Population-scale Social Network Analysis

Frank Takes

7th International Conference on Computational Social Science - IC2S2 2021

POPNET

Universiteit van Amsterdam

Universiteit Leiden

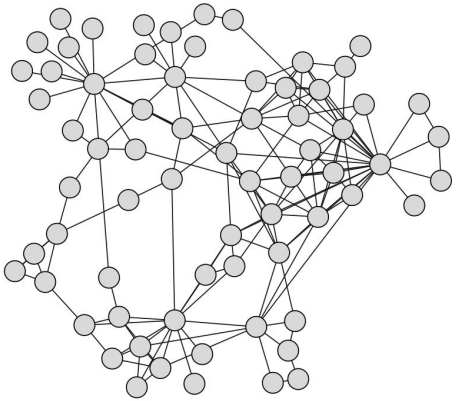
Centraal Bureau voor de Statistiek

Platform Digitale Infrastructuur Social Sciences & Humanities

Talk outline

- What is **population-scale social network analysis**?
- What **challenges** does it bring at the intersection of
  - theory,
  - methods, and
  - data?
- The **POPNET** project on population-scale social network analysis
  - Network data on 17 million inhabitants of the Netherlands
  - Sourced from administrative register data
- First empirical results of the project; revisiting a seminal social network phenomenon from a **computational social science** perspective

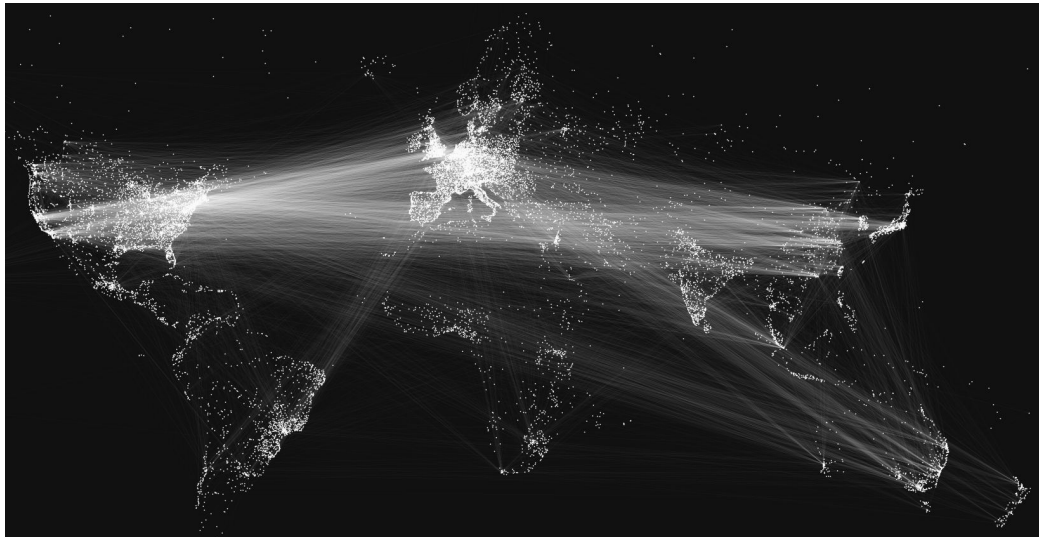
Social network analysis (SNA)



SNA ≈ network science
SNA ≈ complex networks
SNA ~ social complexity
SNA ~ (computational) social science

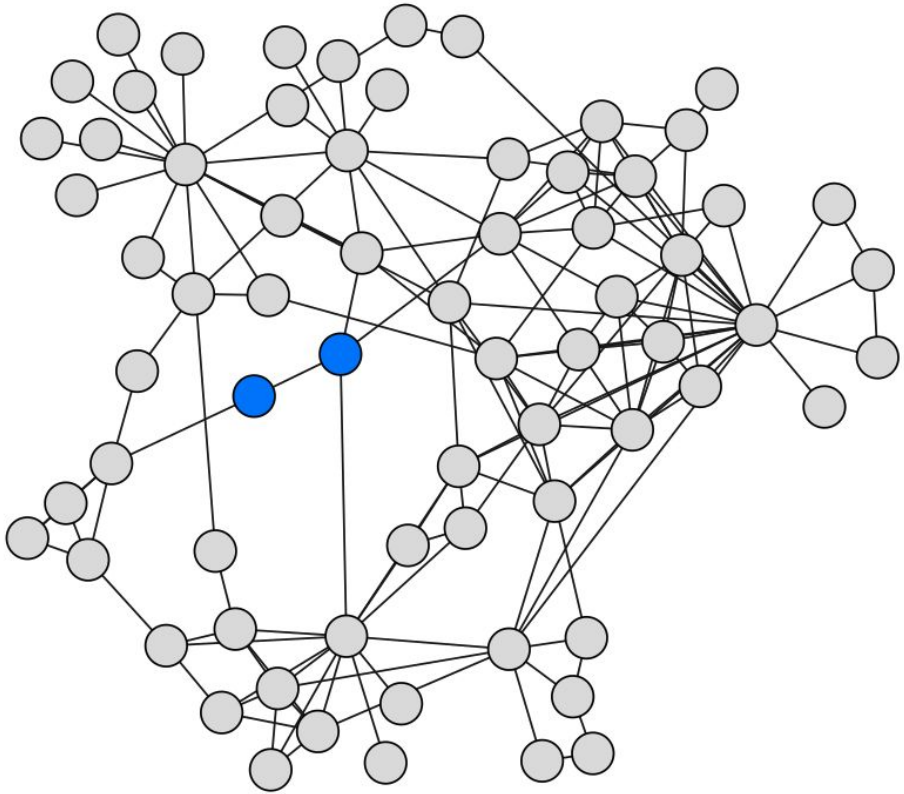
nodes are people

Population-scale?



more than “using some big data”…

Social network analysis



What is a social tie?

kinship, friendship, acquaintance, communication, proximity, …

And: are they all “the same”?

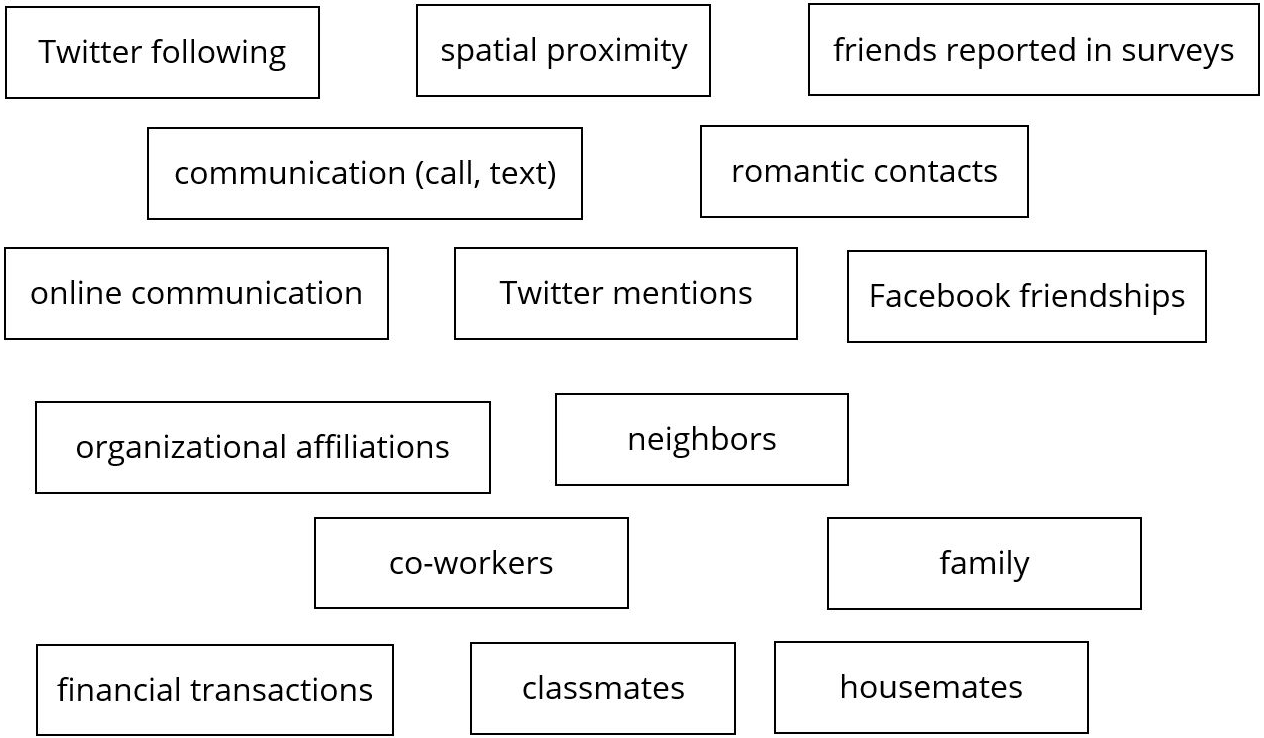
What is a social tie?

- Computational/data scientist: *"I will make a multilayer network model out of whatever connectivity data on people you give me..."*
- Social scientist: *"I think it depends on which grand social science challenge you want to address with your network analysis..."*

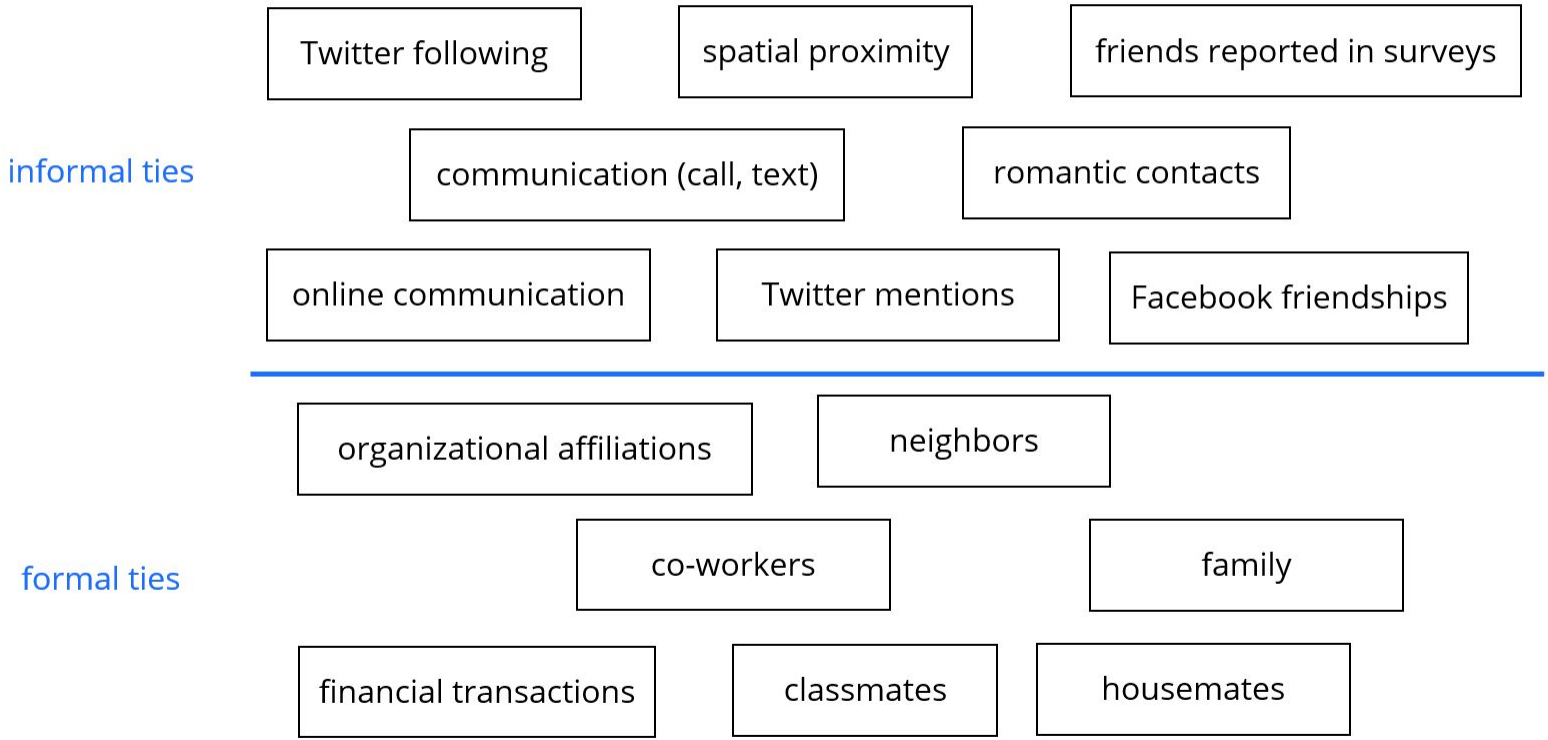
- Computational social scientist: consider, from both a substantive and data-aware point of view, the fundamental **unit of analysis**. This pertains:
  a. the individual
  b. the social tie

Butts, C. T. (2009). Revisiting the foundations of network analysis. *Science*, 325(5939), 414-416.

Social ties


Twitter following
spatial proximity
friends reported in surveys
communication (call, text)
romantic contacts
online communication
Twitter mentions
Facebook friendships
organizational affiliations
neighbors
co-workers
family
financial transactions
classmates
housemates

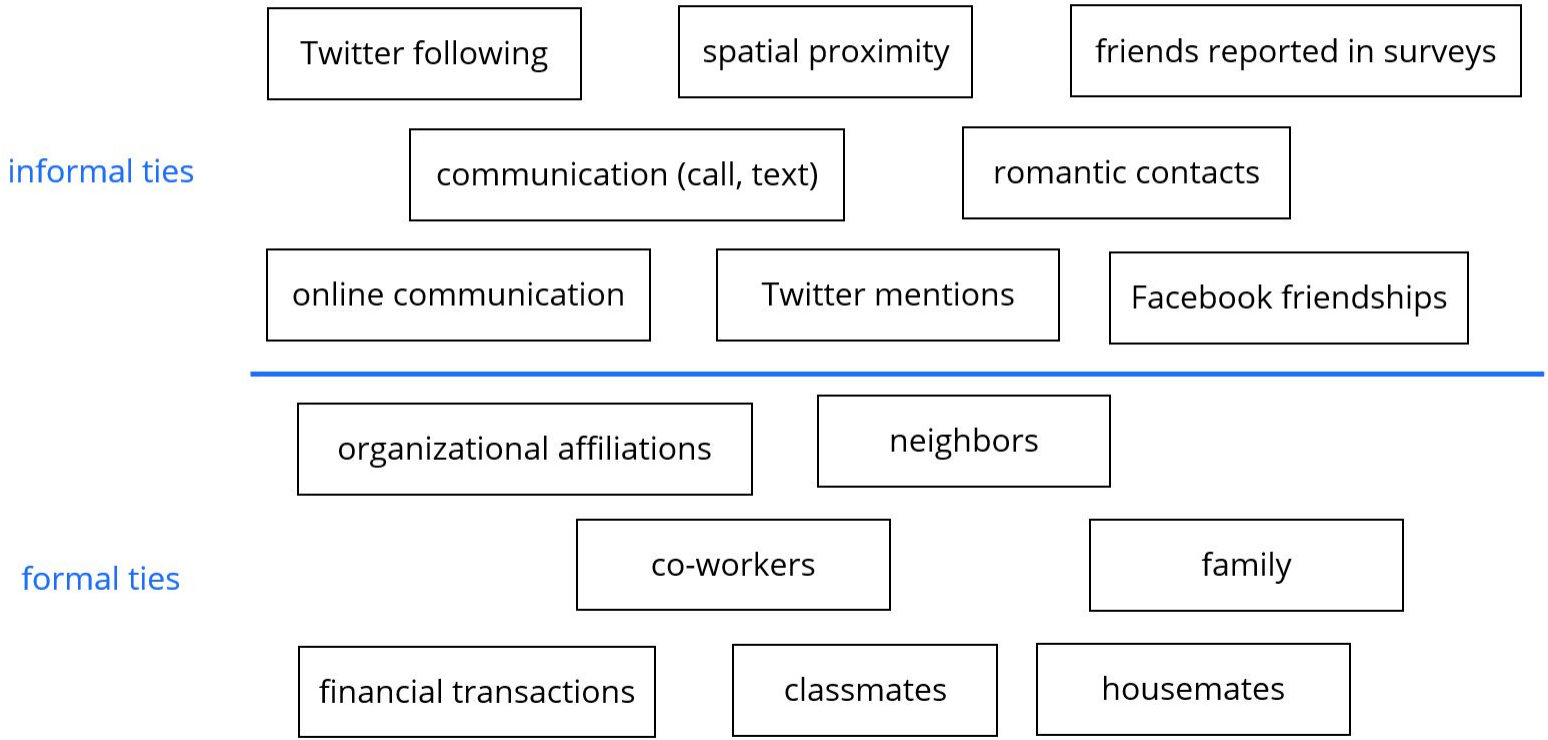
Social ties

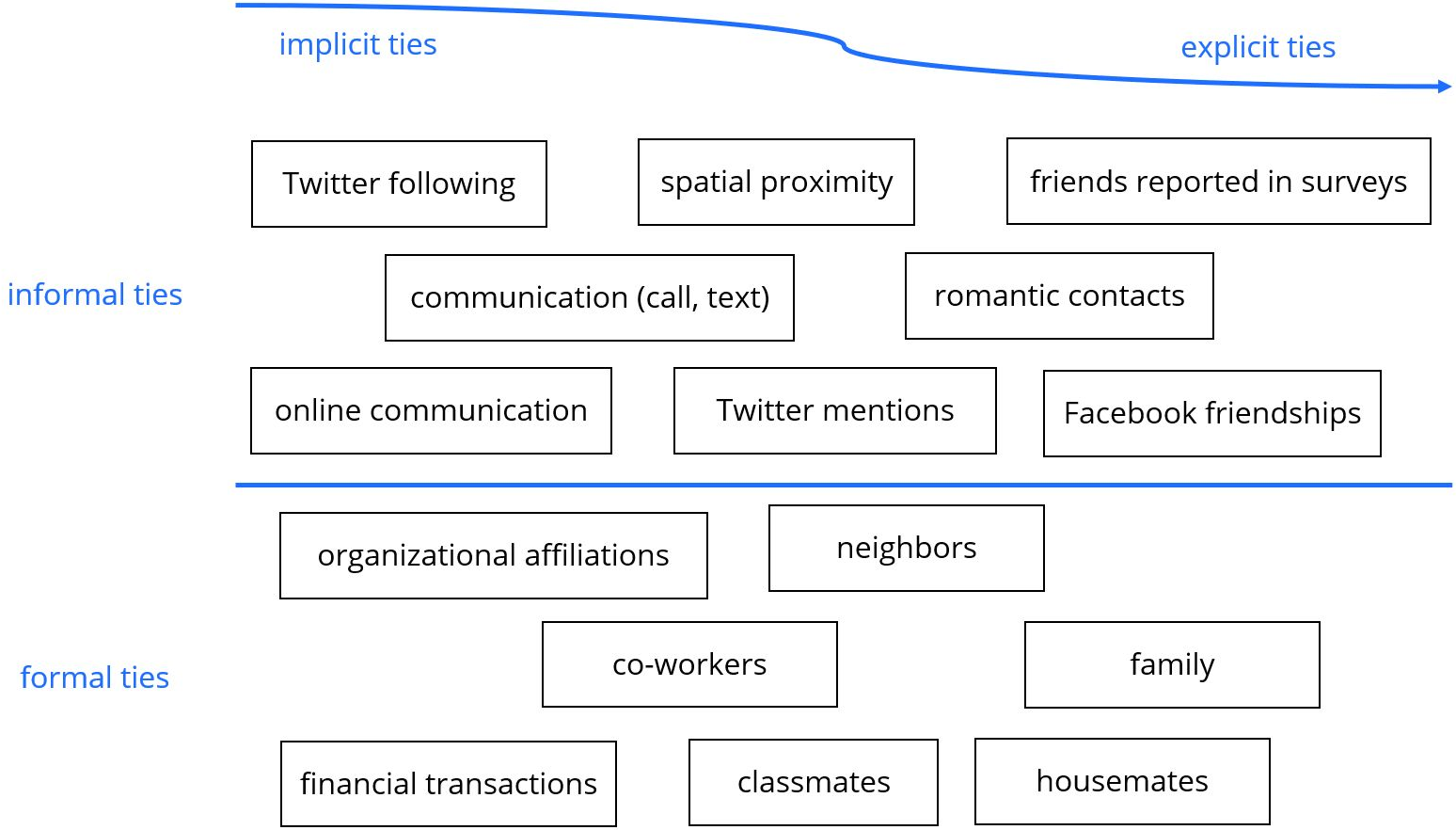

Twitter following
spatial proximity
friends reported in surveys
informal ties
communication (call, text)
romantic contacts
online communication
Twitter mentions
Facebook friendships
organizational affiliations
neighbors
co-workers
formal ties
family
financial transactions
classmates
housemates

Social ties

Formal vs. informal ties

1. Formal ties represent affiliation or connectivity of individuals originating from a well-defined context and accompanying data source (e.g., income tax filings or municipal archives), over which the individual has limited control
2. Informal ties represent relationships caused, created and/or to some extent controlled by the individual(s) involved in the particular tie


Twitter following
spatial proximity
friends reported in surveys
informal ties
communication (call, text)
romantic contacts
online communication
Twitter mentions
Facebook friendships
organizational affiliations
neighbors
co-workers
family
formal ties
financial transactions
classmates
housemates


implicit ties
explicit ties
Twitter following
spatial proximity
friends reported in surveys
informal ties
communication (call, text)
romantic contacts
online communication
Twitter mentions
Facebook friendships
organizational affiliations
neighbors
co-workers
family
formal ties
financial transactions
classmates
housemates

Social ties

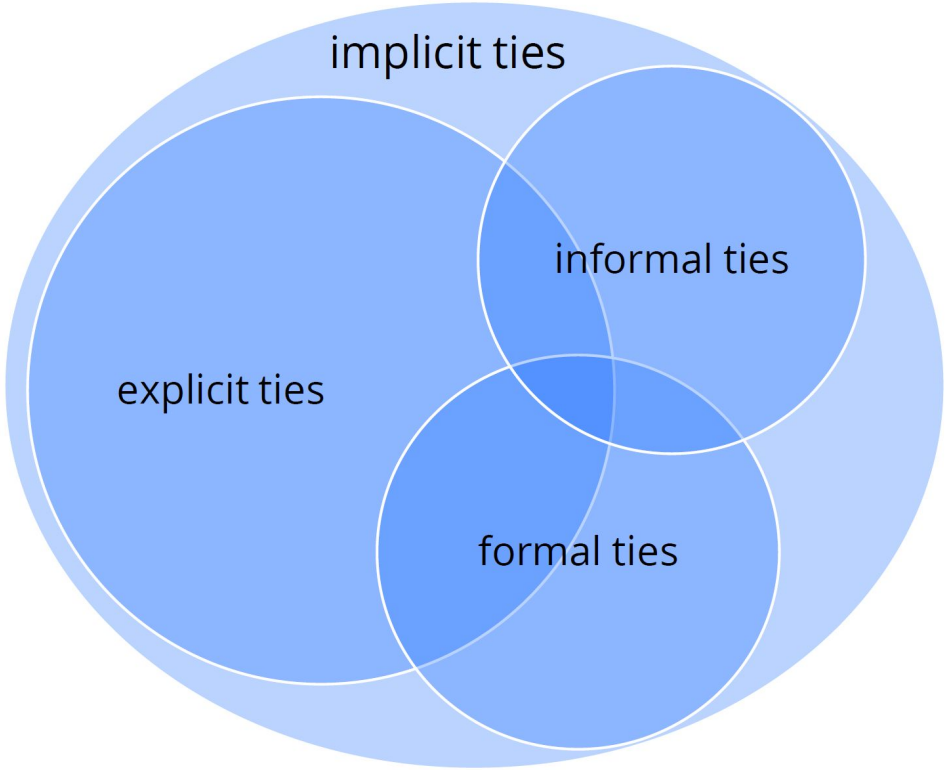
Formal vs. informal ties

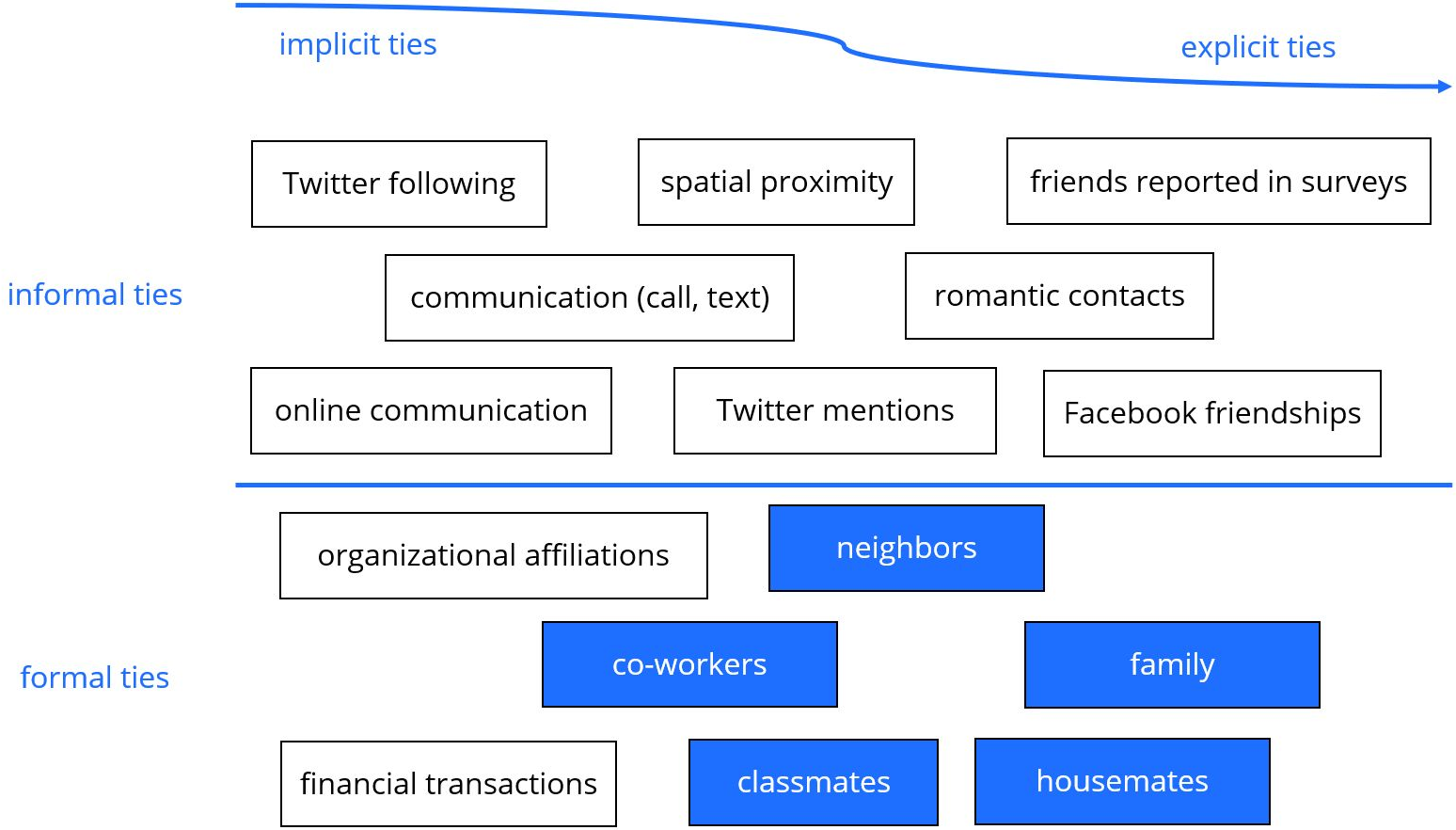
1. Formal ties represent affiliation or connectivity of individuals originating from a well-defined context and accompanying data source (e.g., business registers or family archives), over which the individual has limited control
2. Informal ties represent relationships caused, created and/or to some extent controlled by the individual(s) involved in the particular tie

Implicit vs. explicit ties

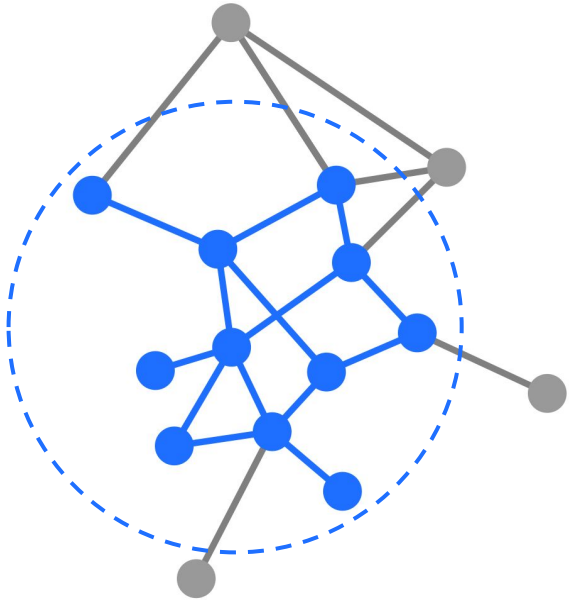
1. Implicit ties are inferred by the researcher (e.g., a social tie is inferred from frequent proximity in a human proximity sensor study)
2. Explicit ties are reported on by the individual (e.g., a person names social ties / friends, or lists these on some social media platform)

Social ties

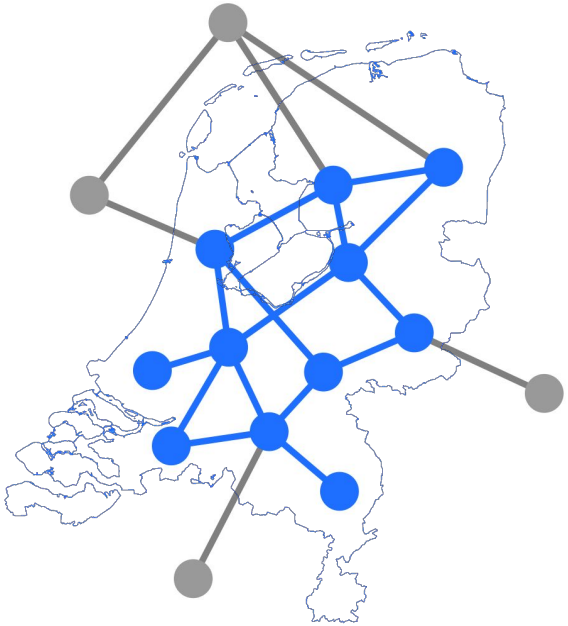



implicit ties
explicit ties
Twitter following
spatial proximity
friends reported in surveys
informal ties
communication (call, text)
romantic contacts
online communication
Twitter mentions
Facebook friendships
organizational affiliations
neighbors
co-workers
family
formal ties
financial transactions
classmates
housemates

Boundary specification



Sampling-induced boundary



Geography-based exact boundary

E.O. Laumann et. al (1989). The boundary specification problem in network analysis. In *Research Methods in Social Network analysis*, pp. 61-87.
D. Lazer et al. (2021). Meaningful measures of human society in the twenty-first century. *Nature* 595: 189-196.

Population-scale social network

- A network of **people** sourced from highly curated register data,
- With unique identifiers for all individuals and their affiliations, and therewith little to no node measurement errors, i.e., **high node accuracy**
- Within a clearly demarcated boundary (typically geography-defined), and therewith **high completeness of nodes** / individuals
- Consisting out of **formal ties** originating from precise contexts with **high completeness**
- Which links are considered (explicit) can be controlled and therewith **control over link accuracy**

E.M. Heemskerk, et al. (2018). The promise and perils of using big data in the study of corporate networks: Problems, diagnostics and fixes, *Global Networks* 18(1): 3-22.
D.J. Wang, et al. (2012). Measurement error in network data: A re-classification. *Social Networks* 34(4): 396-409.
G. Kossinets (2006). Effects of missing data in social networks. *Social networks*, 28(3), 247-268.

POPNET

- **POP**ulation-scale social **NET**work Analysis: **POPNET**
- Started April 1, 2021; duration: 4 years
- Team to be extended further in 2022
- Partnership with Statistics Netherlands (CBS), providing data access
- Network data on 17 million Dutch inhabitants, sourced from register data
- Two main project goals:
  - **Computational social science research on population-scale social networks**
  - Development of sustainable digital research infrastructure
- (all subject to standardized statistical disclosure control procedures and a constant focus on data security, anonymization, privacy and ethics)

POPNET team

POPNET











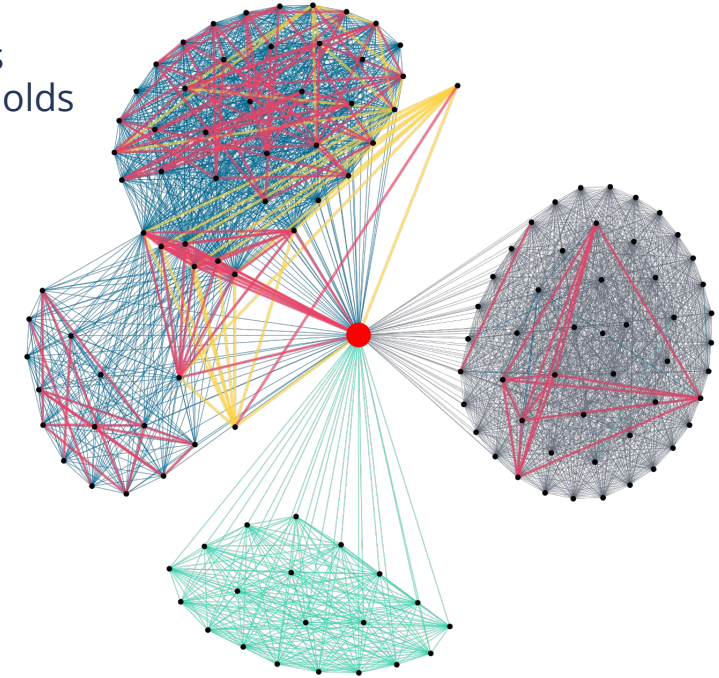


Eelke Heemskerk, Frank Takes, Eszter Bokányi, Yuliia Kazmina, Gert Buiten, Tobias Blanke, Jayshri Murli

Rachel de Jong, Bart de Zoete, Helena Rauxloh, Jan van der Laan, Mark van der Loo, Edwin de Jonge

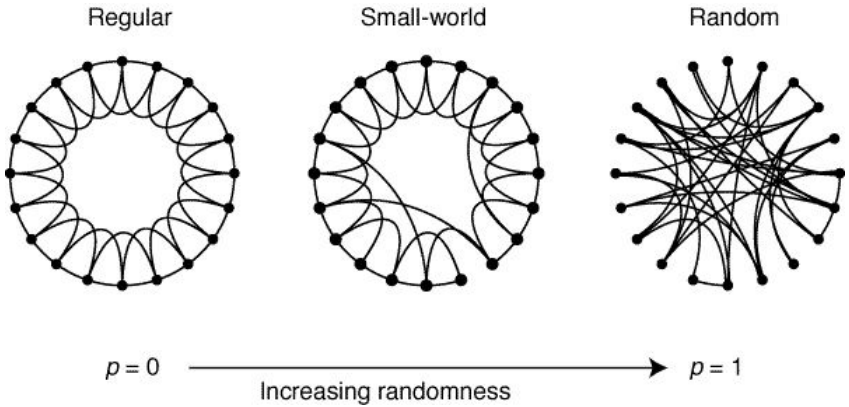
POPNET data

- **Links**
  - **Family**: directed parent relationships and inferred second degree family relations
  - **Households**: people registered at the same address
  - **Neighborhood**: links to people in 10 closest households
  - **Work**: co-workers employed at same organization
  - **School**: classmates at primary school, high school, special, applied and higher education
- **Nodes**
  - Age
  - City / neighborhood
  - Education level
  - Ethnicity
  - Gender
  - Income



A small-world population?

- Small-world networks: a) high local clustering, b) low node-node distances

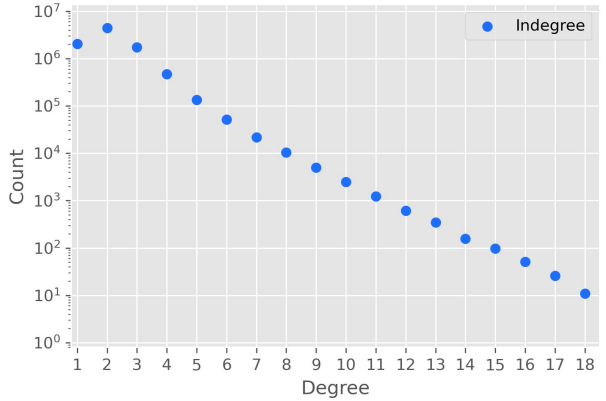


- **Is a population-scale social network a small-world network?**
  - Country-specific OSNs are often found to be small-world, but may suffer from spurious links, low node completeness and sampling bias
  - Population scale social network data contains a lot of missing informal links

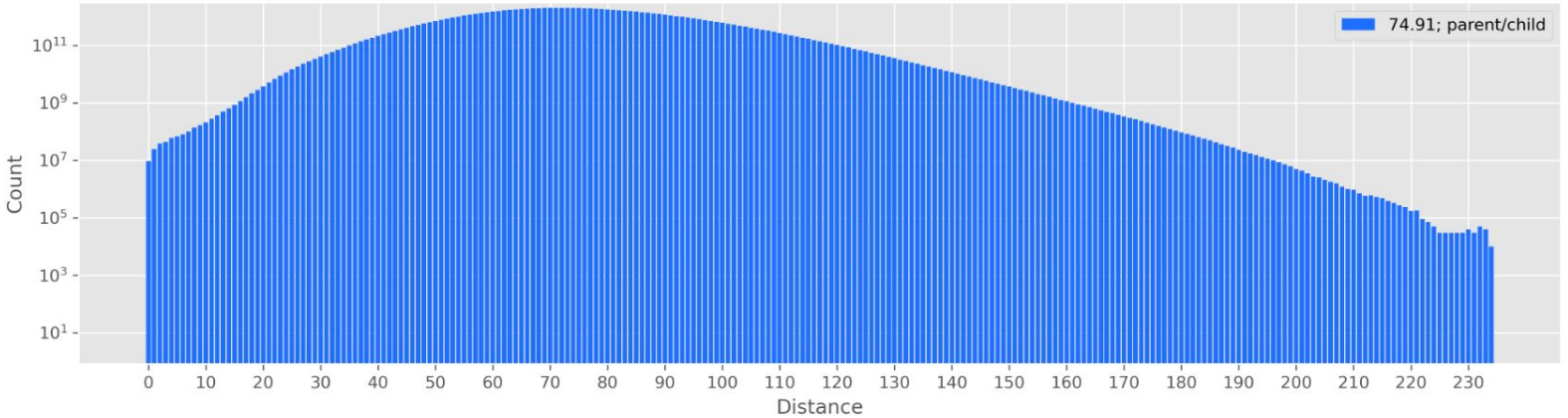
Image credit: D. Watts, S. Strogatz, Collective dynamics of 'small-world' networks. Nature 393, 440–442 (1998)

Parent links

- Directed “family tree of the population”
- 15.76 M nodes, no triangles (no clustering)
- Directed paths of length 1-5 (generations)
- Giant component containing 58% of nodes

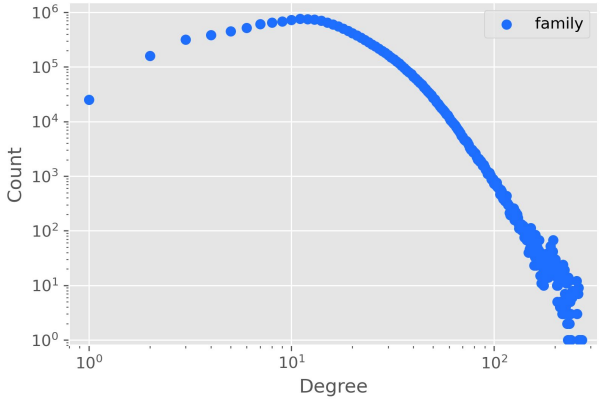


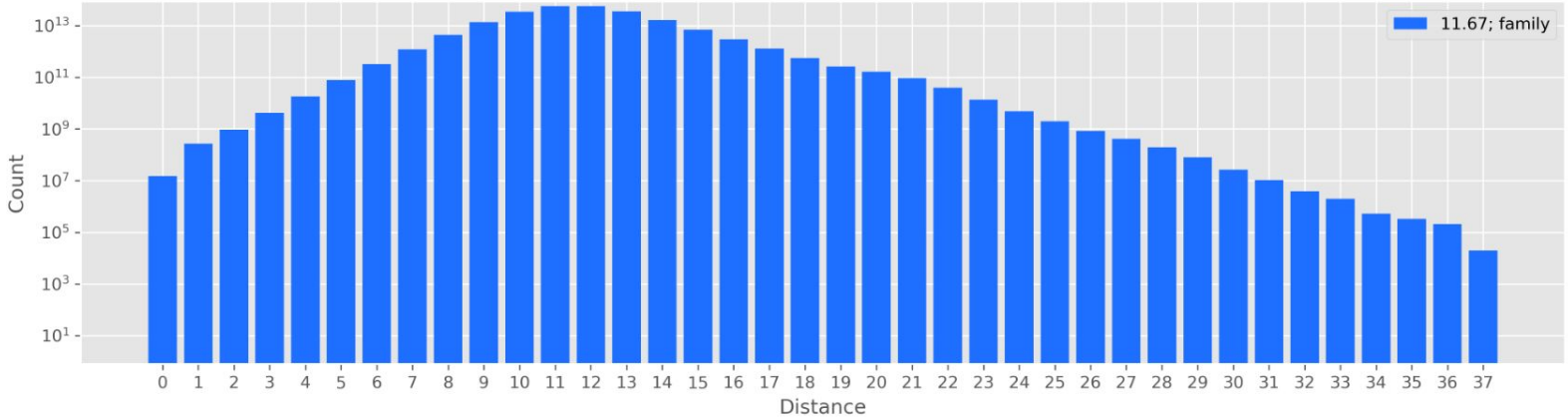
(approximated by computing distance between *200 x n* randomly sampled node pairs using `teexgraph`)



Family network

- Undirected network of all **second degree family relations** (parent/child, co-parents, (half-) siblings, aunts/uncles, cousins, nephews/nieces, grandparents)
- Average clustering coefficient of 0.7
- Giant component containing 97.8% of nodes



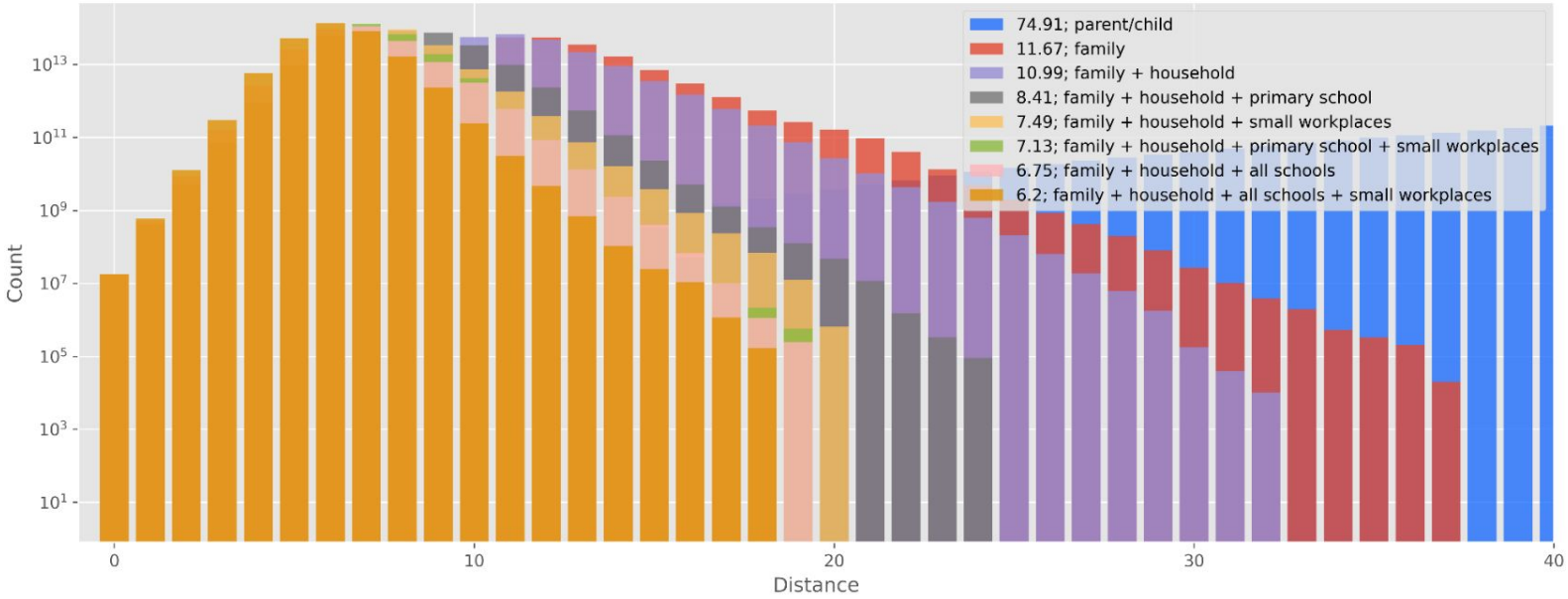


Households, work, school, neighbors

- **Household**: fully connected components of all people in the same house
- **School**: fully connected components of all people in the same class
  - Small classes: only primary school
  - All school levels
- **Work**: fully connected components of people with the same employer
  - Small workplaces: companies with less than 50 people employed
  - All workplaces
- **Neighbors**: connections to individuals in top-*k* closest households

How do each of these layers contribute to the small-world aspect?

Small-world population-scale social networks



6 to 7 degrees of (formal social tie) separation

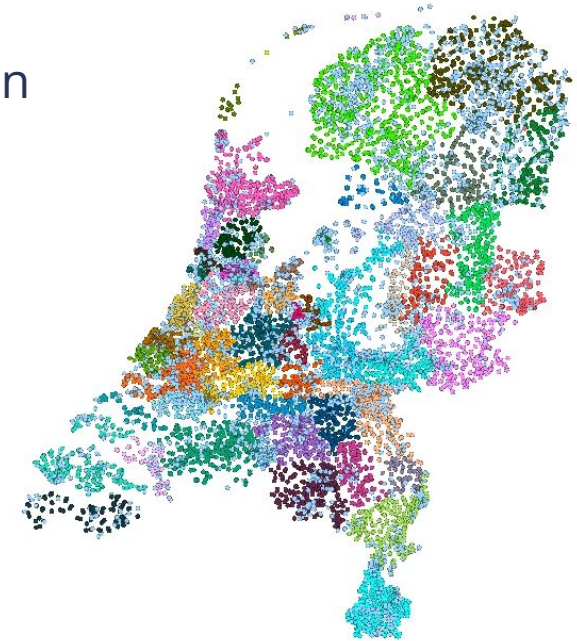
	nodes	edges	components	% nodes in giant comp.	average clust. coef.	diam.	average distance
parent	15.76M	19.16M	963.75K	0.58	0.000	288	74.91
family	16.44M	135.10M	318.02K	0.92	0.701	42	11.67
family + household	16.80M	137.12M	317.30K	0.93	0.663	38	10.99
family + household + primary school	16.80M	164.84M	227.00K	0.96	0.674	29	8.41
family + household + small workplaces	16.85M	174.98M	218.35K	0.96	0.674	27	7.49
family + household + primary school + small workplaces	16.86M	202.69M	157.69K	0.97	0.664	25	7.13
family + household + all schools	16.85M	253.97M	163.64K	0.97	0.673	24	6.75
family + household + all schools + small workplaces	16.89M	291.80M	118.86K	0.98	0.665	23	6.20
family + household + all schools + small workplaces + neighbors	17.25M	491.58M	1.67K	0.99	0.445	30	5.21

- Local clustering from (second degree) family and household relations
- Distant links from school, work and (spurious?) neighbor relations

(clustering coefficient approximated using a 1000 node sample, diameter computed exactly using `teexgraph`)

Ongoing POPNET projects

- Community-driven measures of social segregation
- Network-driven inequality measures
- Multi-layer family network analysis
- Atlas of social capital
- Offspring mobility
- Survey-based link validation
- Network-driven official statistics
- Measuring anonymity in networks
- Ethics in population-scale network analysis



Conclusion and outlook

- **Population-scale social network analysis** requires a critical reconsideration of the fundamental unit of analysis, the effect of measurement errors and the boundary specification problem
- Depending on the research goal, a careful analysis of the link source (**formal** vs. **informal**) and link type (**implicit** vs. **explicit**) is required
- Even with just relatively explicit formal ties, the population-scale social network of the Netherlands exhibits a **small-world structure**
- The **POPNET** project has an exciting time ahead :-)

Thank you!

POPNET

Frank Takes

✉ takes@liacs.nl

https://franktakes.nl

@franktakes

POPNET project

✉ popnet@uva.nl

https://popnet.io

@popnet_research

Thoughts / ideas / suggestions?
Please reach out!

- Biweekly "**POPNET** *Connects*" seminars (online & free!)
- Want to learn more? Subscribe to our mailistinglist: https://popnet.io/staytuned

Slide/image credits:
Hanjo Boekhout, Eszter Bokányi, Eelke Heemskerk, Yuliia Kazmina