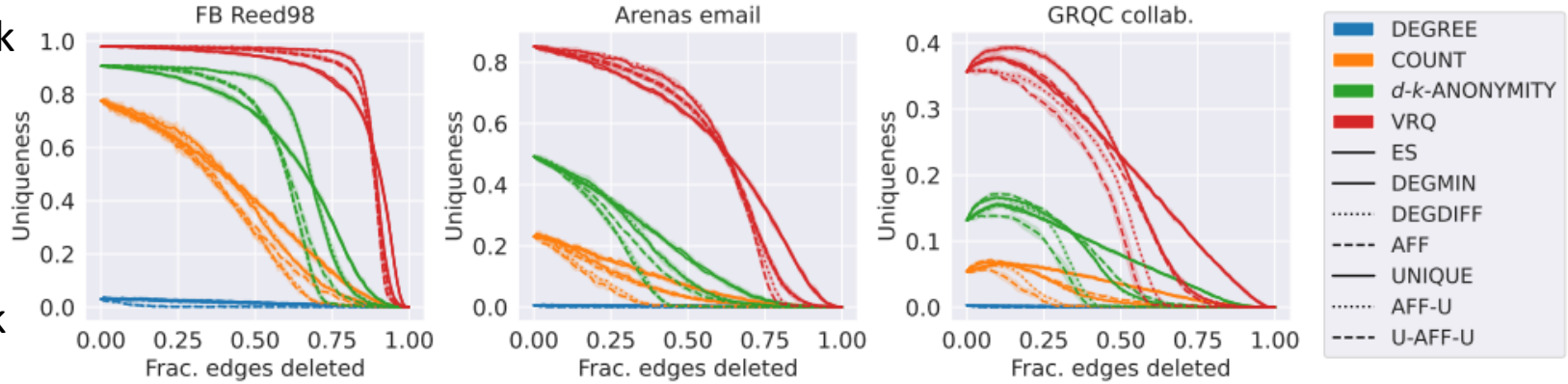
COUNT(*d*) is a good approximation for *d-k* anonymity and faster at d=1

De Jong *et al.* (submitted) https://arxiv.org/abs/2407.02290

Experimenting with anonymization

High risk

Low risk



Fraction of links deleted

- **Link deletion is better** than adding or swapping
- **Targeted deletion works *much* better** than random deletion

De Jong *et al.* (submitted) https://arxiv.org/abs/2409.16163

Take-home messages

Network data is highly interesting for Statistical Institutes, which often have unique data to construct them.

There already are interesting use cases, **but we are just scratcing the surface!**

We are working on **measuring and mitigating disclosure risk** when sharing network data

Thank you!

computationalnetworkscience.org
Computational Network Science at Leiden University

markvanderloo.eu
mpj.vanderloo@cbs.nl