Annals of Epidemiology 102 (2025) 28–35

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.sciencedirect.com/journal/annals-of-epidemiology

Invited commentary

Considerations for social networks and health data sharing: An overview

Dana K. Pasquale [a,b,*,1], Tom Wolff [b,c,1], Gabriel Varela [b,d], jimi adams [e], Peter J. Mucha [f], Brea L. Perry [g,h], Thomas W. Valente [i], James Moody [b,d]

[a] Department of Population Health Sciences, Duke University, Durham, NC, USA
[b] Duke Network Analysis Center, Duke University, Durham, NC, USA
[c] Medical Social Sciences, Northwestern University, Evanston, IL, USA
[d] Department of Sociology, Duke University, Durham, NC, USA
[e] Department of Sociology, University of South Carolina, Columbia, SC, USA
[f] Department of Mathematics, Dartmouth College, Hanover, NH, USA
[g] Department of Sociology, Indiana University, Bloomington, IN, USA
[h] Irsay Institute for Sociomedical Sciences, Indiana University, Bloomington, IN, USA
[i] Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

ARTICLE INFO

Keywords
Data sharing
Network data
Reproducibility of Results
Secondary Data Analysis
Data security
Privacy

ABSTRACT

The use of network analysis as a tool has increased exponentially as more clinical researchers see the benefits of network data for modeling of infectious disease transmission or translational activities in a variety of areas, including patient-caregiving teams, provider networks, patient-support networks, and adoption of health behaviors or treatments, to name a few. Yet, relational data such as network data carry a higher risk of deductive disclosure. Cases of reidentification have occurred and this is expected to become more common as computational ability increases. Recent data sharing policies aim to promote reproducibility, support replicability, and protect federal investment in the effort to collect these research data by making them available for secondary analyses. However, typical practices to protect individual-level clinical research data may not be sufficiently protective of participant privacy in the case of network data, nor in some cases do they permit secondary data analysis. When sharing data, researchers must balance *security, accessibility, reproducibility,* and *adaptability* (suitability for secondary analyses). Here, we provide background about applying network analysis to health and clinical research, describe the pros and cons of applying typical practices for sharing clinical data to network data, and provide recommendations for sharing network data.

Introduction

Over the last 40 + years, network analysis has established itself as a prominent data-intensive area of research. [1–3] Federal research investment in the field has grown dramatically: in 2022, the National Science Foundation (NSF) invested nearly 26 million dollars in 74 projects covering social, mathematical, statistical and related programs using network analysis. During this same period, the National Institutes of Health (NIH) invested over 93 million dollars in 209 projects (Fig. 1). Similar investment trends can be seen at the Centers for Disease Control and Prevention (CDC) and US Department of Defense (DoD) [for security in particular, see the Future Directions of Network Science report [4]].

Network analysis studies the relations between entities, often people, organizations, or political entities, and how those relations affect behaviors and outcomes. [5] Relationships between people in a network often include but are not limited to social, sexual, familial, or economic ties. [6,7] For example, a common network study would proceed from data that asked students in schools to name their closest friends. [8–11] Network analysis has produced novel and significant insights in fields such as health sciences, [5,12–16] social and economic modeling, [17] environmental modeling, [18,19] security studies, [20,21] and education [22] among others. For example, patient networks influence disease contagion, [23] physician networks impact care quality and prescribing behavior, [24–27] and community healthcare networks shape care delivery, [28] which in turn impact health disparities, [29] health analytics, and precision medicine. Social networks shape markets (both

* Corresponding auhtor at: Department of Population Health Sciences, Duke University, Durham, NC, USA.
E-mail address: dana.pasquale@duke.edu (D.K. Pasquale).
[1] These authors contributed equally

https://doi.org/10.1016/j.annepidem.2024.12.014
Received 24 October 2024; Received in revised form 20 December 2024; Accepted 28 December 2024
Available online 30 December 2024

hiring and performance) [30] and propel innovation (ideas and investments) [31] as well as its diffusion. [32–35] Education and innovation hinge on cross-linked social networks of scientists, [36] teacher networks support school performance, [37] and student networks affect each other's academic performance. [38–41] Network methods have also been central to major agenda-setting government reports, such as the U.S. Surgeon General's 2023 advisory on the healing effects of social connection. [42]

Effective January 25, 2023, NIH's new Data Management and Sharing (DMS) policy emphasizes data sharing and requires a DMS plan for all submissions for funding, [43] as part of NIH's increased focus on rigor and reproducibility. The DMS plan describes the scientific data generated from NIH funding, its management, and a plan for sharing the data with other researchers. As of December 2024, none of the sample DMS plans on NIH's DMS page pertain to network data. [44]

Yet, network data present uniquely problematic cases for sharing. By definition, network data is relational, with a correspondingly distinctive format consisting of an individual-level dataset describing each study member's relevant characteristics and measures along with a relational dataset describing the connections between study population members. *Egocentric* networks describe relationships between a focal respondent (called the *ego*) and their named contacts (called *alters*), alter and relationship characteristics, and alter-alter relations, [45] representing the linkages within the ego's immediate personal network (e.g., caregiving, social support, or needle-sharing networks, etc.). *Sociocentric* networks describe the relationships between all actors (called *nodes*) in a given context of interest (classroom, hospital, city, sexual network, etc.). In both types of network data, characteristics of the actors and their connections are typically collected, in addition to information describing the nature of the relationships. [46,47]

Consequently, there is a high risk for deductive disclosure of confidential information with network data, as the relational information makes it easier to reidentify participants. [47,48] Deductive disclosure occurs whenever one can reidentify unique individuals based on attributes, [49] even if explicit identifiers are removed. In the case of network data, a banal set of attributes which may not be sufficient to identify any single individual become more distinguishing once the relations between individuals are known, such as household members, romantic partners, or co-workers, and this risk increases with each additional person and their relationship linked into the egocentric or sociocentric network. Re-identification of people in network datasets has been shown to be possible, [48] potentially revealing private attributes such as income, sexual identity, or drug use. Sociocentric network studies pose deductive disclosure risk because the analytic sample is a census of the relevant context, meaning that anyone with some knowledge of setting actors can use that knowledge to identify study members and then learn any confidential information included in the survey. Ego-network data has the advantage of sampling from populations (leading to inclusion ambiguity), but often includes clusters of close contacts with identifying attributes since each respondent provides detailed information about their alters and their connections. These uniquely elevated risks require enhanced data security when sharing network data.

The risk of deductive disclosure for relational data is compounded by another feature of network data – that subjects provide both their own data and data about the relationships (edges) elicited from them, often

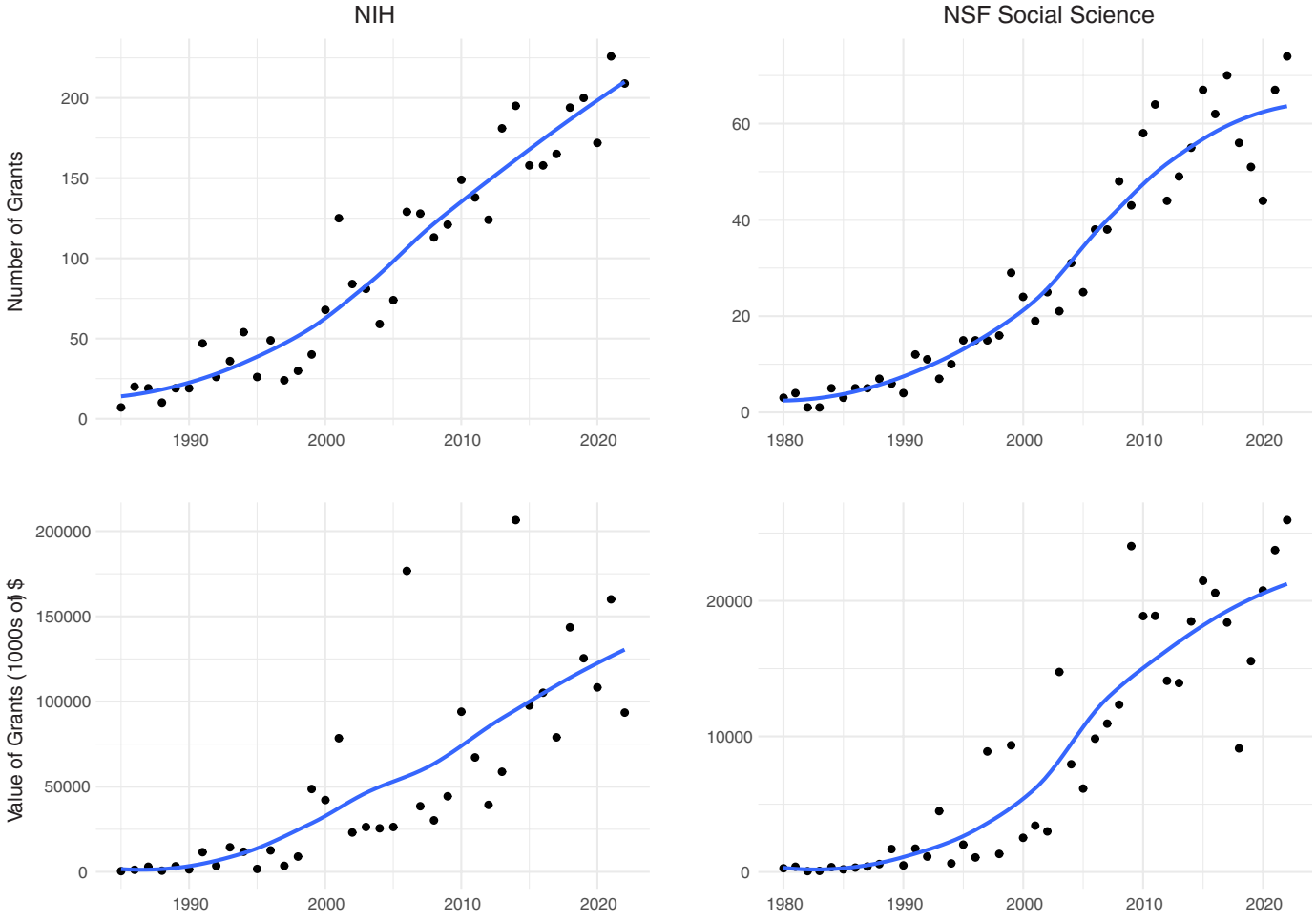

Fig. 1. illustrates grant funders ongoing interest and investment in projects that include network data and analysis.

supplying information about people who may not have consented to their participation in the study, including various types of potentially identifying information such as names or initials. This proxy reporting process is a necessary feature of sociocentric and longitudinal egocentric data collection, where it is essential to be able to link or match individuals across participants or across time. As such, the typical individual-oriented consent procedures for collecting, storing, and sharing data must evolve to meet these unique cases.

As issues with re-identification of network members have largely occurred in computational and social science spaces, clinical scientists have yet to thoroughly grapple with it. However, this is now a pressing issue in light of NIH's new data sharing policy, as more clinical researchers see the benefits of network data for translational activities in a variety of areas.

As a community of researchers, we have to balance potentially competing ethical duties. On one hand, it is imperative to protect the privacy of our research participants, who provide intimate information about private health, economic, and social behaviors. On the other hand, we have an obligation to scientific integrity and inquiry that is best served by open and transparent analytic practice. [50,51] This generally requires making data available to other researchers for independent inspection. As we describe in the next section, many traditional practices for protecting privacy while openly sharing research data cannot be directly applied to network data without compromising its utility or increasing the risk of privacy violations.

To address these challenges, we review common data sharing practices and their applicability when sharing network data and suggest that the best strategy for ethically sharing sensitive network data requires a multi-level access approach that leverages the utility-privacy tradeoff inherent in identifiability. At one extreme, we provide guidelines for minimal data sharing that allows simple analyses but hides much of the attribute data that poses re-identification risk. At the other extreme, full data can be shared, enhancing replicability and secondary data analysis but increasing risk of privacy loss for participants and the individuals about whom proxy data were provided. Several proposed techniques aim to balance privacy and data utility, though determining a standard practice is complicated by variation across networks. [52–55] Ideally, these techniques should balance *security* (protection of privacy), *accessibility* (openness to the other researchers), *reproducibility* (sharing to ensure good practices), and *adaptability* (a protection of the investment made to collect the data in such a way that secondary analyses can be conducted). However, balancing these four features can be complicated in practice.

Applicability of data sharing practices to network data

The aforementioned NIH policy targets analytical replicability, yet the resulting practices also facilitate sharing data to address questions beyond their original purpose. Common practices to ensure safe clinical data sharing practices come with different implications when sharing network data. [56] The clinical data sharing practices we cover include aggregate reporting, partial reporting, reducing data sensitivity, perturbing the data or adding noise, or sharing via a repository. We describe these practices, provide examples, discuss the advantages and disadvantages of applying these strategies to network data, and provide actionable recommendations to balance *security*, *accessibility*, *reproducibility*, and *adaptability* (SARA) when sharing network data.

Aggregated data

As with other data types, reporting aggregate network data – or summarized statistics about the network as a whole derived from information about relationships, or the individual nodes or alters – provides the highest level of *security* and *accessibility*. For networks, aggregated statistics often appear in research publications and provide an idea of the network's structure (such as the number of nodes and edges, density, connectivity, composition, centralization, or the modularity of communities / groups). [45,57] These statistics may also summarize node-level attributes across the network, such as the proportion of nodes with specific characteristics or differences in network statistics by group characteristics of interest (e.g., gender, race). [58,59] Aggregated statistics are commonly used in standard regression models or in building network-based simulations.

While commonly used, aggregated network statistics share disadvantages with other types of aggregate reporting. Analyses are not reproducible, and the richness of the data is lost, reducing potential for *adaptability*. [60] This is particularly important in sociocentric datasets that consist of a *single* network, where analyses of individual nodes are often required to understand how the behavior or outcomes of those nodes are influenced by the broader network structure and a person's position in it. [5] Such analyses are typically not possible using aggregated sociocentric network data. Thus, using aggregated sociocentric data for novel research questions generally requires simulation of networks tuned to the set of network statistics provided. [61–64] Moreover, given the probabilistic nature of simulations, full replication of published research findings using node-level data may not be possible.

In contrast, egocentric network analysis allows researchers to access hundreds or thousands of independent networks. [45] Thus, aggregated network statistics are more commonly used to simplify statistical modeling because much of the useful information about the networks, especially network structure, is contained in the aggregate. Nonetheless, certain kinds of egocentric analyses, such as multilevel modeling of the behavior of alters within ego networks, are not feasible with aggregated data. In some cases, broader "global" structures in which the egocentric networks exist can be simulated from the available information. [65,66]

Partial reporting

Partial reporting, or depositing a dataset with some individuals omitted, is another technique often used to reduce the risk of identifying participants and provide a structure to promote *accessibility*. Partial reporting is one of the recommendations for sharing genetic or pathogen genomic sequences [67] [a 2014 Genomic Data Sharing Policy requires sharing of human and non-human genomic data generated from NIH funding via the appropriate genomic data repository [68]]. The goal of partial reporting is to reduce the risk of deductive disclosure by leaving some of the potentially linked sample members out of the dataset. However, the suitability of this practice for network data is unclear because it may still be possible to identify egos and alters based on the combination of attributes among reported nodes in both egocentric and sociocentric networks. This is compounded for persons with rare nodal, relational, or dyad-level attributes.

While this strategy promotes high *accessibility*, it hinders the *reproducibility* and *adaptability* of network datasets where the strength is in identifying and measuring relationships between people. There is an inverse relationship between *security* and *adaptability/reproducibility* dependent on the amount of data reported; for network data in which the richness lies in the combination of attributes and relations, holding back enough data to enhance *security* may reduce *adaptability* or *reproducibility*. *Reproducibility* is particularly affected if overall network statistics or positions are skewed.

Partial reporting of network data can change the broader structure of sociocentric networks depending on which nodes or edges are kept out of the reported dataset (Figure 2). Removing nodes or edges which affect the degree distribution or motif distribution (dyad/triad census) affects *reproducibility* in sociocentric networks. [69–71] Furthermore, the possibility for partial reporting could selectively remove relational information, rather than "sampling" for deletion from the individuals; this difference in (multiple) units of analysis is one of the unique considerations to appropriately sharing *network* data.

In egocentric networks, removing alters can change the structural (e.g., network size or density) or compositional (e.g., proportion women,

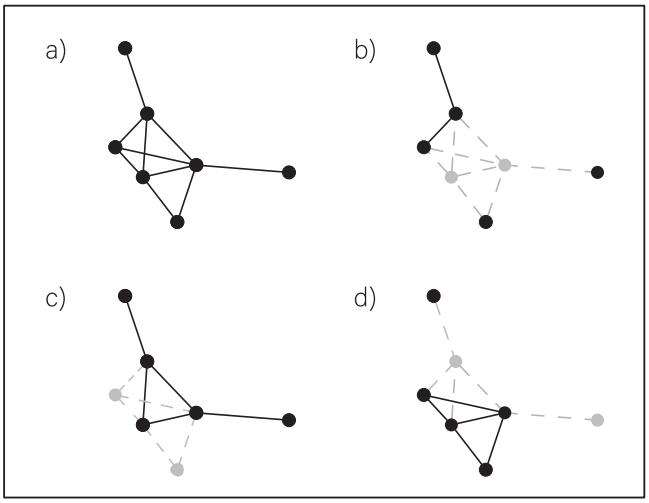

Fig. 2. Potential loss of relational information in partially reported data. Panel a) represents a network with full reporting of nodes and edges. Panels b), c), and d) illustrate the same network with two nodes randomly omitted in partial reporting. One sees that the removal of select nodes can lead to different measures and interpretations of the same network, with negative consequences for reproducibility.

mean tie strength) properties of networks, though this problem is minimized if the research adheres to proper procedures regarding randomization and number of alters dropped. [72] *Adaptability* is impeded if the relationships between nodes or if the broad network structure is altered. Additionally, partial reporting may not fully promote *security*. The risk of deductive disclosure is not minimized in the case of ego networks retained in the reported dataset if few of the alters are probabilistically selected to be left out of the shared data. Leaving out a sample of complete egocentric networks also does not protect the privacy of the remaining reported egocentric networks since egos are typically sampled from a population and thus independent.

Reducing data sensitivity

Reducing data sensitivity involves adjusting coding schemas, such as removing dates or changing continuous variables like age to categorical ones. There are similar ways that data can be categorized or generalized in ways to reduce sensitivity that operates at the relational level (e.g., illicit affairs) without acting on individual-level data sensitivity. The goal is to create a dataset with limited sensitivity that can still be used for statistical analyses. *Accessibility* remains unaffected so long as data are made publicly available or shared by the initial principal investigator or primary data collector (hereafter "primary user" as this is the person likely to be responsible for decisions about data sharing). Depending on the analysis, this strategy may still enable *reproducibility* — but often only for questions that have already been addressed or that require levels of measurement anticipated by the primary user. *Adaptability* is then inhibited, especially if new analytic questions rely on continuous comparison between subjects or finely detailed measurements. This strategy may also obscure some of the salient features driving connections, for instance age differences between sexual partners in a network [73,74] or designating relations as kin vs. non-kin in an egocentric network. [75] From a *security* standpoint, identification of network members may still be possible with rare outcomes since the addition of relations between people adds a new dimension through which to identify cohort members. Reducing data sensitivity may minimize the precision of measurements linked to specific individuals, but providing a coded or limited dataset can still put network members at risk of re-identification, especially when there are rare nodal, relational, or dyad-level attributes which can still lead to re-identification (e.g., relations across age bands).

Data perturbation

Data perturbation consists of adding "noise" to measurements in a dataset to reduce the risk of deductive disclosure. For network data, noise is added by changing a percentage of edges (relationships) [76] – for instance, moving a certain percentage of edges from one pair of nodes to another or within a partition [77] in ways that preserve estimates of network structure, thereby curtailing network re-identification and promoting *security*. As with reducing data sensitivity, *accessibility* depends on the willingness of primary users to make data publicly available.

In other fields which use relational data, such as geospatial analysis, introducing noise has been shown to be protective while adding little bias. [78] However, this approach may have negative consequences for network data depending on the extent of perturbation. Data perturbation can induce significant bias in network data specifically, although methods to improve network estimates after adding noise have been developed [79] and differential privacy approaches to balance privacy and utility can be applied at local or centralized levels. [52] Relatedly, it can negatively affect *reproducibility* if path structures, motif distributions, or degree distributions change. This is evident with sociocentric data depending on the goal of analysis: changes to path structures can affect diffusion, especially at the levels needed to protect privacy, and changes to the motif or degree distribution can affect network description. [80] Perturbing egocentric data – particularly dropping/adding edges between alters or perturbing alter attributes – can also affect *reproducibility*, especially when modeling the network. *Adaptability* may still be high in both cases depending on the magnitudes and types of changes and whether overall network structure is impacted.

An alternative approach to preserve *reproducibility* and promote *adaptability*, permuting node attributes, can reduce this bias, though it risks hiding these attributes' relationships to tie formation and other outcomes of interest. Moreover, data perturbation may still permit identification of nodes with rare attributes or relationships, or in a network with high population depending on the data access level. A perturbation method that increases *security* and supports *accessibility* and *adaptability* includes permuting some critical node-level attributes and applying an exponential random graph model to the permuted data to create a faux network based on the empirical data. [81,82] However, this method only reproduces past analyses insofar as the network generated resembles its empirical counterpart.

Data sharing via data repositories

Specialized data repositories are adapted to distinct formats and sensitivity risks. Examples include the genomic data repositories (Gene Expression Omnibus (GEO), [83] Sequence Read Archive (SRA), [84] GenBank, [85] Los Alamos National Laboratory (LANL), [86] etc.) listed in the 2014 NIH Genomic Data Sharing Policy as well as a set of generalist or domain-specific NIH-supported scientific data repositories adhering to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. [87] However, many of these repositories maximize *accessibility* by significantly limiting personal identifiers and, consequently, limit *adaptability* for secondary research. Existing repositories for network data include the Stanford Large Network Dataset Collection (SNAP), [88] Network Data Repository, [89] Colorado Index of Complex Networks (ICON), [90] and IEEE DataPort. [91] These repositories typically list important measures associated with each dataset to help users determine which one is appropriate for their needs. However, most network data sets in these repositories have limited personal information.

Some domain-specific repositories follow set processes to share fuller network data, opting to maximize *reproducibility* and *adaptability*. Commonly, prior to data deposit, the primary user can set criteria for

who can request the data and how they can be used, as with the Inter-university Consortium for Political and Social Research (ICPSR). [92] Institutional review board (IRB) approval is communicated to the repository owner and if the data requestors meet the requirements of the primary users then the data will be shared. Limitations of sharing full network data via a repository echo those for other types of data: it relies on the data requestor to protect the highly sensitive data shared. Full data sharing of many network data sets carries a significant *security* risk, and when repositories do prioritize *security*, *accessibility* becomes a function of whether there are processes in place to facilitate secure sharing.

Several network studies, including Add Health, share data in this manner through restricted-use contractual data sharing. [93] Such processes typically exist as formal agreements between the primary user and requestor, facilitated by the data requestor's IRB to approve secondary analyses of the data, provided that a data requestor fits criteria set by the primary user. IRB approval is communicated to the primary user before the dataset is shared with the requestor. This process puts the onus of *security* on the stringency of the IRB and leverages trust in ethical practices to share and protect data for analysis. Data use agreements (DUAs) are generated and stored by the primary user. This process is frequently manual rather than automated and thus inefficient and difficult to scale, which can be burdensome and inadvertently limit *accessibility*.

Other limitations of online repositories have implications for *accessibility* and *security* even when users follow best practices. Online data repositories are frequently hosted within a specific institution; if the institution changes investments in a way that threatens the continuity of the repository then *accessibility* is affected while the primary users re-submit data to an alternative repository. Another limitation emerges from the potential for credential counterfeiting or user credential phishing; these potential *security* concerns are distinctive to online settings and should be appropriately considered.

Despite these limitations, full datasets hosted in repositories may be the best solution for network data. Many older datasets now exist electronically, can be shared easily with no restriction on number of copies, and could exist in perpetuity. Accordingly, online repositories may be the most effective path to sharing secure and versatile network data provided *security* can be assured. Furthermore, as computing power increases, and as relational social media data becomes more ubiquitous, re-identification of individuals in these data may be possible via methods that did not exist when the data were collected, forcing a duty to consider the future ramifications of present data sharing. Thoughtfulness and care are needed regarding collecting, storing, and sharing these data regardless of their ability to foster *reproducibility* and *adaptability*. Such repositories could achieve this by following the practices of repositories for other types of data, [94,95] including stringent DUAs, relying on the ethical practices to which we as a community of researchers are bound. DUAs in this vein include restrictions on use to what is specified in the agreement and by whom, restrictions on sharing, a time-limited period of data use, instructions for both storage and destruction of data upon close of the agreed-upon analyses, and a requirement of IRB oversight. Ideally, electronic repositories will utilize an automated DUA with an honest data broker to ease the burden and facilitate *accessibility*.

Effects of data sharing practices by network type

Egocentric and sociocentric network data have distinct structure and analysis goals, leading to different considerations for sharing practices. Each practice varies in its ability to ensure the *security, accessibility, reproducibility,* and *adaptability* by network type. Figure 3 provides an overview of how these practices promote each of these qualities.

Considerations and recommendations specific to egocentric data. As the most accessible option, aggregated data at the ego network level can be provided with minimal risk to security. This includes summary statistics for the structural and compositional characteristics of each ego network. However, a DUA is recommended for sharing alter-level data to ensure that these more sensitive data, which carry a possibility of deductive disclosure for a person who has not consented, are not identifiable. Strategies to protect alters' confidentiality should always include removing any alter names, including first names or initials, and using only numeric identifiers to link alters across waves of data. Moreover, security can be further enhanced by either providing only aggregated information at the network level for alter demographics or reducing sensitivity by broadly categorizing or dichotomizing information (e.g., relationship as kin/non-kin, education as college/non-college). Exceptions to these alter-level data sharing guidelines arise when including egocentric network data collected in a small, knowable, or at-risk community could make individuals identifiable (e.g., drug use networks in a rural community); [96] in such cases, alter-level characteristics should not be provided regardless of a DUA.

Considerations and recommendations specific to sociocentric data. As with egocentric data, we recommend that sociocentric aggregated data and network statistics be made publicly available. [97] Restrictions are not needed unless the aggregated statistics are combined in such a way that a small or rare group could be identified. [98] Beyond DUA standards,

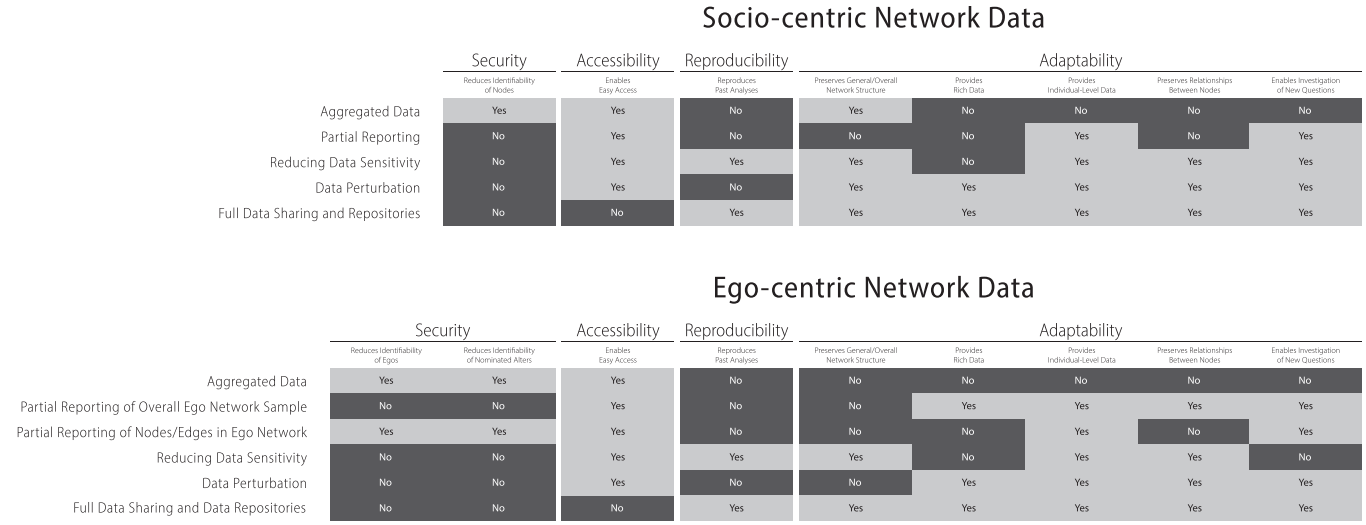

Fig. 3. Clinical data sharing practices and their efficacy in promoting *security, accessibility, reproducibility,* and *adaptability* for network data.

one must consider both the identification issues alongside the potential uses of shared data. While individual-level protections are important, given the "complete" population nature of sociocentric network studies, the implications of these considerations must be addressed at both the collective and individual level. To this end, sociocentric network data sharing must address protections for the metadata about the study context, protections of nodal attribute *and* relational information, as well as appropriate uses at the analytic level. With the exception of individual attribute data (which in the case of sociocentric data is typically self-provided), each of these require network-specific considerations beyond the "standard" operating procedures.

Conclusion

Network data have unique features that do not adhere to typical conceptualizations of data privacy associated with traditional (non-relational) data, resulting in a significantly higher risk of deductive disclosure. Best practices at a minimum should include protocols to protect privacy (*security*), clear guidelines for data depositors and requestors (*accessibility, reproducibility*), an automated process to ease sharing (also *accessibility*), and provision of the data for robust secondary analyses (*adaptability*). Maximizing *reproducibility* and *adaptability* protects investments and promotes open science. [99] Balancing all four principles requires a clear pipeline for data depositors to share their data in the fullest form possible depending on sensitivity. As with other types of clinical or public health data, a data repository can have different protocols to deposit and share network data based on sensitivity. Sharing network data is crucial for transparency in science and maximizing federal research investments. To achieve these goals, sharing data in its full form while protecting security and sensitive information is essential. To that end, a centralized data repository designed specifically for network data would provide a focal point for searching and sharing network data, protecting the investment made to collect the data.

Sources of financial support

D.K.P., T.W., G.V., and J.M. were supported by the U.S. National Science Foundation (NSF) award #2024271 made to Duke University and P.J.M. was supported by NSF award #2140024 made to Dartmouth University. J.M and G.V were partially supported by the U.S. National Institutes of Health (NIH) Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) award R25 HD079352 made to Duke University. B.L.P. was supported by the U.S. NIH National Institute on Aging award R01 AG057739 made to Indiana University.

CRediT authorship contribution statement

Tom Wolff: Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **Gabriel Varela:** Writing – review & editing, Writing – original draft, Conceptualization. **Brea L Perry:** Writing – review & editing, Writing – original draft. **Thomas W Valente:** Writing – review & editing, Writing – original draft. **jimi adams:** Writing – review & editing, Writing – original draft, Conceptualization. **Peter J Mucha:** Writing – review & editing, Writing – original draft, Funding acquisition. **James Moody:** Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization. **Dana Kristine Pasquale:** Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization.

Declarations of Interest

Authors D.K.P, T.W., G.V., P.J.M., and J.M are currently developing a new network data repository and analysis platform, IDEANet ideanet — The IntegratedIntegrating Data Exchange and Analysis for Networks (https://cran.r-project.org/package=ideanet, supported by NSF awards #2024271 and #2140024). The remaining authors have none to declare.

References

[1] Borgatti SP, Mehra A, Brass DJ, Labianca G. Network analysis in the social sciences. Science 2009;323:892–5. https://doi.org/10.1126/science.1165821.
[2] Butts CT. Revisiting the foundations of network analysis. Science 2009;325(5939): 414–6. https://doi.org/10.1126/science.1171022.
[3] Lazer D, Pentland A, Adamic L, Aral S, Barabasi A-L, Brewer D, et al. Life in the network: The coming age of computational social science. Science 2009;323 (5915):721–3. https://doi.org/10.1126/science.1167742.
[4] Coronges K, Barabasi A-L, Vespignani A. A workshop report on the emerging science of networks. *FDNS: Future Dir Netw Sci* 2016. Available from: ⟨https://basicresearch.defense.gov/Portals/61/Documents/future-directions/Network_Sciences.pdf⟩.
[5] Valente TW. Social networks and health: Models, methods, and applications. New York, NY: Oxford University Press, Inc; 2010. p. 277.
[6] Bearman PS, Moody J, Stovel K. Chains of affection: The structure of adolescent romantic networks. Am J Sociol 2004 (in press.).
[7] Peng Y. Kinship networks and entrepreneurs in china's transitional economy. Am J Sociol 2004;109(5):1045–74. https://doi.org/10.1086/382347.
[8] Coleman JS. The adolescent society: The social life of the teenager and its impact on education. Free Press of Glencoe; 1961.
[9] Shrum W, Cheek NHJ, Hunter SM. Friendship in school: Gender and racial homophily. Sociol Educ 1988;61(4):227–39.
[10] Copeland M, Alqahtani RT, Moody J, Curdy B, Alghamdi M, Alqurashi F. When friends bring you down: Peer stress proliferation and suicidality. Arch Suicide Res 2021;25(3):672–89. https://doi.org/10.1080/13811118.2020.1746939.
[11] Doehne M, McFarland DA, Moody J. Network ecology: Tie fitness in social context (s). Soc Netw 2024;76:174–90. https://doi.org/10.1016/j.socnet.2023.09.005.
[12] Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology 2016;41(7): 1691–6. https://doi.org/10.1038/npp.2016.7.
[13] Kirk JM, Kim SO, Inoue K, Smola MJ, Lee DM, Schertzer MD, et al. Functional classification of long non-coding RNAs by k-mer content. Nat Genet 2018;50: 1474–82. https://doi.org/10.1038/s41588-018-0207-8.
[14] Pasquale DK, Doherty IA, Sampson LA, Hue S, Leone PA, Sebastian J, et al. Leveraging phylogenetics to understand HIV transmission and partner notification networks. J Acquir Immune Defic Syndr 2018;78(4):367–75. https://doi.org/10.1097/QAI.0000000000001695.
[15] Robinson JI, Weir WH, Crowley JR, Hink T, Reske KA, Kwon JH, et al. Metabolomic networks connect host-microbiome processes to human clostridioides difficile infections. J Clin Invest 2019;130:3792–806. https://doi.org/10.1172/JCI126905.
[16] Pasquale DK, Doherty IA, Miller WC, Leone PA, Sampson LA, Ledford SL, et al. Factors associated with human immunodeficiency virus infections linked in genetic clusters but disconnected in partner tracing. Sex Transm Dis 2020;47(2):80–7. https://doi.org/10.1097/olq.0000000000001094.
[17] Keister LA. Exchange structures in transition: A longitudinal study of lending and trade relations in Chinese business groups. Am Sociol Rev 2001;66:336–60.
[18] Gomez JM, Nunn CL, Verdu M. Centrality in primate-parasite networks reveals the potential for the transmission of emerging infectious diseases to humans. Proc Natl Acad Sci USA 2013;110(19):7738–41. https://doi.org/10.1073/pnas.1220716110.
[19] Pilosof S, Morand S, Krasnov BR, Nunn CL. Potential parasite transmission in multi-host networks based on parasite sharing. PLoS One 2015;10(3):e0117909. https://doi.org/10.1371/journal.pone.0117909.
[20] Krebs VE. Mapping networks of terrorist cells. Connections 2002;24(3):43–52.
[21] Cranmer SJ, Menninga EJ, Mucha PJ. Kantian fractionalization predicts the conflict propensity of the international system. Proc Natl Acad Sci 2015;112(38): 11812–6. https://doi.org/10.1073/pnas.1509423112.
[22] Vest Ettekal A, Adams J, Teti DM, Cleveland HH, Rullison KL. Handbook of research methods in developmental science. Wiley; 2022. Applications of social network analysis in developmental science.
[23] Karkada UH, Adamic LA, Kahn JM, Iwashyna TJ. Limiting the spread of highly resistant hospital-acquired microorganisms via critical care transfers: A simulation study. Intensive Care Med 2011;37(10):1633–40. https://doi.org/10.1007/s00134-011-2341-y.
[24] Iyengar R, Van den Bulte C, Valente TW. Opinion leadership and social contagion in new product diffusion. Mark Sci 2011;30(2):195–212. https://doi.org/10.1287/mksc.1100.0566.
[25] Onnela JP, O'Malley AJ, Keating NL, Landon BE. Comparison of physician networks constructed from thresholded ties versus shared clinical episodes. Appl Netw Sci 2018;3(1):28. https://doi.org/10.1007/s41109-018-0084-1.
[26] Trogdon JG, Chang Y, Shai S, Mucha PJ, Kuo TM, Meyer AM, et al. Care coordination and multispecialty teams in the care of colorectal cancer patients. Med Care 2018;56(5):430–5. https://doi.org/10.1097/MLR.0000000000000906.
[27] Trogdon JG, Weir W, Shai H, S, Mucha PJ, Kuo TM, Meyer AM, et al. Comparing shared patient networks across payers. J Gen Intern Med 2019;34:2014–20. https://doi.org/10.1007/s11606-019-04978-9.
[28] Unnikrishnan KP, Patnaik D, Iwashyna TJ. Spatio-temporal structure of US critical care transfer network. AMIA Jt Summits Transl Sci Proc 2011;2011:74–8. PMC3248748.
[29] Schaefer DR, adams j. Coevol Netw Health Netw Sci 2017;5(3):249–56. https://doi.org/10.1017/nws.2017.24.

[30] Dubos R, Erickson BH. In: Lin N, Cook K, Burt RS, editors. *Good networks and good jobs: The value of social capital to employers and employees*, in *Social capital: Theory and research*. New York: Routledge; 2001. p. 333.

[31] Reed M. A study of social network effects on the stock market. J Behav Financ 2016;17(4):342–51. https://doi.org/10.1080/15427560.2016.1238371.

[32] Valente TW. Network models of the diffusion of innovations. Cresskill, NJ: Hampton Press; 1995.

[33] Valente TW, Davis RL. Accelerating the diffusion of innovations using opinion leaders. Ann Am Acad Political Soc Sci 1999;566:55–67.

[34] Valente TW, Fosados R. Diffusion of innovations and network segmentation: The part played by people in promoting health. Sex Transm Dis 2006;33(7):S23–31.

[35] Valente TW, Dyal SR, Chu KH, Wipfli H, Fujimoto K. Diffusion of innovations theory applied to global tobacco control treaty ratification. Soc Sci Med 2015;145:89–97. https://doi.org/10.1016/j.socscimed.2015.10.001.

[36] Moody J. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. Am Sociol Rev 2004;69(2):213–38.

[37] Carolan BV. Social network analysis and education: Theory, methods & applications. Thousand Oaks, CA: SAGE Publications Inc; 2014. https://doi.org/10.4135/9781452270104.

[38] Kreager DA, Rulison K, Moody J. Delinquency and the structure of adolescent peer groups. Criminology 2011;49(1):95–127. https://doi.org/10.1111/j.1745-9125.2010.00219.x.

[39] Moody J, Brynildsen WD, Osgood DW, Feinberg ME, Gest S. Popularity trajectories and substance use in early adolescence. Soc Netw 2011;33(2):101–12. https://doi.org/10.1016/j.socnet.2010.10.001.

[40] Copeland M, Fisher JC, Moody J, Feinberg ME. Different kinds of lonely: Dimensions of isolation and substance use in adolescence. J Youth Adolesc 2018;47(8):1755–70. https://doi.org/10.1007/s10964-018-0860-3.

[41] McCabe JM. Connecting in college: How friendship networks matter for academic and social success. Chicago: University of Chicago Press; 2016. https://doi.org/10.7208/9780226409665-003.

[42] *Office of the Surgeon General (OSG). Our epidemic of loneliness and isolation: The U.S. Surgeon general's advisory on the healing effects of social connection and community, in Publications and Reports of the Surgeon General, US Department of Health and Human Services*, Editor. 2023, US Department of Health and Human Services: Washington (DC).

[43] *National Institutes of Health. Final NIH policy for data management and sharing (notice number: Not-od-21-013)*, O.o.S. Policy, Editor. n.d., National Institutes of Health,: Bethesda, MD. Available from: ⟨https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.htm⟩⟨l⟩.

[44] *National Institutes of Health. Writing a data management & sharing plan: Applications for receipt dates on/after jan 25 2023*. Scientifc Data Sharing n.d. [cited 2024 20-Dec-2024]; Available from: ⟨https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-for-data-management-and-sharing/writing-a-data-management-and-sharing-plan#sample-plans⟩.

[45] Perry BL, Pescosolido BA, Borgatti SP. Egocentric network analysis: Foundations, methods, and models. New York, N.Y.: Cambridge University Press; 2018. https://doi.org/10.1017/9781316443255.

[46] Wasserman S, Faust K. Social network analysis: Methods and applications. Structural analysis in the social sciences. New York, NY: Cambridge University Press; 1994. p. 825.

[47] adams j. Gathering social network data. Quantitative applications in the social sciences. Vol. Book, 180. SAGE Publications, Inc; 2019.

[48] Narayanan, A. and V. Shmatikov. *De-anonymizing social networks*. arXiv, 2009: p. 0903.3276v1.

[49] Na L, Yang C, Lo CC, Zhao F, Fukuoka Y, Aswani A. Feasibility of reidentifying individuals in large national physical activity data sets from which protected health information has been removed with use of machine learning. JAMA Netw Open 2018;1(8):e186040. https://doi.org/10.1001/jamanetworkopen.2018.6040.

[50] Hansson MG, Lochmuller H, Riess O, Schaefer F, Orth M, Rubinstein Y, et al. The risk of re-identification versus the need to identify individuals in rare disease research. Eur J Hum Genet 2016;24(11):1553–8. https://doi.org/10.1038/ejhg.2016.52.

[51] Ohmann C, Banzi R, Canham S, Battaglia S, Matei M, Ariyo C, et al. Sharing and reuse of individual participant data from clinical trials: Principles and recommendations. BMJ Open 2017;7(12):e018647. https://doi.org/10.1136/bmjopen-2017-018647.

[52] Jiang H, Pei J, Yu D, Yu J, Gong B, Cheng X. Applications of differential privacy in social network analysis: A survey. IEEE Trans Knowl Data Eng 2021:1. https://doi.org/10.1109/tkde.2021.3073062.

[53] Boedihardjo M, Strohmer T, Vershynin R. Covariance's loss is privacy's gain: Computationally efficient, private and accurate synthetic data. Found Comput Math 2022. https://doi.org/10.1007/s10208-022-09591-7.

[54] Boedihardjo, M., T. Strohmer, and R. Vershynin. *Covariance loss, Szemerédi regularity, and differential privacy*. arXiv, 2023. 2301.02705. DOI: ⟨10.48550/arXiv.2301.02705⟩.

[55] Yazdanjue, N., H. Yazdanjouei, H. Gharoun, M.S. Khorshidi, M. Rakhshaninejad, and A.H. Gandomi. A comprehensive bibliometric analysis on social network anonymization: Current approaches and future directions. arXiv, 2023. 2307.13179. DOI: ⟨10.48550/arXiv.2307.13179⟩.

[56] Lubarsky B. Re-identification of "anonymized data". 1 Geo L Tech Rev 2017. ⟨https://perma.cc/86RR-JUFT⟩ (Available from).

[57] McCarty C, Lubbers MJ, Vacca R, Molina JL. Conducting personal network research: A practical guide. Methodology in the social sciences series. Guilford Publications; 2019.

[58] Bearman PS, Moody J, Stovel K. The Add Health network variable codebook (p). University of North Carolina at Chapel Hill; 1997.

[59] Carolina Population Center - University of North Carolina at Chapel Hill. *National longitudinal study of adolescent to adult health - wave i: Network variables*. 2001: Chapel Hill, NC. Available from: ⟨https://addhealth.cpc.unc.edu/wp-content/uploads/docs/restricted_use/School-Network-Data.zip⟩.

[60] Rawlings CM, Smith JA, Moody J, McFarland DA. Network analysis: Integrating social network theory, method, and application with r. 1 ed. Cambridge: Cambridge University Press; 2023. https://doi.org/10.1017/9781139794985.

[61] Badham J, Abbass H, Stocker R. Parameterization of keeling's network generation algorithm. Theor Popul Biol 2008;74(2):161–6. https://doi.org/10.1016/j.tpb.2008.06.002.

[62] adams j, Schaefer DR. How initial prevalence moderates network-based smoking change: Estimating contextual effects with stochastic actor-based models. J Health Soc Behav 2016;57(1):22–38. https://doi.org/10.1177/0022146515627848.

[63] Jenness SM, Goodreau SM, Morris M. Epimodel: An r package for mathematical modeling of infectious disease over networks. J Stat Softw 2018;84. https://doi.org/10.18637/jss.v084.i08.

[64] Valente TW, Vega Yon GG. Diffusion/contagion processes on social networks. Health Educ Behav 2020;47(2):235–48. https://doi.org/10.1177/1090198120901497.

[65] Smith JA. Macrostructure from microstructure: Generating whole systems from ego networks. Socio Method 2012;42(1):155–205. https://doi.org/10.1177/0081175012455628.

[66] Krivitsky PN, Morris M. Inference for social network models from egocentrically sampled data, with application to understanding persistent racial disparities in HIV prevalence in the US. Ann Appl Stat 2017;11(1):427–55. https://doi.org/10.1214/16-AOAS1010.

[67] Dawson L, Benbow N, Fletcher FE, Kassaye S, Killelea A, Latham SR, et al. Addressing ethical challenges in US-based HIV phylogenetic research. J Infect Dis 2020;222(12):1997–2006. https://doi.org/10.1093/infdis/jiaa107.

[68] National Institutes of Health. *NIH genomic data sharing policy (notice number: Not-od-14-124)*, O.o.S. Policy, Editor. 2014, National Institutes of Health,: Bethesda, MD. Available from: ⟨https://grants.nih.gov/grants/guide/notice-files/not-od-14-124.html⟩.

[69] Smith JA, Moody J. Structural effects of network sampling coverage i: Nodes missing at random. Soc Netw 2013;35(4). https://doi.org/10.1016/j.socnet.2013.09.003.

[70] Smith JA, Moody J, Morgan J. Network sampling coverage ii: The effect of non-random missing data on network measurement. Soc Netw 2017;48:78–99. https://doi.org/10.1016/j.socnet.2016.04.005.

[71] Smith JA, Morgan JH, Moody J. Network sampling coverage iii: Imputation of missing network data under different network and missing data conditions. Soc Netw 2022;68:148–78. https://doi.org/10.1016/j.socnet.2021.05.002.

[72] Peng S, Roth AR, Perry BL. Random sampling of alters from networks: A promising direction in egocentric network research. Soc Netw 2023;72:52–8. https://doi.org/10.1016/j.socnet.2022.09.004.

[73] Birkett M, Kuhns LM, Latkin C, Muth S, Mustanski B. The sexual networks of racially diverse young men who have sex with men. Arch Sex Behav 2015;44(7):1787–97. https://doi.org/10.1007/s10508-015-0485-5.

[74] Weiss KM, Goodreau SM, Morris M, Prasad P, Ramaraju R, Sanchez T, et al. Egocentric sexual networks of men who have sex with men in the United States: Results from the artnet study. Epidemics 2020;30:100386. https://doi.org/10.1016/j.epidem.2020.100386.

[75] Offer S, Fischer CS, Alwin DF, Felmlee DH, Kreager DA. Social networks and the life course: Integrating the development of human lives and social relational networks. Cham: Springer; 2018. p. 117–38. https://doi.org/10.1007/978-3-319-71544-5_6. Calling on kin: The place of parents and adult children in egocentric networks.

[76] Liu L, Wang J, Liu J, Zhang J. Privacy preservation in social networks with sensitive edge weights. in *Proceedings of the 2009 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics; 2009.

[77] Minello G, Rossi L, Torsello A. K-anonymity on graphs using the Szemerédi regularity lemma. IEEE Trans Netw Sci Eng 2021;8(2):1283–92. https://doi.org/10.1109/tnse.2020.3020329.

[78] Hampton KH, Fitch MK, Allshouse WB, Doherty IA, Gesink DC, Leone PA, et al. Mapping health data: Improved privacy protection with donut method geomasking. Am J Epidemiol 2010;172(9):1062–9. https://doi.org/10.1093/aje/kwq248.

[79] Karwa V, Krivitsky PN, Slavković AB. Sharing social network data: Differentially private estimation of exponential family random-graph models. Appl Stat 2017;66(3):481–500. https://doi.org/10.1111/rssc.12185.

[80] Moody J, adams j, Morris M. Epidemic potential by sexual activity distributions. Netw Sci (Camb Univ Press) 2017;5(4):461–75. https://doi.org/10.1017/nws.2017.3.

[81] Resnick MD, Bearman P, Blum RW, Bauman KE, Harris KM, Jones J, et al. Protecting adolescents from harm: Findings from the national longitudinal study on adolescent health. J Am Med Assoc 1997;9(10):832–43.

[82] Hunter DR, Goodreau SM, Handcock MS. Goodness of fit of social network models. J Am Stat Assoc 2008;103(481):248–58. https://doi.org/10.1198/016214507000000446.

[83] Edgar RC, Domrachev M, Lash A. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30(1):207–10.

[84] Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. The sequence read archive. Nucleic Acids Res 2011;39:D19–21. https://doi.org/10.1093/nar/gkq1019.

[85] Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Res 2020;48(D1):D84–6. https://doi.org/10.1093/nar/gkz956.
[86] Los Alamos National Laboratory. *Los Alamos national laboratory LANL*. 1996: United States. Available from: ⟨https://www.loc.gov/item/lcwaN0014171⟩/.
[87] Trans-NIH BioMedical Informatics Coordinating Committee (BMIC). Repositories for sharing scientific data. n.d., National Institutes of Health,: Bethesda, MD. Available from: ⟨https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/repositories-for-sharing-scientific-data⟩.
[88] Leskovec J, Krevl A. SNAP datasets: Stanford large network dataset collection. Stanford University; 2014. ⟨https://snap.stanford.edu/data⟩ (Available from).
[89] Rossi RA, Ahmed NK. The network data repository with interactive graph analytics and visualization. Moutain View, CA 2015. Available from: ⟨http://networkrepository.com⟩.
[90] Clauset, A., E. Tucker, and M. Sainz. *The Colorado index of complex networks (icon)*. 2016; Available from: ⟨https://icon.colorado.edu⟩.
[91] IEEE.org. *IEEE dataport*. n.d. Available from: ⟨https://ieee-dataport.org⟩/.
[92] ICPSR Data Stewardship Policy Committee. *ICPSR access policy framework*. 2010 Updated 13-August-2018; Version 3:[Available from: ⟨https://www.icpsr.umich.edu/web/pages/datamanagement/preservation/policies/access-policy-framework.html⟩.
[93] Harris KM, Halpern CT, Smolen A, Haberstick BC. The national longitudinal study of adolescent health (Add Health) twin data. Twin Res Hum Genet 2012;9(6):988–97. https://doi.org/10.1375/twin.9.6.988.
[94] Panel Study of Income Dynamics. *Produced and distributed by the survey research center*. Institute for Social Research, University of Michigan: Ann Arbor, MI.
[95] Centers for Disease Control and Prevention. Pregnancy risk assessment monitoring system (prams). US Health and Human Services: Atlanta, GA. Available from: ⟨https://www.cdc.gov/prams/index.htm⟩.
[96] Small ML. Someone to talk to. New York: Oxford University Press; 2017.
[97] Bagrow J, Ahn Y-Y. Network cards: Concise, readable summaries of network data. Appl Netw Sci 2022;7(1). https://doi.org/10.1007/s41109-022-00514-7.
[98] Zimmer M. But the data is already public": On the ethics of research in facebook. Ethics Inf Technol 2010;12(4):313–25. https://doi.org/10.1007/s10676-010-9227-5.
[99] Neal ZP, Almquist ZW, Bagrow J, et al. Recommendations for sharing network data and materials. Netw. Sci. 2024;12(4):404–17. https://doi.org/10.1017/nws.2024.16.