RESEARCH ARTICLE

Subdivisions and crossroads: Identifying hidden community structures in a data archive's citation network

Sara Lafia[1], Lizhou Fan[2], Andrea Thomer[2], and Libby Hemphill[1,2]

[1]ICPSR, University of Michigan, Ann Arbor, MI
[2]School of Information, University of Michigan, Ann Arbor, MI

Keywords: archival science, community detection, data citation, data reuse, network analysis

ABSTRACT

Data archives are an important source of high-quality data in many fields, making them ideal sites to study data reuse. By studying data reuse through citation networks, we are able to learn how hidden research communities—those that use the same scientific data sets—are organized. This paper analyzes the community structure of an authoritative network of data sets cited in academic publications, which have been collected by a large, social science data archive: the Interuniversity Consortium for Political and Social Research (ICPSR). Through network analysis, we identified communities of social science data sets and fields of research connected through shared data use. We argue that communities of exclusive data reuse form "subdivisions" that contain valuable disciplinary resources, while data sets at a "crossroads" broadly connect research communities. Our research reveals the hidden structure of data reuse and demonstrates how interdisciplinary research communities organize around data sets as shared scientific inputs. These findings contribute new ways of describing scientific communities to understand the impacts of research data reuse.

1. INTRODUCTION

Data are essential resources for social science research, and data creators' contributions should be rewarded (Alter & Gonzalez, 2018). In addition to ensuring credit, measures of data reuse such as downloads and citations can reveal a data set's role in a research community and provide insights into how researchers engage with data (Cousijn, Feeney et al., 2019). Analyzing data citations reveals data citation practices and provides a way to quantify the analytical utility and disciplinary reach of data collections (Buneman, Dosso et al., 2021). However, it has typically been challenging to find these measures because download data is not widely available, and researchers inconsistently cite data (Buneman, Christie et al., 2020; Lowenberg, Chodacki et al., 2019). Incomplete or opaque research data citations fail to include persistent identifiers, which create obstacles to tracking data use and fail to give appropriate credit to data creators (Moss & Lyle, 2018).

Data archives—particularly domain-specific archives with robust curation services—are ideal sites to study data reuse. They provide data services that make reuse easier, making them sites of research convergence. Archives anticipate data sets that have high analytical potential for long-term preservation and community impact as "topical collections" (Fenlon, 2017; Palmer, Weber, & Cragin, 2011). Additionally, some maintain bibliographies of papers that

an open access journal

Citation: Lafia, S., Fan, L., Thomer, A., & Hemphill, L. (2022). Subdivisions and crossroads: Identifying hidden community structures in a data archive's citation network. *Quantitative Science Studies*, *3*(3), 694–714. https://doi.org/10.1162/qss_a_00209

DOI: https://doi.org/10.1162/qss_a_00209

Received: 16 May 2022
Accepted: 22 June 2022

Corresponding Author: Sara Lafia
slafia@umich.edu

Handling Editor: Ludo Waltman

The MIT Press

reuse data from the archive, therefore tracking "citations" even when they are not formally included in a paper (e.g., NASA's Data Archive Centers: DAACs[1]; biodiversity data aggregators such as Global Biodiversity Information Facility: GBIF[2]; and Data Observation Network for Earth: DataONE[3]). There has been relatively little analysis of the intercitation networks resulting from research data reuse in academic literature, however.

Citations of data in these archives create networks of data sets with attributes that help us understand data reuse and its implications. For instance, understanding the context of data discovery and reuse may help us understand the distribution of ideas or topics within and between research domains, and identify data sets that exhibit exceptional long-term analytical potential (Palmer et al., 2011). Like "hibernators" among research papers (Hu & Rousseau, 2019), valuable data sets may lay dormant for years until they are discovered and "awakened" through reuse. Identifying the different functions that data serve within knowledge communities can help us ensure data creators receive appropriate credit for their contributions.

Additionally, looking for new patterns of data couse or reuse would help identify hidden communities that use archived data in novel ways. Data reuse can be viewed as an indirect form of cooperation and collaboration between researchers (Sands, Borgman et al., 2012; Thomer, Twidale, & Yoder, 2018; Zimmerman, 2008). Data archives promote research by providing access to data sets, and some of these data sets function as "boundary objects" (Star & Griesemer, 1989) or parts of shared information spaces (Bannon & Schmidt, 1989). The visibility of data reuse depends on the vantage point; while data reuse may be visible to those directly involved, larger patterns of reuse may remain invisible, both to the data archive (e.g., data managers) and to prospective data users from different disciplines. Revealing hidden reuse communities and their structures helps us understand what roles data play in knowledge production and how they function as boundary objects between fields of research.

Despite recent data-sharing mandates, securing data deposits is still a challenge for data archives. Researchers are often wary of sharing data because they fear being "scooped" or are unsure how other researchers might use their data (Borgman, Scharnhorst, & Golshan, 2018; Cragin, Palmer et al., 2010). Mapping the network of data citations provides evidence of data reuse that will help data producers and archives better assess the collaborative utility of data and demonstrate different types of secondary use to researchers and potential depositors.

In this paper, we inspect an authoritative bibliography of social science data sets cited in academic publications from the Inter-University Consortium for Political and Social Research (ICPSR) Bibliography of Data-related Literature[4]. Specifically, we analyzed its citation graph to uncover hidden community structures and identified the different roles data sets play in networked communities. By linking citations to metadata from a scholarly database, Dimensions, we were able to include attributes such as "fields of research"[5] in our analysis (Hook, Porter, & Herzog, 2018). We then used community detection algorithms to identify hidden communities within the network of data citations and identified two types of data sets that unite scientists involved in social science knowledge production: subdivision data sets and crossroads data sets. Subdivisions exclusively function as disciplinary resources used by a narrow set of fields.

[1] https://lpdaac.usgs.gov/resources/publications/.
[2] https://www.gbif.org/resource/search?contentType=literature&relevance=GBIF_USED.
[3] https://search.dataone.org/profile.
[4] https://www.icpsr.umich.edu/web/pages/ICPSR/citations/.
[5] According to Dimensions, the fields of research (FoR) is a hierarchical classification applicable for categorizing all research and development activity.

Crossroads, by contrast, enable interdisciplinary research. The network structures we identify and name acknowledge the variation in reuse and help us recognize the myriad functions that data sets serve in scientific communities.

2. BACKGROUND AND RELATED WORK

2.1. Data Archives as a Site for Understanding Scholarly Communication Practices

Data archives support data-intensive research by providing long-term data stewardship, access, and high-quality data curation. Notable examples of data archives with high levels of curation include GenBank, a rich repository of genetic sequence data; SESAR, a repository of metadata describing physical samples in the earth sciences, as well as links to derived data sets; and PANGAEA, a publisher for georeferenced data sets linked to earth system studies. Data sharing through archives enables researchers to find and reuse data that they did not collect. In other words, data created for one purpose can be used by new audiences to answer new questions (Brown, 2003; Wilkinson, Dumontier et al., 2016). Researchers can use existing data to validate previous findings, extend their data collections, or form the basis for new studies via integration or independent reuse (Gregory, Groth et al., 2020; King, 1995; Pasquetto, Randles, & Borgman, 2017; Thomer, 2022). Additionally, as more funders and journals mandate that data from grants and papers be shared openly, data archives are only growing in importance as sites of scholarly communication.

The data held in these repositories often have untapped reuse potential across disciplinary boundaries (Hey, Tansley, & Tolle, 2009; Palmer et al., 2011). Such interdisciplinary research using archived data can lead to breakthrough discoveries (National Academy of Sciences, 2005; Tenopir, Allard et al., 2011). Fields of research may share an interest in explaining different aspects of the same phenomenon, giving rise to interfield theories that bridge fields of science (Darden & Maull, 1977). "Borderland disciplines" sometimes form where fields of research collide over shared resources, such as instruments or data, leading to the evolution of new techniques (Gökalp, 1987). Data sets that facilitate interactions between research areas therefore function as "boundary objects," carrying multivalent analytical potential across research communities (Star & Griesemer, 1989) and facilitating knowledge exchange across boundaries. However, there has been little research on the prevalence of such data sets-as-boundary-objects. We know little about which features of data sets promote boundary crossing, or how to measure their collaborative potential.

2.2. Data Citation Standards and Emerging Data Citation Networks

One way of exploring interdisciplinary data reuse—and therefore, the extent to which data sets function as boundary objects between communities—is by studying data citation networks. Efforts to promote data citation over the last 20 years have led to the adoption of new data citation practices in many communities. Milestones formalizing data citation include the Joint Declaration of Data Citation Principles (Data Citation Synthesis Group, 2014), Data Citation Roadmap for Scholarly Data Repositories (Fenner, Crosas et al., 2016), and Data Citation Roadmap for Scientific Publishers (Cousijn, Kenall et al., 2018). Data citation counts provide a foundation for studying the scholarly impact of scientific data and the value of data curation efforts.

The adoption of data citation principles makes it possible to analyze emerging data reuse behavior and structures of hidden research communities in data citation networks. Citation networks generally represent documents as vertices and citations of one document by another

as edges (Leicht, Clarkson et al., 2007). Citation networks can highlight central nodes such as influential institutions; heavy edges between nodes indicate important connections and processes, such as the diffusion of ideas (Chen, 2017). Prior studies of citation networks have provided insights into ties between individual researchers and collaborations between research disciplines (Tomasello, Vaccario, & Schweitzer, 2017). Studies of publication citation networks (e.g., papers or journals) have also identified novel papers, measured the impact of papers and their authors, and attributed discoveries to authors (Newman, 2004).

Whereas publication citations broadly enable lineage retrieval for ideas, data citations indicate the origins and processing history of the data sets that have been used in an analysis (Bose & Frew, 2005). Data citation networks reflect connections between disciplinary literature and the research data that they draw from. They reveal the reach of research data and support the computation of bibliometrics that show the relationships and impacts of scientific products (Buneman et al., 2021). The interactional context of data production and citation also reflects relationships between data producers and consumers in a broader data economy (Vertesi & Dourish, 2011).

Quantifying the scholarly impact of data archives and other research infrastructures relies on proxy measures for data usage, such as downloads and citations (Mayernik, Hart et al., 2017). However, a number of recent studies have highlighted the limitations of studies that rely on current data citation tracking infrastructures. Platforms such as DataCite have the potential to enable large-scale studies of data production and its scholarly impact (e.g., citations); however, a lack of consensus on the definition of "data" and alignment of metadata across providers limits DataCite's analytical potential (Robinson-Garcia, Mongeon et al., 2017). An analysis of publication-data set networks constructed from GenBank and Figshare found that authors tend to cite publications over data sets, suggesting that historically, data sets have not been regarded as first-class research objects and that data use inferred from citation networks may undercount data use (Zeng, Wu et al., 2020). We avoid concerns about data and metadata quality by focusing exclusively on a curated bibliography linking social science studies to publications held by a single data archive. A recent study of ICPSR's metadata records described the thematic and temporal dimensions of social science data sets and their citing literature separately (Lee & Jeng, 2019). We jointly analyze data and publications by constructing an interdisciplinary cocitation network. To tap the potential of shared data sets, we examine the role that data citations play in the production and dissemination of knowledge in the social sciences.

2.3. Exclusive and Inclusive Communities in Knowledge Organizations

The analysis of citation networks can reveal hidden organizational structures. Cocitation analysis studies the structure of science and the emergence of specialties in bibliometric networks by examining how frequently pairs of documents are invoked (Small, 1973). Author cocitation analysis reveals individual contributions to specialty areas and paradigm shifts in the research landscape (White & McCain, 1998). Citation analysis can be used to identify exclusionary community structures, such as "invisible colleges" (Price & Beaver, 1966)—in-groups that control scientific discourse, which are defined by strong ties and informal communication (Crane, 1977). Similar analyses can also detect "citation cartels" of authors who cite each other exclusively, and effectively shut out other authors who work on the same subject (Franck, 1999). In addition to exclusionary practices, citation analysis can also identify convergence in research communities. Studies of cross-field citation networks have found that fields of science tend to become more integrated, rather than exclusive, over long periods of time (Varga, 2019), albeit incrementally across neighboring disciplines (Porter & Rafols, 2009).

While the notion of "community" is central to these analytical methods, it is a difficult concept to operationalize (Orthia, McKinnon et al., 2021); communities may take many forms, and may play many roles. Identifying communities via data citation is further complicated by the interdisciplinary nature of data analysis and citation (Heidorn, 2008). However, we take inspiration from prior work showing that data reuse can be viewed as an indirect form of cooperation and collaboration between researchers—and groups that commonly reuse the same data might be considered communities-at-a-distance (Sands et al., 2012; Thomer et al., 2018; Zimmerman, 2008). Research data is a primary input for scientific knowledge production, making data archives important sites for identifying nascent research communities. We use community detection to reveal patterns of data reuse and examine the structure of research communities that use data as shared scientific inputs.

3. DATA AND METHODS

We analyzed the ICPSR Bibliography, an authoritative source of high-quality, manually curated links between 8,071 social science studies and 101,674 publications that have cited them. An additional 2,420 studies (23%) do not have any data-related publications and so are not represented in the Bibliography. At ICPSR, each study consists of one or more data files and metadata. Table 1 provides an example of available metadata for a highly cited ICPSR study.

Curation of the ICPSR Bibliography is labor-intensive, so the current coverage of the ICPSR Bibliography is uneven[6]. Bibliography staff search broadly for academic literature that references ICPSR studies and add literature to the Bibliography only if it analyzes ICPSR data or includes an extensive discussion of data-related methodology. Publications in the Bibliography are a mixture of materials published by the original data creator and publications that analyze existing data. The majority of materials are journal articles, reports, conference proceedings, theses, books, and book chapters. We restricted our analysis to materials published since the inception of ICPSR as an archive in 1962.

We analyzed citations for all of ICPSR's currently available studies. Many ICPSR studies have institutional principal investigators (PIs) including U.S. government agencies (e.g., U.S. Census Bureau, Department of Justice, Department of Education, Department of Health and Human Services), news outlets (CBS News, the New York Times), and university research centers (e.g., University of Michigan's Survey Research Center). Teams of individual researchers also deposit data with ICPSR. Studies in our analysis included both restricted and public data files. The terms of use for restricted data prohibit linking it to other data, so studies that include restricted data may be undercounted in terms of their potential use.

The majority of ICPSR's studies (62%) are also part of a series, meaning that they are part of a recurring collection with new data archived over time (e.g., repeated cross-sectional studies or longitudinal studies). ICPSR provides access to 278 series. We used a natural breaks classification (Jenks, 1963) to find highly cited series, which are reported in Table 2.

3.1. Network Definitions

We constructed citation networks from the ICPSR Bibliography, which are summarized in Table 3. Given that studies from the same series have been created intentionally to be

[6] The process of retrieving citations for all studies is ongoing. Because staff are actively searching for publications that reference ICPSR data sets, these measures are minimum counts, which likely underestimate the number of papers and their relationships.

Table 1. Example of available metadata for an ICPSR study

Study name	Series title	Release	Citations	Subject terms
Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 1996	Monitoring the Future (MTF) Public-Use Cross-Sectional Datasets	1998-10-05	251	attitudes, demographic characteristics, drug use, family life, high school students, life plans, lifestyles, social behavior, social change, values, youths

analyzed together (e.g., across years), we grouped studies by their series and referred to the resulting unit as a "data set"—either one series with multiple studies or one study that is not part of a series. Grouping studies into ICPSR-defined series allowed us to distinguish data that were *designed* to be used together (e.g., by their project sponsor, funder, archive) from data that have been *discovered* to be useful together (e.g., by researchers who cocite them in literature).

Because publications and data sets are two different classes of objects in the ICPSR Bibliography, we modeled the connections between them in a bipartite network (B), consisting of publication nodes, data set nodes, and edges linking publications to the data sets that they cite. Citations are based on the total number of publications that use data from a study or series. From network B, we projected data set nodes to create a weighted data set cocitation network (S). Edge weight in S indicates the total number of times that a pair of data sets have been used together in publications. We removed low-frequency data cocitations from our analysis to focus on data sets that were used together across multiple publications; we removed edges from S with a weight less than 2, meaning that those data sets were only used together once. This reduced edges by 87% (from 24,942 to 3,208) and nodes by 70% (from 3,363 to 998).

Table 2. Features of highly cited ICPSR series data

Series title	Lead investigators	Studies in series	Combined citations
American National Election Study (ANES) Series	Warren E. Miller et al. and the National Election Studies	92	16,771
Uniform Crime Reporting Program Data Series	Federal Bureau of Investigation	263	13,041
Monitoring the Future (MTF) Public-Use Cross-Sectional Datasets	Lloyd D. Johnston et al.	76	11,808
Current Population Survey Series	US Bureau of the Census	296	11,012
National Health and Nutrition Examination Survey (NHANES) and Followup Series	Kathleen Mullan Harris et al.	3	6,951
National Survey on Drug Use and Health (NSDUH) Series	United States Department of Health and Human Services; National Institutes of Health; National Institute on Drug Abuse	29	5,893
National Electronic Injury Surveillance System (NEISS) Series	United States Department of Health and Human Services; Centers for Disease Control and Prevention; National Center for Injury Prevention and Control	38	5,255
National Crime Victimization Survey (NCVS) Series	Bureau of Justice Statistics	85	4,472

Table 3. Summary of network definitions and metrics

Network	*B*	*S*	*F*
Nodes	Publications, data sets	Data sets	Fields of research
Edges	Publication cited ICPSR data set	ICPSR data sets cited in the same publication	Publication tagged with both fields cited one ICPSR study
***N* (nodes)**	90,922 publications; 3,363 data sets	998 data sets	129 research fields
***N* (edges)**	102,580	3,208	4,238
Node size	Constant	Constant	$\log(N_{papers})$
Edge weight	n/a	1 for each publication in which the pair of ICPSR data sets is cited	1 for each ICPSR study a publication cites
Components	1,687	80	1
Density	$2.3e^{-5}$	$6.4e^{-3}$	0.51
Transitivity	n/a	0.28	0.74
Degree assortativity	n/a	−0.02	−0.30

Next, we used a similar process to define a field of research network (*F*) (Cunningham, Smyth, & Greene, 2022). We gathered supplementary publication metadata for a subset of 44,639 publications in the ICPSR Bibliography (45% of the total) that were available in the Dimensions database (Hook et al., 2018). We retrieved field of research (FoR) codes for each publication. FoR codes consist of 22 high-level divisions and their subgroups (e.g., Curriculum and Pedagogy is a subgroup of Education). We linked FoR codes to ICPSR data sets through their corresponding publications in an unweighted bipartite network (*B′*). We then projected the FoR nodes to create a weighted cocitation network (*F*). In *F*, edges are data sets that are cocited between fields of research. Because each study could be cited by many different combinations of fields of research, we did not group studies by their series, allowing for the observation of different cocitation patterns in the same series of studies. Edge weight indicates the total number of times a pair of data sets have been used together in publications. We simplified *F* by removing low-frequency FoR cocitations, which correspond to edges with a weight less than five.

3.2. Community Detection

We applied community detection algorithms to each network as summarized in Table 4. Community detection identifies nodes that have a high probability of interacting based on the network structure (Fortunato & Hric, 2016). We selected detection approaches based on the desired representation of communities in each type of network (Lancichinetti & Fortunato, 2009; Yang, Algesheimer, & Tessone, 2017). We allowed communities to overlap in the data set cocitation network because we wanted to identify data sets with multiple roles. However, we did not allow overlap in the field of research network because we wanted to find communities defined by members with the strongest ties.

Table 4. Summary of community detection approaches

Network	Definition	Community detection method	Community definition	Communities detected
S	Data sets (studies or series)	*k*-clique ($k = 3$)	Data sets used in the same paper	41
F	Fields of research (FoR) in papers	Louvain	Fields of research that use the same study-level data	4

We applied a *k*-clique percolation method to the data set cocitation network (*S*) using the corresponding implementation from the NetworkX Python library (Hagberg, Swart, & S Chult, 2008). A clique is a complete subgraph of a defined size (*k*) that can be reached from the cliques of the same community through a series of adjacent cliques, meaning that the cliques share $k - 1$ nodes (Palla, Derényi et al., 2005). Each node may belong to more than one clique, resulting in overlapping communities. We selected a minimum clique size of three and labeled each community with the three most common ICPSR subject terms for all studies in each clique. Subject terms uniformly describe topics covered by the data and are defined by a controlled vocabulary of social science concepts in the ICPSR Subject Thesaurus, which are assigned during data curation.

We then selected an aggregation-based method to represent communities in our field of research network. We applied the Louvain algorithm to the FoR network (*F*) using the corresponding implementation from the Louvain Python library (Hagberg et al., 2008). The algorithm uses modularity to discover communities in large networks by moving nodes locally to create a network aggregation; communities are merged until the resulting modularity of the overall partition can no longer increase (Blondel, Guillaume et al., 2008). This method results in nonoverlapping communities that show the most densely connected fields of research that cocite ICPSR data sets. The networks (*S*, *F*) were then arranged with a spring layout, which places nodes with high degrees at the center of the graph.

4. RESULTS

We used two network measures—centrality and betweenness—to interpret the importance of data sets and fields of research in their respective cocitation networks (Newman, 2003). First, we calculated each node's degree as the number of connections it shares with all other nodes in the network. High-degree nodes are prominent in the network because they are highly connected. We also calculated each node's betweenness centrality by measuring all shortest paths passing through a given node. Nodes with high betweenness function as hubs and connect disparate parts of the network.

We also assessed structural features of the network—number of components, assortativity, density, and transitivity—to compare the data set and field of research cocitation networks (Table 3). The FoR network is connected, meaning that all of its nodes are in the same component, while the other two networks have multiple components or disconnected subgraphs. This suggests that the FoR network is less complex than the data set cocitation network. Both *S* and *F* exhibit negative degree assortativity, meaning that their nodes are less likely to be connected to nodes in the network with a similar degree value. This pattern is stronger in *F* (−0.30) than in *S* (−0.02). Finally, networks *B* and *S* have low density ($2.3e^{-5}$ and $6.4e^{-3}$, respectively), while network *F* is far denser (0.51), indicating that *B* and *S* have comparatively fewer edges linking nodes and are not as easily traversed as *F*.

4.1. Data Set Cocitations

The data set cocitation network (*S*) has a periphery of data sets that have been used together only a few times and a denser core of highly connected data sets, which are often used together. Figure 1 highlights important, central data sets, which are all found in the largest subgraph at the core of the network. We used natural breaks to determine six data sets with high betweenness and degree centrality, which play important roles in the network (Table 5).

The important data sets we identified are long-running series made up of multiple studies. Of these, the Uniform Crime Reporting Program Data Series has the highest degree and betweenness. It has been used with 115 other data sets from studies or series across the citation network. The other data sets have strong ties to many other data sets and connect components of the network. Half of these data sets are highly cited, with more than 10,000 citations each; the others are less cited, yet play an important role in connecting the network. Finally, the lead investigators for these important data sets include institutional PIs, meaning that one of the study's principal investigators or depositors is an institution (e.g., the US Bureau of the Census), and noninstitutional PIs.

To find collections of data sets that are often used together in publications, we performed community detection on the data set cocitation network (Figure 2). Not all studies belong to a cocitation community. Only a fraction of data sets in the analysis (N = 632; 63%) belong to cliques of size three or larger; these data sets are often analyzed with at least two additional ICPSR data sets. The data sets that fell out of our analysis were used independently and were not combined with other data sets. We labeled each community with the three most common ICPSR subject terms for all data sets within it. The largest clique has 461 data set members and is topically broad (e.g., "demographic characteristics, employment, income") while smaller cliques tend to have narrower focuses (e.g., "terrorism, terrorists, radicalism").

We also identified 20 data sets (3% of all nodes in the network) that belong to more than one community, which may facilitate analyses across topics. Of these, we summarized data

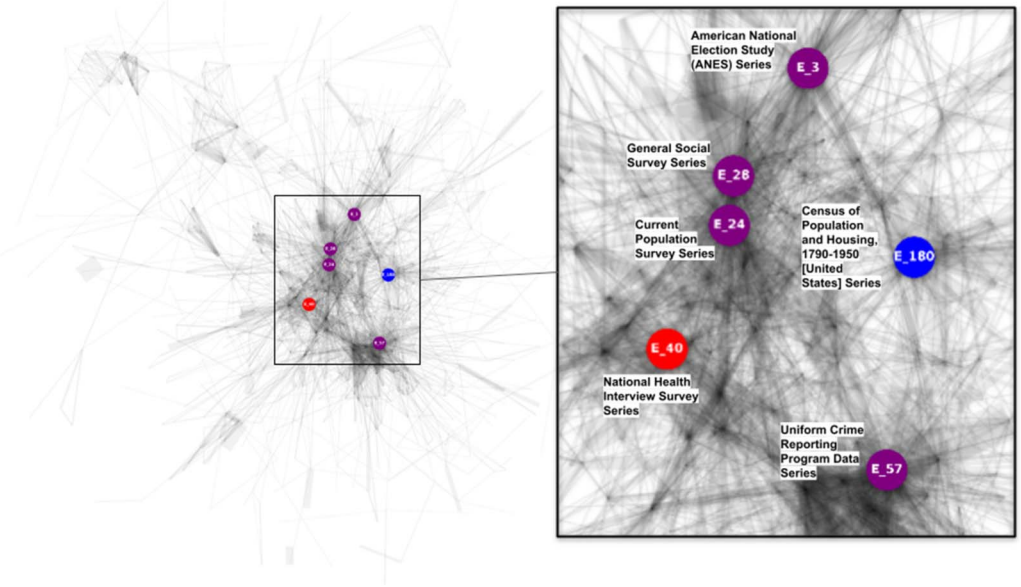


Figure 1. Overview of data set cocitation network featuring data sets functioning as hubs. Inset: High *degree* (red), high *betweenness* (blue), and high *degree* and *betweenness* (purple) nodes.

Table 5. Data sets with high betweenness and degree centrality in cocitation network

Data set name	Investigators	Betweenness	Degree	Studies in series	Combined citations
Uniform Crime Reporting Program Data Series	Federal Bureau of Investigation	0.17	115	263	13,041
General Social Survey Series	National Opinion Research Center; Davis et al.	0.12	113	15	1,551
American National Election Study (ANES) Series	Miller et al.; National Election Studies	0.11	109	92	16,771
Current Population Survey Series	US Bureau of the Census	0.11	117	296	11,012
Census of Population and Housing, 1790–1950 [United States] Series	Haines et al.; US Bureau of the Census	0.10	72	2	818
National Health Interview Survey Series	National Center for Health Statistics	0.05	80	155	4,448

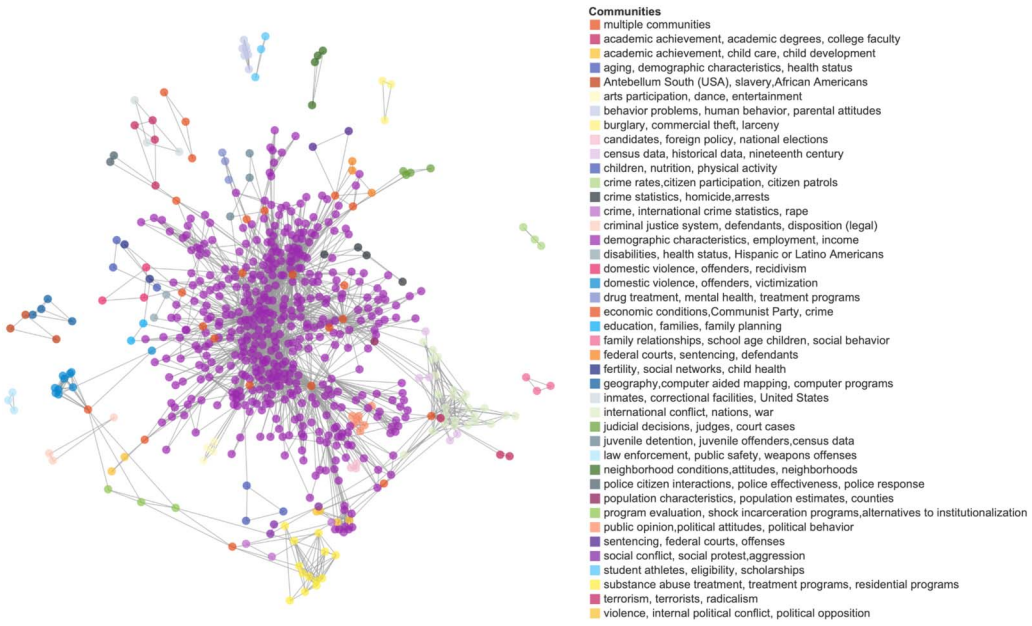


Figure 2. Result of community detection (41 communities detected at $k = 3$) with labels generated from the three most frequent subject terms for the data sets in each community. An interactive graph with detailed node information is available in Tableau[7].

sets that belong to more than two communities, along with examples of other data sets that they have been cocited with, and a representative publication that has cited the same data Table 6. For example, the Census of Population and Housing, 1790–1950 [United States] Series appears in three different data set communities. It has been used with other ICPSR data sets to study topics such as industrial development and urbanization in the United States; conflict and international trade; and social movements and elections.

[7] https://public.tableau.com/app/profile/lizhou/viz/Study_communities_v2/Study_Communities_2_Dashboard.

Table 6. Data sets in more than two communities, their cocited data sets, and publications

Data set	Community label terms	Example of cocited data sets	Example of citing publication
American National Election Study (ANES) Series	demographic characteristics, employment, income	National Black Politics Study, [United States], 1993	Wiegand, A. W. (1999). *Differences in public opinion between blacks and whites: A social psychological perspective.* University of California, Santa Cruz.
	public opinion, political attitudes, political behavior	Swedish Election Test-Data Series: Swedish Election Study, 1979	Granberg, D., & Holemberg, S. (1991). Election campaign volatility in Sweden and the United States. *Electoral Studies*, *10*(3), 208–230.
	candidates, foreign policy, national elections	American Representation Study, 1958: Candidate and Constituent, Incumbency	Hill, K. Q., & Hurley, P. A. (1979). Mass Participation, Electoral competitiveness, and issue-attitude agreement between congressmen and their constituents. *British Journal of Political Science*, *9*(4), 507–511.
Census of Population and Housing, 1790–1950 [United States] Series	demographic characteristics, employment, income	United States Agriculture Data, 1840–2012	Kitchens, C. T., & Rodgers, L. P. (2020). *The impact of the WWI agricultural boom and bust on female opportunity cost and fertility* (No. w27530). National Bureau of Economic Research.
	international conflict, war, nations	Direction of Trade	McKeown, T. J. (1991). A liberal trade order? The long-run pattern of imports to the advanced capitalist states. *International Studies Quarterly, 35*(2), 151–172.
	census data, historical data, 19th century	National Samples from the Census of Manufacturing: 1850, 1860, and 1870	Dobis, E. A. (2016). *The evolution of the American urban system: history, hierarchy, and contagion.* Doctoral dissertation, Purdue University.
Monitoring of Federal Criminal Sentences Series	demographic characteristics, employment, income	Federal Justice Statistics Program Data Series	Bureau of Justice Statistics. (2021). Tribal crime data collection activities. *Technical Report.* NCJ 301061, Washington, DC: Bureau of Justice Statistics.
	federal courts, sentencing, defendants	Court Workforce Racial Diversity and Racial Justice in Criminal Case Outcomes in the United States, 2000–2005	Ward, G., Farrell, A., & Rousseau, D. (2009). Does racial balance in workforce representation yield equal justice? Race relations of sentencing in federal court organizations. *Law & Society Review, 43*(4), 757–806.
	sentencing, federal courts, offenses	Impact of Sentencing Guidelines on the Use of Incarceration in Federal Criminal Courts in the United States, 1984–1990	Tonry, M. (1991). Mandatory minimum penalties and the US Sentencing Commission's mandatory guidelines. *Federal Sentencing Reporter, 4*(3), 129–133.

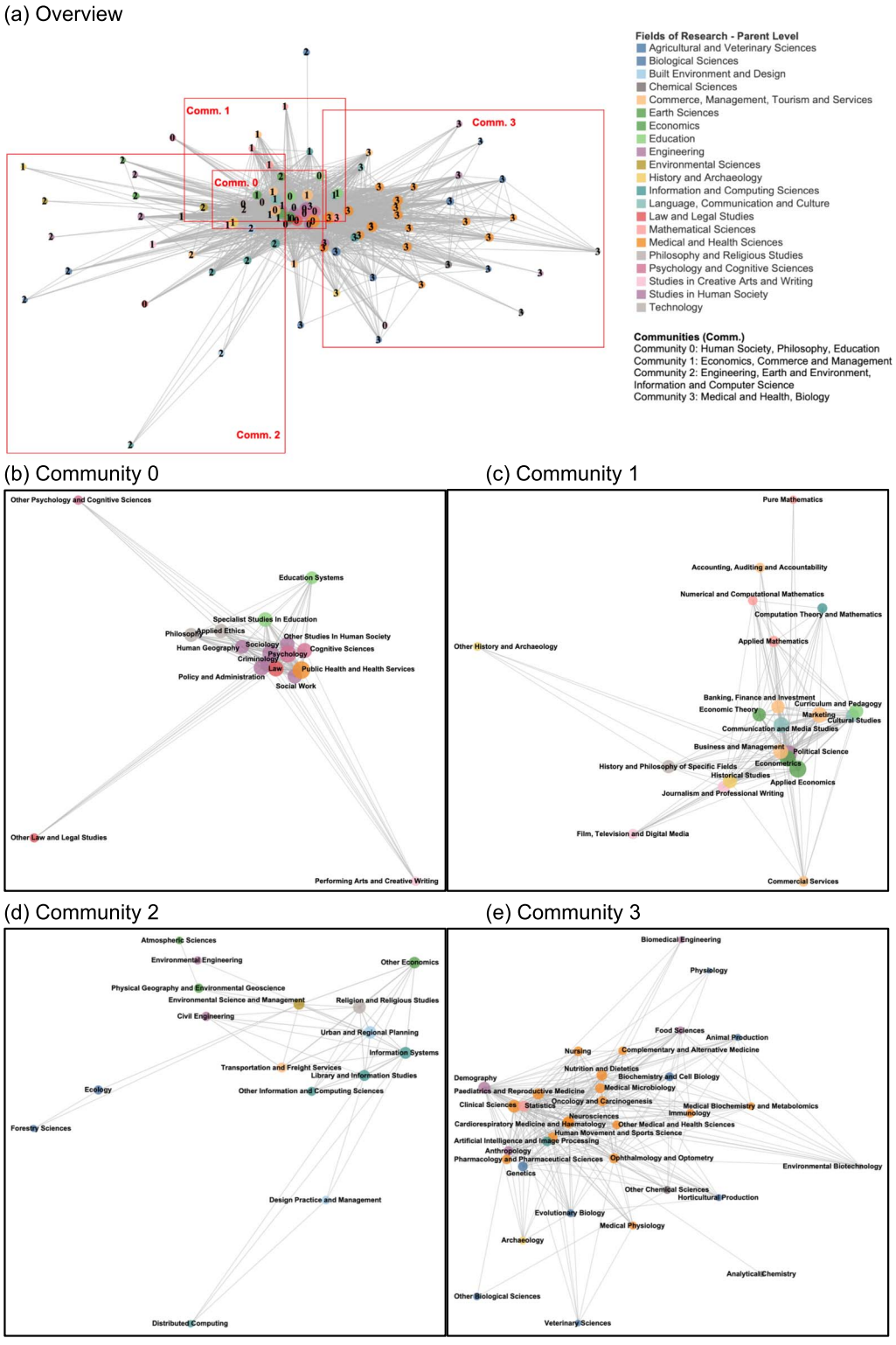


Figure 3. Results of community detection in the field of research network (*F*, with nodes connected by edges of size ≥ 5). The interactive graph with detailed node and edge information in size and study numbers is available in Tableau[8].

[8] https://public.tableau.com/app/profile/lizhou/viz/CommunitiesinFieldsofResearchFoR/CommunitiesinFieldsofResearch.

4.2. Fields of Research

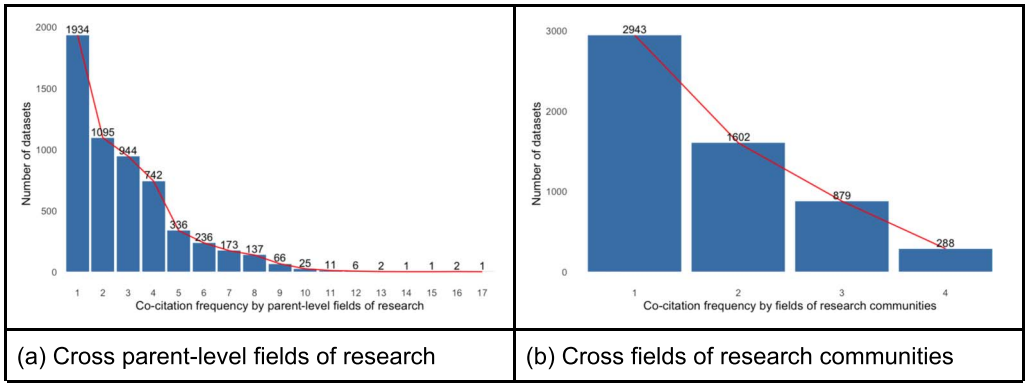
To find fields of research (FoR) that often use the same data sets, we performed community detection on the FoR cocitation network (*F*). Nodes in *F* are color-coded by their parent-level divisions and labeled by their child-level code. We detected four large communities, which are summarized in Figure 3(a). The primary fields of research in each community are Human Society, Philosophy and Education (Community 0); Economics, Commerce and Management (Community 1); Engineering, Earth and Environment, Information and Computer Science (Community 2); and Medical and Health, Biology (Community 3).

Fields in the center of *F* have more cocitations, meaning that they are highly connected to other fields. The central red frame in Figure 3(a) shows the major domains of research that cite ICPSR data sets: Human Society, Philosophy and Education (Community 0). These central domains are consistent with the idea that most items in the ICPSR Bibliography are social science publications. Indeed, social science (e.g., Study of Human Society) and methodological research fields (e.g., Statistics) are found in the core of the network while humanities and other fields (e.g., Creative Writing, Performing Arts) exist mostly on the periphery.

Figures 3(b)–(e) show the composition of each of the four communities in greater detail. We found that the communities tend to divide along disciplinary lines. For example, members within each community are similar, in that they tend to share the same parent-level field of research. For example, "Human Geography" and "Sociology" share the same parent-level field of Human Society and are grouped into the same community (Community 0).

To examine the extent to which similar fields of research use the same data sets, we calculated citation statistics based on network *F*. We consider fields "similar" if they belong to the same parent-level field (e.g., "Civil Engineering" and "Environmental Engineering" are both classified under Engineering) or the same community. We found that similar fields of research cocite a limited range of data sets. The distribution of the aggregated numbers of data sets for cocitation frequency by parent-level fields of research roughly follows a Poisson distribution with $\lambda = 1$, indicating that as the number of parent-level fields citing the data set increases, the number of cocitations decreases (Figure 4(a)). More than half (2,943 of 5,712) of the data sets in *F* are cocited by only one community, further suggesting that data set use tends not to cross community boundaries (Figure 4(b)).

We also observed core and periphery structures in the FoR network shown in Figure 3(a). Table 7(a) shows examples of fields of research located at the core of each community



(a) Cross parent-level fields of research (b) Cross fields of research communities

Figure 4. Data sets cited by parent-level fields of research. The *y*-axis indicates how many data sets were cited by the number of parent-level fields on the *x*-axis. Most data sets are cited by a single parent-level field of research.

Table 7. Examples of nonsocial science fields of research with core and periphery structures

(a) Fields in the core of each community subgraph

Community membership	Field of research	Number of connected fields of research—degree centrality of nodes
0	Psychology	1,833
0	Cognitive Sciences	543
0	Law	1,308
1	Applied Mathematics	26
2	Library and Information Studies	66
2	Information Systems	138
3	Statistics	363
3	Artificial Intelligence and Image Processing	110

(b) Fields in the periphery of each community subgraph

Community membership	Field of research	Example of frequently cocited data set and corresponding fields of research
0	Performing Arts and Creative Writing	"National Crime Victimization Survey: School Crime Supplement, 2011", cocited by fields including Policy and Administration, Criminology, Sociology, Specialist Studies in Education, Psychology, Public Health and Health Services, Cognitive Sciences
1	Curriculum and Pedagogy	"Midlife in the United States (MIDUS 2), 2004–2006", cocited by fields including Applied Mathematics, Banking, Finance and Investment, Economic Theory, Communication and Media Studies, Business and Management, Political Science, Econometrics, Applied Economics, Commercial Services
2	Transportation and Freight Services	"American Time Use Survey (ATUS): Arts Activities, [United States], 2003–2018", cocited by fields including Environmental Science and Management, Other Economics, Religion and Religious Studies, Urban and Regional Planning, Information Systems, Library and Information Studies
3	Archaeology	"National Health and Nutrition Examination Survey III, 1988–1994", cocited by fields including Anthropology, Demography, Clinical Sciences, Statistics, Artificial Intelligence and Image Processing, Human Movements and Sports Science, Neurosciences, Nutrition and Dietetics, Other Medical and Health Sciences, Biochemistry and Cell Biology

subgraph. They include a wide range of subfields such as Psychology, Statistics, and Library and Information Studies, which often advance methodological practices and make data-related contributions. These nodes are highly connected to other fields of research and have a much higher degree centrality compared to the average degree of nodes in *F*, which is 9.

Fields of research in the periphery of each community subgraph (Table 7(b)) reveal hidden connections among disciplines through the data sets that they cocite. For example, Archaeology was cocited by 10 fields—while some of the cocitations are from social science disciplines such as Anthropology and Demography, many others are related to biological and physical sciences, including Clinical Sciences, Neurosciences, and Nutrition and Dietetics, which are found in Community 3.

5. DISCUSSION

In this article we have applied metaphors from the built environment to interpret the hidden research communities that we detected, and labeled the structures *subdivisions* and *crossroads*. These metaphors remind us that these communities of data use have emerged through patterns of interaction in the research landscape and can be reshaped through intentional design. We refer to data sets in research as *subdivisions* if they are inward-facing, exclusive, and not well connected to other data sets or fields. Conversely, we refer to data sets that are often traversed by communities and fields as *crossroads*. We find 632 research data sets in *subdivisions* that function as disciplinary resources; 20 research data sets at *crossroads* in the network that function as boundary objects by facilitating interdisciplinary research; and nonsocial science fields that engage with social science data.

5.1. Subdivisions: Disciplinary Research Community Resources

We refer to data sets that serve a single disciplinary community as *subdivisions* because they are inward-facing, exclusive, and not well connected. The largest data communities we detected focus on international conflict, substance abuse, victimization, and public opinion polls. Despite the topical breadth of the data set network (*S*), it partitioned into coherent cliques with a structure better described as a patchwork of subdivisions than a melting pot. By comparison, the FoR network (*F*) had high density and high transitivity, suggesting that its nodes tended to be clustered together. Given its cohesive structure, we partitioned *F* into a small number of meaningful communities.

To understand the communities that function as subdivisions, we drew from a combination of metrics computed for each network, which are summarized in Table 3. Overall, the data set cocitation network (*S*) isn't well connected. It has low density and low transitivity, is nonassortative based on degree, and contains many components. By comparison, the field of research network (*F*) has a negative degree assortativity, meaning that high-degree fields of research nodes tend to attach to low-degree nodes. The network is not fractured compared with the data set cocitation network (*S*) and has only one component.

In the data set network (*S*) shown in Figure 2, we found instances of isolated cliques with data sets that were exclusively used together. For example, we detected a clique of three data sets described by the terms "Antebellum South (USA), slavery, slave labor." These data sets ("Southern Farms Study, 1860"; "Mortality in the South, 1850"; and "New Orleans Slave Sale Sample, 1804–1862") have different investigators and were produced for different purposes, yet have been used together numerous times in academic articles. These three studies function like a collection even though ICPSR did not designate them as one (i.e., by naming them a series). In general, the analytic utility of data sets in subdivisions is limited to specific areas of research. The notion of "thematic research collection"—a set of materials on a related theme (Fenlon, 2017; Palmer, 2004)—may be useful for data archives to adopt; finding groups of data used together is one way to identify candidate collections.

We also found examples of cliques that shared topics, yet were disconnected from each other (e.g., "domestic violence, offenders, recidivism" and "domestic violence, offenders, victimization"). While these data sets may be topically similar, researchers have not yet used these data together. Cliques may be exclusive or disjointed for discovery reasons (i.e., researchers outside of the user group are not aware of these data) or their data may be discoverable but unsuitable (e.g., due to variables, geography, or other properties). For example, one community with data about "drug treatment" is composed of studies funded by the U.S. Department of Health and Human Services, while a separate community of "substance abuse" data sets is funded by the U.S. Department of Justice. These distinct communities may have stances toward a research topic that are not interoperable and may even conflict.

In the field of research network (*F*) in Figure 3, we observed a subdividing tendency and an in-group cocitation pattern for similar fields of research. These patterns of connection suggest that each field of research cites a limited range of ICPSR data sets and supports the idea that ICPSR data use divides along disciplinary boundaries (e.g., social science disciplines such as economics and education tend to cite the same data sets, but this is less common across nonsocial science fields, such as engineering or nursing). Data sets in *subdivisions* have high analytic potential for narrow communities of research; surfacing them and increasing their visibility may also help unlock hidden potential for new uses beyond those narrow communities.

5.2. Crossroads: Engagement Across Research Communities

Data sets that facilitate interdisciplinary research are *crossroads* because they are often traversed in connecting communities; in comparison to subdivision data sets, they are rare. For instance, ICPSR is well known for large series data sets (e.g., American National Election Study [ANES]), which attract data users to the archive. We found several of these series in the largest clique (see Table 5), which overlaps with the largest subgraph of the network. These series are well known and have high engagement across multiple research communities. In particular, the ANES Series and the Uniform Crime Reporting Program Data Series are institutionally funded, highly cited, and connect a network of researchers who use them.

Prior work found a correlation between data sets with at least one institutional PI and higher data reuse (Hemphill, Pienta et al., 2022). When we examine data reuse based on citations rather than downloads, however, the relationship between data sets with institutional PIs and reuse is less clear. Some institutional data sets already link multiple data sources into a single data set and are useful on their own; they may not need to be combined with other data sets to be analytically powerful. Among the crossroads data sets we found, the Census of Population and Housing data set is unique because the individual investigator who constructed the data set combined multiple years and data sources into a single data set, which has been broadly useful across many applications.

In addition to the three connective data sets described in Table 6, we found 17 additional data sets that function as crossroads between research communities. Many of these data sets were often used with less cited data sets, explaining the negative associativity observed in network *F*. For example, the less cited "Vietnam Longitudinal Survey, 1995–1998," is used with the highly cited "India Human Development Survey (IHDS) Series" and "Chitwan Valley [Nepal] Family Study Series" to study education, families, and family planning. Researchers who seek data from a well-known study may traverse the citation network to find complementary data sets from lesser-known studies. While a single data series such as the IHDS might meet only some users' needs, given its limited geographic coverage, the data sets linked through its connections offer opportunities for comparative analysis.

In the field of research network (*F*), we found two dominant patterns of cocitation, summarized in Table 7. Fields in the core of the network are highly connected and operate at an interdisciplinary crossroads; they tend to use more data sets in common with other fields. These fields, such as Statistics and Applied Mathematics, are not in the social sciences. Rather, the data sets that they use function as crossroads, activating sites for research convergence. In Community 3 (Figure 3(e)) for example, Statistics cocites many of the same data sets as Biology, Neuroscience, and Medical Sciences. Statistical methods are often applied in data analysis and can advance the development of methodologies in these areas. Fields on the periphery of the network also seem to indicate new forms of engagement with social science data. For example, the field of Transportation and Freight Services uses data from "American Time Use Survey (ATUS): Arts Activities, [United States], 2003–2018," along with Environmental Science and Management, Economics, Religion and Religious Studies, Urban and Regional Planning, Information Systems, Library and Information Studies. Connections between fields on the periphery of each community subgraph appear to maintain weak ties among fields of research (Granovetter, 1973).

5.3. The Role of Research Data in Scientific Communities

These two structures suggest unique roles for data in scientific communities. Data sets in *subdivisions* and *crossroads* are two types of essential resources supporting social science research; subdivisions may have high disciplinary impact for the specific research domains that use them, while data sets at a crossroads may provide connectivity across domains. For instance, data at *crossroads* enable a kind of "arm's length" cooperative work where the work is loosely coupled, but depends on a "shared information space" that includes common data; much like community efforts to maintain taxonomies across time and space, the shared analysis of data sets contributes to cooperative "conceptual infrastructures" of scientific knowledge (Bannon & Schmidt, 1989, p. 361; Thomer et al., 2018).

While most ICPSR data is used by many disciplines within the bounds of social science, data reuse outside of the social sciences tends to engage with data in two main ways. First, fields such as statistics and artificial intelligence are central in the field of research cocitation network; these fields may reuse social science data to develop new research and analytic methods. Second, fields such as performing arts and creative writing are peripheral in the network; while they tend to reuse ICPSR's data less overall, they may provide novel inroads for "awakening" cross-disciplinary data reuse in new application areas (Hu & Rousseau, 2019).

Identifying hidden communities and their structures within the data citation graph helps us understand how data promotes knowledge production (Buneman et al., 2021; Lowenberg et al., 2019). It is likely that data sets occupying these different structures offer different types of "analytical potential." Palmer et al. (2011) describe "analytic potential" as "possible analytic contributions for the range of possible user communities" (p. 4), and our method exposes those possible communities and their structure. Research communities are beginning to recognize the importance of contributing to data resources, and the citation graph enables us to assign credit for different kinds of contributions (Alter & Gonzalez, 2018; Cousijn et al., 2019). Naming these different structures provides an accessible, extensible language for discussing the functions of data and assigning credit for their creation. Creating and sharing data that are used widely within one's discipline ought to afford researchers credit among their peers, sometimes for facilitating disciplinary depth—*subdivisions*—and at other times for creating multidisciplinary resources—*crossroads*.

5.4. Limitations and Outlook

Our analysis relied on a hand-curated resource, the ICPSR Bibliography of Data-Related Literature, which limits the generalizability of our findings. While other data archives also maintain bibliographies, few are as comprehensive and cross-cutting as ICPSR's, which is maintained by dedicated staff who capture instances of data reuse across a wide variety of media types and scientific disciplines. However, our network analysis method is generally applicable to study data reuse and highlights incentives for data archives to maintain comprehensive bibliographies, which support the long-term study of data impact.

We also used the Dimensions database's existing classification scheme for fields of research. This was a pragmatic choice given that codes were assigned at the level of publications rather than journals. However, the granularity of fields of research may be too coarse for interpreting finer disciplinary patterns of data use within domain archives. Adopting other domain analysis approaches could enhance our understanding of scientific knowledge production (Hjørland & Albrechtsen, 1995). In addition, we could compare the reuse of curated social science data at ICPSR to self-archived data (e.g., from the Dutch National Centre of Expertise and Repository for Research Data: DANS).

We were able to identify data sets that served different purposes within scientific communities, but our data do not allow us to comment on how credit for creating different types of data resources ought to be awarded to data creators and providers. Future research should examine the relationship between data creation, reputation, and careers to understand how to recognize data creators' contributions. Because of the different roles they play in connecting and supporting scholarly communities, data creators who produce *subdivisions* or *crossroads* likely deserve different types of credit for their contributions. For instance, creating a data set that operates as a subdivision should afford data creators substantial credit within their discipline, while creating crossroads may award creators a broader reputation that is less well-recognized within a single discipline. Data creators' academic careers depend on how they receive credit for their work and could impact the types of data resources they create and share.

Our data is essentially a snapshot in time, and they do not enable us to investigate the processes of community formation. The ICPSR Bibliography is a dynamic database; new citations are added continuously as they are discovered. The fact that a study does not have any citations, or has very few, does not mean that its data have never been used; rather, it may mean that any existing records of its use have not yet been discovered. More exhaustive searches for references to ICPSR data are under way. It is unclear whether data sets mentioned in literature (e.g., "Data from the American National Election Survey (ANES) is restricted in its geographic coverage but contains valuable direct questions on the subjective evaluation of racial groups ...") imply that the authors have analyzed the data or are mentioning the data for other purposes. Finer distinctions between types of data set references will enable future studies of factors that contribute to the analytical potential and end-users' decision to use data.

Given that data citation is a dynamic process, we are also interested in studying community formation to better understand how social ties, data curation, or other factors shape data citation networks. For example, temporal citation dynamics provide rich insights into the formation of research communities (Chubin, 1976). Extending the idea of "hibernation" to research data sets that have not yet been "awakened" through reuse (Hu & Rousseau, 2019) and detecting bursts of citations following long periods of dormancy would allow us to detect discovery events in the network. Understanding factors associated with novel data

reuse would provide evidence to recommend underutilized research data and prioritize funding and credit for specific data curation activities.

6. CONCLUSION

Data citation networks contain hidden information about communities of data users and the roles data play as primary inputs for scientific knowledge production. Through network analysis, we revealed these communities and identified 41 communities of social science data sets, along with four interdisciplinary research communities that use these data. Six important data series connect the cocitation network. Data sets that are used together exclusively form research *subdivisions*, which are valuable data collections for particular disciplines. Other data sets or fields that connect research communities are *crossroads* and have high topical or analytical versatility. Research data sets that are produced for different purposes, such as long-running series data and single-purpose study data, are often used together. Similar fields of research also tend to use the same combinations of data. In conclusion, these findings contribute new ways of seeing scientific communities and make the impacts of research data reuse visible.

ACKNOWLEDGMENTS

Many thanks to Elizabeth Moss and the ICPSR Bibliography staff (Homeyra Banaeefar, Sarah Burchart, and Eszter Palvolgyi-Polyak), David Bleckley, Elizabeth Yakel, Amy Pienta, and Dharma Akmon of the MICA team at the Inter-university Consortium for Political and Social Research (ICPSR) for their support of this research. We are also grateful to Sagar Kumar and Andrew Schrock for providing feedback on our earlier drafts.

AUTHOR CONTRIBUTIONS

Sara Lafia: Analysis and interpretation of data, Conceptualization, Methodology, Visualization, Writing—original draft, Writing—review & editing. Lizhou Fan: Analysis and interpretation of data, Methodology, Visualization, Writing—original draft. Andrea Thomer: Conceptualization, Funding acquisition, Supervision, Writing—original draft. Libby Hemphill: Conceptualization, Funding acquisition, Methodology, Supervision, Writing—original draft.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This material is based upon work supported by the National Science Foundation under grant 1930645.

DATA AVAILABILITY

Citation data were derived from the ICPSR Bibliography in February 2022. Code for this article's analysis is available in a GitHub repository (Lafia, 2022b) and data is available on openICPSR (Lafia, 2022a). Access to licensed metadata from Dimensions was granted to subscription-only data sources under a license agreement with Digital Science through the University of Michigan.

REFERENCES

Alter, G., & Gonzalez, R. (2018). Responsible practices for data sharing. *The American Psychologist*, *73*(2), 146–156. https://doi.org/10.1037/amp0000258, PubMed: 29481108

Bannon, L. J., & Schmidt, K. (1989). CSCW: Four characters in search of a context. ECSCW 1989. In *Proceedings of the First European Conference on Computer Supported Cooperative Work*. https://www-ihm.lri.fr/~mbl/ENS/CSCW/2012/papers/Bannon-ECSCW-89.pdf

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2018). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, *70*(8), 888–904. https://doi.org/10.1002/asi.24172

Bose, R., & Frew, J. (2005). Lineage retrieval for scientific data processing: A survey. *ACM Computing Surveys*, *37*(1), 1–28. https://doi.org/10.1145/1057977.1057978

Brown, C. (2003). The changing face of scientific discourse: Analysis of genomic and proteomic database usage and acceptance. *Journal of the American Society for Information Science and Technology*, *54*(10), 926–938. https://doi.org/10.1002/asi.10289

Buneman, P., Christie, G., Davies, J. A., Dimitrellou, R., Harding, S. D., ... Wu, Y. (2020). Why data citation isn't working, and what to do about it. *Database: The Journal of Biological Databases and Curation*, *2020*, baaa022. https://doi.org/10.1093/database/baaa022, PubMed: 32367113

Buneman, P., Dosso, D., Lissandrini, M., & Silvello, G. (2021). Data citation and the citation graph. *Quantitative Science Studies*, *2*(4), 1399–1422. https://doi.org/10.1162/qss_a_00166

Chen, C. (2017). Science mapping: A systematic review of the literature. *Journal of Data and Information Science*, *2*(2), 1–40. https://doi.org/10.1515/jdis-2017-0006

Chubin, D. E. (1976). State of the field the conceptualization of scientific specialties. *The Sociological Quarterly*, *17*(4), 448–476. https://doi.org/10.1111/j.1533-8525.1976.tb01715.x

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., & Simons, N. (2019). Bringing citations and usage metrics together to make data count. *Data Science Journal*, *18*(1), 9. https://doi.org/10.5334/dsj-2019-009

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., ... Clark, T. (2018). A data citation roadmap for scientific publishers. *Scientific Data*, *5*, 180259. https://doi.org/10.1038/sdata.2018.259, PubMed: 30457573

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023–4038. https://doi.org/10.1098/rsta.2010.0165, PubMed: 20679120

Crane, D. (1977). Social structure in a group of scientists: A test of the "invisible college" hypothesis. In *Social Networks* (pp. 161–178). Elsevier. https://doi.org/10.1016/B978-0-12-442450-0.50017-1

Cunningham, E., Smyth, B., & Greene, D. (2022). Navigating multidisciplinary research using field of study networks. In *Complex Networks & Their Applications X* (pp. 104–115). https://doi.org/10.1007/978-3-030-93409-5_10

Darden, L., & Maull, N. (1977). Interfield theories. *Philosophy of Science*, *44*(1), 43–64. https://doi.org/10.1086/288723

Data Citation Synthesis Group. (2014). *Joint declaration of data citation principles*. Force11. https://doi.org/10.25490/a97f-egyk

Fenlon, K. (2017). Thematic research collections: Libraries and the evolution of alternative digital publishing in the humanities. *Library Trends*, *65*(4), 523–539. https://doi.org/10.1353/lib.2017.0016

Fenner, M., Crosas, M., Grethe, J., Kennedy, D., Hermjakob, H., ... Clark, T. (2016). A data citation roadmap for scholarly data repositories. *Scientific Data*, *6*, 28. https://doi.org/10.1038/s41597-019-0031-8, PubMed: 30971690

Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, *659*, 1–44. https://doi.org/10.1016/j.physrep.2016.09.002

Franck, G. (1999). Scientific communication—A vanity fair? *Science*, *286*(5437), 53–55. https://doi.org/10.1126/science.286.5437.53

Gökalp, I. (1987). On the dynamics of controversies in a borderland scientific domain: The case of turbulent combustion. *Social Sciences Information*, *26*(3), 551–576. https://doi.org/10.1177/053901887026003005

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380. https://doi.org/10.1086/225469

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Lost or found? Discovering Data needed for research. *Harvard Data Science Review*, *2*(2). https://doi.org/10.1162/99608f92.e38165eb

Hagberg, A., Swart, P., & S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy)* (pp. 11–15). https://www.osti.gov/biblio/960616

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57*(2), 280–299. https://doi.org/10.1353/lib.0.0036

Hemphill, L., Pienta, A., Lafia, S., Akmon, D., & Bleckley, D. (2022). How do properties of data, their curation, and their funding relate to reuse? *Journal of the Association for Information Science and Technology*, *73*(10), 1432–1444. https://doi.org/10.1002/asi.24646

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: Data-intensive scientific discovery*. Microsoft Research.

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American Society for Information Science*, *46*(6), 400–425. https://doi.org/10.1002/(SICI)1097-4571(199507)46:6<400::AID-ASI2>3.0.CO;2-Y

Hook, D. W., Porter, S. J., & Herzog, C. (2018). Dimensions: Building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, *3*, 23. https://doi.org/10.3389/frma.2018.00023

Hu, X., & Rousseau, R. (2019). Do citation chimeras exist? The case of under-cited influential articles suffering delayed recognition. *Journal of the Association for Information Science and Technology*, *70*(5), 499–508. https://doi.org/10.1002/asi.24115

Jenks, G. F. (1963). Generalization in statistical mapping. *Annals of the Association of American Geographers*, *53*(1), 15–26. https://doi.org/10.1111/j.1467-8306.1963.tb00429.x

King, G. (1995). Replication, replication. *PS: Political Science & Politics*, *28*(3), 444–452. https://doi.org/10.2307/420301

Lafia, S. (2022a). *ICPSR Bibliography Citation Network (February 2022)* [Data set]. Inter-university Consortium for Political and Social Research (ICPSR).

Lafia, S. (2022b). *ICPSR/data-communities (Version v1.0.0)* [Computer software]. https://doi.org/10.5281/zenodo.6799127

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E*, *80*(5), 056117. https://doi.org/10.1103/PhysRevE.80.056117, PubMed: 20365053

Lee, J., & Jeng, W. (2019). The landscape of archived studies in a social science data infrastructure: Investigating the ICPSR metadata records. *Proceedings of the Association for Information Science and Technology*, *56*(1), 147–156. https://doi.org/10.1002/pra2.62

Leicht, E. A., Clarkson, G., Shedden, K., & Newman, M. E. J. (2007). Large-scale structure of time evolving citation networks. *European Physical Journal B*, *59*(1), 75–83. https://doi.org/10.1140/epjb/e2007-00271-7

Lowenberg, D., Chodacki, J., Fenner, M., Kemp, J., & Jones, M. B. (2019). Open data metrics: Lighting the fire. *Zenodo*. https://doi.org/10.5281/zenodo.3525349

Mayernik, M. S., Hart, D. L., Maull, K. E., & Weber, N. M. (2017). Assessing and tracing the outcomes and impact of research infrastructures. *Journal of the Association for Information Science and Technology*, *68*(6), 1341–1359. https://doi.org/10.1002/asi.23721

Moss, E., & Lyle, J. (2018). *Opaque data citation: Actual citation practice and its implication for tracking data use*. https://deepblue.lib.umich.edu/handle/2027.42/142393

National Academy of Sciences. (2005). *Facilitating interdisciplinary research*. National Academies Press. https://doi.org/10.17226/11153

Newman, M. E. J. (2003). Mixing patterns in networks. *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, *67*(2 Pt 2), 026126. https://doi.org/10.1103/PhysRevE.67.026126, PubMed: 12636767

Newman, M. E. J. (2004). Who is the best connected scientist? A study of scientific coauthorship networks. In E. Ben-Naim, H. Frauenfelder, & Z. Toroczkai (Eds.), *Complex networks* (pp. 337–370). Springer. https://doi.org/10.1007/978-3-540-44485-5_16

Orthia, L. A., McKinnon, M., Viana, J. N., & Walker, G. (2021). Reorienting science communication towards communities. *Journal of Science Communication*, *20*(03), A12. https://doi.org/10.22323/2.20030212

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*(7043), 814–818. https://doi.org/10.1038/nature03607, PubMed: 15944704

Palmer, C. L. (2004). Thematic research collections. In S. Susan, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities*. Blackwell. https://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-4-5&toc.depth=1&toc.id=ss1-4-5&brand=default. https://doi.org/10.1002/9780470999875.ch24

Palmer, C. L., Weber, N. M., & Cragin, M. H. (2011). The analytic potential of scientific data: Understanding re-use value. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–10. https://doi.org/10.1002/meet.2011.14504801174

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the reuse of scientific data. *Data Science Journal*, *16*, 8. https://doi.org/10.5334/dsj-2017-008

Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719. https://doi.org/10.1007/s11192-008-2197-2

Price, D. J. de Solla, & Beaver, D. (1966). Collaboration in an invisible college. *American Psychologist*, *21*(11), 1011–1018. https://doi.org/10.1037/h0024051, PubMed: 5921694

Robinson-Garcia, N., Mongeon, P., Jeng, W., & Costas, R. (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations. *Journal of Informetrics*, *11*(3), 841–854. https://doi.org/10.1016/j.joi.2017.07.003

Sands, A., Borgman, C. L., Wynholds, L., & Traweek, S. (2012). Follow the data: How astronomers use and reuse data. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–3. https://doi.org/10.1002/meet.14504901341

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269. https://doi.org/10.1002/asi.4630240406

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, "translations" and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science*, *19*(3), 387–420. https://doi.org/10.1177/030631289019003001

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., ... Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLOS ONE*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101, PubMed: 21738610

Thomer, A. K. (2022). Integrative data reuse at scientifically significant sites: Case studies at Yellowstone National Park and the La Brea Tar Pits. *Journal of the Association for Information Science and Technology*, *73*(8), 1155–1170. https://doi.org/10.1002/asi.24620

Thomer, A. K., Twidale, M. B., & Yoder, M. J. (2018). Transforming taxonomic interfaces. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–23. https://doi.org/10.1145/3274442

Tomasello, M. V., Vaccario, G., & Schweitzer, F. (2017). Data-driven modeling of collaboration networks: A cross-domain analysis. *EPJ Data Science*, *6*(1), 22. https://doi.org/10.1140/epjds/s13688-017-0117-5

Varga, A. (2019). Shorter distances between papers over time are due to more cross-field references and increased citation rate to higher-impact papers. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(44), 22094–22099. https://doi.org/10.1073/pnas.1905819116, PubMed: 31611374

Vertesi, J., & Dourish, P. (2011). The value of data: considering the context of production in data economies. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (pp. 533–542). https://doi.org/10.1145/1958824.1958906

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the Association for Information Science and Technology*, *49*(4), 327–355. https://doi.org/10.1002/(SICI)1097-4571(19980401)49:4<327::AID-ASI4>3.0.CO;2-W

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18, PubMed: 26978244

Yang, Z., Algesheimer, R., & Tessone, C. J. (2017). A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, *7*, 46845. https://doi.org/10.1038/srep46845, PubMed: 28650447

Zeng, T., Wu, L., Bratt, S., & Acuna, D. E. (2020). Assigning credit to scientific datasets using article citation networks. *Journal of Informetrics*, *14*(2), 101013. https://doi.org/10.1016/j.joi.2020.101013

Zimmerman, A. S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology, & Human Values*, *33*(5), 631–652. https://doi.org/10.1177/0162243907306704