Population-scale Social Network Analysis: Advances and Opportunities

Frank W. Takes[0000–0001–5468–1030]

1 Introduction

Social networks are generally seen as an interesting object of study from both the perspective of the social sciences [26] as well as of the natural sciences [1, 23]. The advent of computational social science [20] marked the joining of forces of these two disciplines, among other things in search for new network analysis methods and mining algorithms to address pressing social scientific questions. The relatively recent availability of digital trace data sparked a number of studies focused on the analysis of large-scale social networks. Very frequently, these social networks are obtained from social media data [21] or telecommunication data [3]; oftentimes sourced from organizations that operate the infrastructure behind these networks through data sharing agreements or by means of scraping or sampling. While interesting due to their large size, these networks are inherently subject to bias in terms of who is included. For example, the set of users of a social media platform is not necessarily a uniformly random sample of the population [24]. Moreover, the connections represented in such datasets are not per se an accurate display of one's social circles, and might involve spurious connections that do not represent meaningful ties in the real world [7]. In other words, the so-called *boundary specification problem* [19] depicted in Figure 1 (left) plays its parts in large-scale social network data. This means that it is often unclear according to what inclusion criteria a dataset was sampled. Even more so, there is the risk that the commercial organizations behind the social media platforms unexpectedly change their data access policy, resulting in sudden unavailability of such social network data for research purposes [29].

A recent line of work has proposed to create networks from official government-curated register data at the level of a complete population, e.g., an entire country.

Frank W. Takes
Department of Computer Science (LIACS), Leiden University, e-mail: takes@liacs.nl

Author preprint August 31, 2024

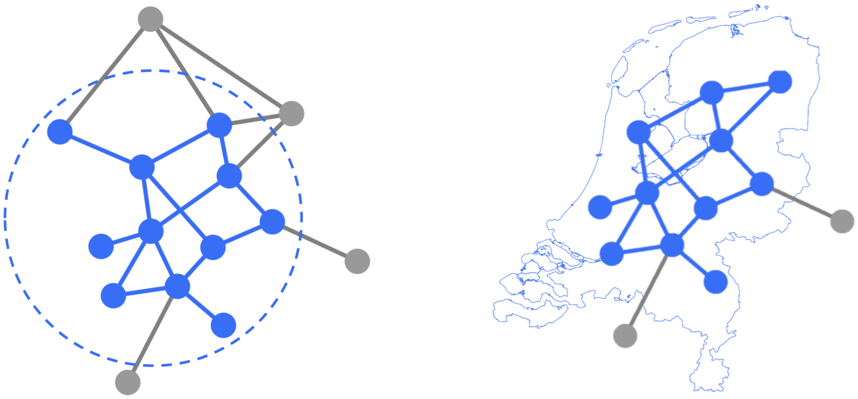

Fig. 1 Boundary specification. Sampling-induced (left) vs. Population-induced boundary (right). Blue nodes are in the sample, grey ones are not. The rightmost outline is that of the Netherlands.

Because such administrative data is based on population registers, it is typically highly complete. This means that the delineation of the set of *nodes* is precisely defined by those people present in the population under consideration. In Figure 1 (right), this situation is contrasted to the aforementioned sampling-induced boundary. In a population-scale network, the set of *edges* is not a collection of self-reported ties, but rather stems from official register data and, e.g., tax filings. For example, birth and marriage registers can be used to derive family connections. Income tax data can be used to obtain employment information, enabling the derivation of an individual's colleagues. In a similar vein, schoolmates can be identified from central school administrations. Housing registers can in turn, when combined with geospatial data, be used to derive household and next-door neighbor ties. The resulting networks are known under different terms, with the most frequently used ones being "population-scale social network"[1] and "nation-scale social network"[2]. The creation of such network data at the country level is typically done by national statistics institutes [2, 30] that are able to link, in a responsible and privacy-preserving manner, data on individuals originating from different governmental data sources.

The remainder of the text starts with a description of the main concepts and terminology related to population-scale social networks in Section 2. We then give an overview of some of the most important findings from the field in Section 3, ending the chapter in Section 4 with an outlook in terms of what opportunities this type of social network data offers researchers in the years to come.

[1] See, e.g., the POPNET project (2020–2025) in the Netherlands: https://www.popnet.io

[2] See, e.g., the Nation-scale Social Networks project (2020–2025) in Denmark: https://sodas.ku.dk/projects/nation-scale-social-networks

2 Population-scale Social Networks

A *population-scale social network* is defined as a graph consisting of:

- a set of nodes representing people in a clearly demarcated population, and
- a set of edges originating from one or more well-defined contexts.

The most common and simple demarcation of a population of people, i.e., the node set, is a country. Data on the place of residence of people can be used to study precise sub-populations of, e.g., provinces, municipalities, or neighborhoods. In addition, demographic data on for example people's age, education, socio-economic status and race can provide contextual information on the individuals in the data in the form of node attributes. Unique about population-scale social networks in terms of data quality is that the set of nodes is often highly *complete*. Concretely, there is no question of sampling bias or error resulting from for example entity resolution problems. This is typically because the data creation process utilizes government-issued identifiers from centralized population registers, that are per definition unique.

When the set of edges stems from multiple different contexts, e.g., different government registers on employment and family of people, a *multi-layer network* representation [17] in which each layer defines the precise context from which the link stems, is often used. An example is given in Figure 2, visualizing one person's ego network (red center node), featuring five different (partially overlapping) contexts

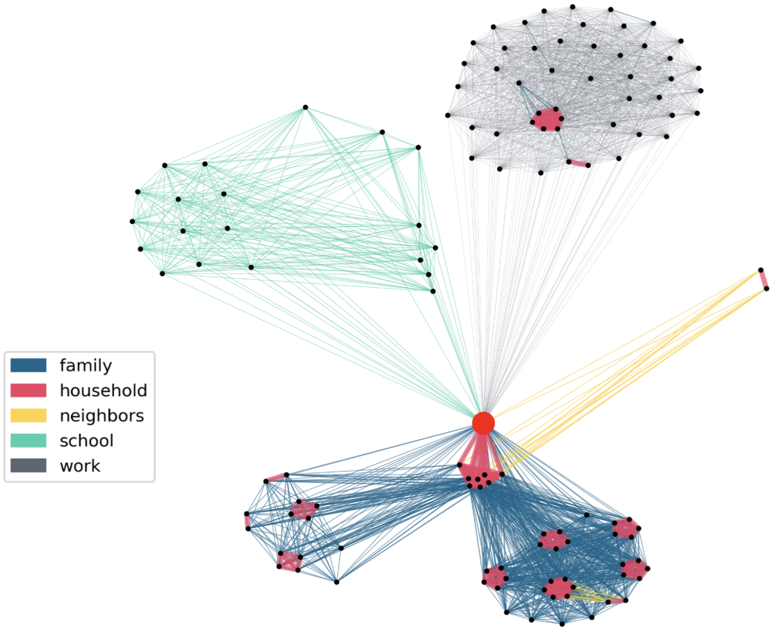

Fig. 2 Ego network of center red node in a 5-layer population-scale social network.

(layers). The set of edges is equally *complete*: for each context represented, the official underlying data source is typically available for the population as a whole. For example, in a layer of work connections, precisely all colleagues associated with the same organization as the ego's employer, are represented. Similarly, a layer of family connections contains all family up to a certain degree of kinship.

One very substantial caveat in all of the above, lies in the word "complete". Given the definition of how the network was constructed from official government registers, the data itself is indeed highly complete. However, this does not say anything about the use of that particular type of connection for a downstream task utilizing the network data. For each context (layer), connections may or may not be meaningful in modelling a particular network process. More generally, it is uncertain whether for a given ego node, the connections in the various layers represent the *true* social circles of that node. Certain layers may *oversample* the number of meaningful connections. This may happen, e.g., when all classmates are used as a proxy for someone's friendship network. On the other hand, only utilizing register-based data may in fact *undersample* friendships acquired via, for example, informal sports clubs.

Be it epidemic spread, knowledge diffusion, social influence or the plain measurement of social capital (for more examples, see Section 3); which links of the population-scale network are "active" or "meaningful" may differ depending on the studied problem. When this is approached as a problem of *data quality*; then it is actually about the *accuracy* of the links present in the data. Here it can be useful to, loosely inspired by [6, 18], distinguish between formal and informal ties. We say that *formal ties* represent connectivity of individuals originating from co-affiliation within a well-defined context and accompanying data source (e.g., business registers or family archives), over which the individual itself has limited control. These are typically the links present in a population-scale social network. In turn, *informal ties* represent relationships caused, created and/or to some extent controlled by the individual(s) involved in that particular tie. Clearly, formal and informal ties are not mutually exclusive: they may overlap, some formal ties may not be reported on by an individual, and the other way around, formal ties may not capture all meaningful social connections, i.e., the informal ties.

Depending on what downstream task is performed on the network, a population-scale social network may thus undersample or oversample certain types of connections. A particularly useful more social-scientifically theoretical way of approaching the links in the population-scale social network, is that of the *social opportunity structure* [5]. Each layer of the network can be seen as a particular context, i.e., opportunity structure, in which the individual is potentially exposed to a certain group of people (e.g., neighbors, colleagues or classmates). This conceptualization has been used in a number of studies, on which the section below elaborates further.

3 Advances

The structure of population-scale social networks has been shown to exhibit universal structural properties also observed in other types of social networks, such as small-world characteristics, high clustering and a skewed degree distribution, as shown in [5]. This paper also shows how the multi-layer structure enables an understanding of how people's connections in different contexts, such as family, work, and neighborhoods, evolve over an individual's life course based on factors like age and socioeconomic status. Population-scale network data has also been compared to online social networks. The work presented in [22] focuses on doing so for the Netherlands, comparing its population-scale social network with that of the country's at that time most dominant online social networking website. Results of applying community detection to these networks revealed historical and sociocultural subgroups. Perhaps equally importantly, the work showed that both types of networks capture similar connectivity patterns, suggesting that aforementioned oversampling and undersampling problems might be equally persistent in both online and register-based population-scale social networks. Whereas above referenced comparison was done using aggregated data, a number of studies utilizes individual level data. In such data, the importance of ensuring privacy of individuals in the population-scale social network has sparked a line of research on measuring and ensuring anonymity in networks [10, 11]. Findings in these works suggest that straightforward knowledge of distant connections significantly reduces overall network anonymity. These insights can be used for building more robust privacy-preserving data sharing and data publication techniques for (population-scale) social network data.

As suggested in [13], the potential of population-scale social network analysis lies not merely in analyzing network structure. This type of data also presents a novel way to study many relevant research topics of interest to social scientists. For example, population-scale social network analysis been instrumental in examining socio-economic segregation [31, 32]. Recent work showed that segregation is often more pronounced when measured in people's social networks, as compared to when measured in the traditional way, being within spatial neighborhoods [16]. Social capital studies have also benefited from this type of analysis, allowing for more precise measurements of bonding and bridging capital within a population [12]. Additionally, population-scale social network analysis has been used to explore people's perceptions on immigration, demonstrating that both direct contacts and broader social networks shape these views, with tipping points that depend on the extent of exposure [15]. Moreover, population-scale network data has been used to examine how social context influences the formation of close intergroup ties between immigrants and natives [28]. It turns out that opportunities for intergroup interaction in private contexts significantly increase the likelihood of forming these ties, with natives who have immigrant parents serving as key brokers. Beyond studies rooted in sociology, socio-economics, and political science, population-scale social network analysis has been applied in the field of family studies to better understand kinship networks [33]. In this domain, a recent work confirmed the generalizability of particular demographic patterns of interest across different country's populations [9].

Population-scale social networks have furthermore been used in the domain of public health, in particular in epidemic spread modelling, where network centrality measures were used to predict infection risks during the COVID-19 pandemic [14].

Overall, the studies referenced above collectively highlight the broad applicability and impact of population-scale social network analysis across diverse fields, from network science, to various strands of the social sciences, to public health.

4 Opportunities

Advances in methods for large-scale network analysis [8] combined with the increasing availability of register data for research purposes [27] has made it possible for population-scale social network analysis to establish itself as a core topic in the domain of computational social science. This has made it possible to study contemporary social scientific problems making use of data at unprecedented scale, providing unique avenues for methodological advances in the field of network science and social network analysis. The magnitude of population-scale data together with rich features on both the nodes and edges presents plenty of directions for development of data processing techniques and efficient network analysis models and algorithms. More specifically, the multi-layer large-scale structure of population-scale social networks has the potential to co-evolve with recent methodological advances in higher-order network models [4], which may help to better capture the complex group structures in population-scale network data.

The most logical yet extremely promising next step forward in population-scale social network analysis is that of longitudinal analyses of the network structure. This allows the field to move from insights related to people's position in the social network at a given point in time, to how real-world phenomena relate to how this network position evolves over time. Studying this at macro scale enables insights into the dynamics of an evolving population. First steps in this regard are to some extent taken in [25], which focuses on the evolution and predictability of human lives based on detailed event sequences. The work suggests how this type of work may allow researchers to "discover potential mechanisms that impact life outcomes as well as the associated possibilities for personalized interventions." Indeed, longitudinal network data is also a first and necessary step to deriving causal insights about the relation between social network structure and contextually meaningful outcome variables.

The future development of this field is dependent on continued intensive synergistic collaboration between researchers and national statistics institutes. This way, reusable and interoperable research infrastructure can be developed, enabling scientific research at population-scale with verifiable and reproducible results, while respecting legal and privacy-related guidelines. Moreover, it requires the availability of co-located population-scale data and compute infrastructure, integrating large-scale high performance computing infrastructure and secure government research data access environments.

In the years to come, scholars should seek to even further integrate this line of research into the broader field of computational social science. It is abundantly clear that population-scale social network data offers decades of research opportunities for both fundamental network science research as well as applied research on the most pressing social scientific challenges that our society will face in the years to come.

Index

References

1. A.-L. Barabási. *Network Science*. Cambridge University Press, 2016.
2. A. Bjerre-Nielsen, J. Cremers, J. Einsiedler, F. Christensen, S. N. Eriksen, S. B. A. Kohler, and L. H. Mortensen. Dataset profile: A Danish nation-scale network. *SocArXiv preprint 9wsx7*, 2023.
3. V. D. Blondel, A. Decuyper, and G. Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):10, 2015.
4. S. Boccaletti, P. De Lellis, C. Del Genio, K. Alfaro-Bittner, R. Criado, S. Jalan, and M. Romance. The structure and dynamics of networks with higher order interactions. *Physics Reports*, 1018:1–64, 2023.
5. E. Bokányi, E. M. Heemskerk, and F. W. Takes. The anatomy of a population-scale social network. *Scientific Reports*, 13(1):9209, 2023.
6. A. T. Cobb. Informal influence in the formal organization: Perceived sources or power among work unit peers. *Academy of Management Journal*, 23(1):155–161, 1980.
7. R. Corten. Composition and Structure of a Large Online Social Network in the Netherlands. *PLOS ONE*, 7(4):e34760, 2012.
8. M. Coscia. The atlas for the aspiring network scientist. *arXiv preprint 2101.00863*, 2021.
9. V. de Bel, E. Bokányi, K. Hank, and T. Leopold. A parallel kinship universe? using dutch kinship network data to replicate kolk et al.'s (2023) demographic account of kinship networks in sweden. *SocArXiv preprint 3k6nq*, 2024.
10. R. G. de Jong, M. P. van der Loo, and F. W. Takes. The effect of distant connections on node anonymity in complex networks. *Scientific Reports*, 14(1):1156, 2024.
11. R. G. de Jong, M. P. van der Loo, and F. W. Takes. A systematic comparison of measures for k-anonymity in networks. *arXiv preprint 2407.02290*, 2024.
12. B. de Zoete. Measuring community social capital through the structure of a population-scale social network. *MSc thesis, Leiden University*, 2022.
13. A. Espinosa-Rada and F. O. Ruiz. Why is the world small? comment on how the popnet project can leverage computational social science. *SocArXiv preprint m4srd*, 2023.
14. C. Hedde-von Westernhagen, A. Bagheri, and J. Garcia-Bernardo. Predicting covid-19 infections using multi-layer centrality measures in population-scale networks. *Applied Network Science*, 9(1):27, 2024.
15. Y. Kazmina, E. M. Heemskerk, E. Bokányi, and F. W. Takes. From contact to threat: A social network perspective on perceptions of immigration. *arXiv preprint 2407.06820*, 2024.
16. Y. Kazmina, E. M. Heemskerk, E. Bokányi, and F. W. Takes. Socio-economic segregation in a population-scale social network. *Social Networks*, 78:279–291, 2024.
17. M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.
18. H. J. Krackhardt D. Informal networks: The company behind the chart. *Harvard Business Review*, 71(4):104–111, 1993.
19. E. O. Laumann, P. V. Marsden, and D. Prensky. The boundary specification problem in network analysis. *Research Methods in Social Network Analysis*, 61(8), 1989.
20. D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.
21. J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceeding of the 17th ACM International Conference on World Wide Web (WWW)*, pages 915–924, 2008.
22. M. Menyhért, E. Bokányi, R. Corten, E. M. Heemskerk, Y. Kazmina, and F. W. Takes. Connectivity and community structure of online and register-based social networks. *arXiv preprint 2406.17752*, 2024.
23. M. Newman. *Networks*. Oxford University Press, 2018.

24. J. Pfeffer, A. Mooseder, J. Lasser, L. Hammer, O. Stritzel, and D. Garcia. This sample seems to be good enough! assessing coverage and temporal reliability of Twitter's academic API. In *Proceedings of the 17th AAAI International Conference on Web and Social Media (ICWSM)*, volume 17, pages 720–729, 2023.
25. G. Savcisens, T. Eliassi-Rad, L. K. Hansen, L. H. Mortensen, L. Lilleholt, A. Rogers, I. Zettler, and S. Lehmann. Using sequences of life-events to predict human lives. *Nature Computational Science*, 4(1):43–56, 2024.
26. J. Scott. *Social Network Analysis*. SAGE, 2017.
27. M. L. Small. The data revolution and the study of social inequality: Promise and perils. *Social Research: An International Quarterly*, 90(4):757–780, 2023.
28. N. Soler, E. Heemskerk, and Y. Kazmina. Contacts in contexts. *SocArxiv preprint axumt*, 2023.
29. Twitter Developers. *Starting February 9, we will no longer support free access to the Twitter API.* https://x.com/XDevelopers/status/1621026986784337922, 2024.
30. D. J. van der Laan. A person network of the Netherlands. Discussion paper, Statistics Netherlands, 2022.
31. D. J. van der Laan and E. de Jonge. Measuring segregation using a network of the Dutch population. In *Proceedings of the 5th International Conference on Computational Social Science (IC2S2)*, 2019.
32. D. J. van der Laan and E. de Jonge. Measuring local assortativity in the presence of missing values. In *Proceedings of the 8th International Conference on Complex Networks and Their Applications*, pages 280–290, 2020.
33. D. van Wijk. From prosperity to parenthood: How employment, income, and perceived economic uncertainty influence family formation. *PhD thesis, University of Groningen*, 2023.