Extracting the interdisciplinary specialty structures in social media data-based research
A clustering-based network approach

Fan, Yangliu; Lehmann, Sune; Blok, Anders; Fan, Yangliu

Published in: Journal of Informetrics

Link to article, DOI: 10.1016/j.joi.2022.101310

Publication date: 2022

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Extracting the interdisciplinary specialty structures in social media data-based research: A clustering-based network approach

Yangliu Fan\textsuperscript{a,\ast}, Sune Lehmann\textsuperscript{a,b}, Anders Blok\textsuperscript{a,c}

\textsuperscript{a}Copenhagen Center for Social Data Science, University of Copenhagen, Copenhagen, Denmark
\textsuperscript{b}DTU Compute, Technical University of Denmark, Lyngby, Denmark
\textsuperscript{c}Department of Sociology, University of Copenhagen, Copenhagen, Denmark

\begin{abstract}
As science is becoming more interdisciplinary and potentially more data driven over time, it is important to investigate the changing specialty structures and the emerging intellectual patterns of research fields and domains. By employing a clustering-based network approach, we map the contours of a novel interdisciplinary domain – research using social media data – and analyze how the specialty structures and intellectual contributions are organized and evolve. We construct and validate a large-scale (\textit{N} = 12,732) dataset of research papers using social media data from the Web of Science (WoS) database, complementing it with citation relationships from the Microsoft Academic Graph (MAG) database. We conduct cluster analyses in three types of citation-based empirical networks and compare the observed features with those generated by null network models. Overall, we find three core thematic research subfields – interdisciplinary socio-cultural sciences, health sciences, and geo-informatics – that designate the main epicenter of research interests recognized by this domain itself. Nevertheless, at the global topological level of all networks, we observe an increasingly interdisciplinary trend over the years, fueled by publications not only from core fields such as \textit{communication} and \textit{computer science}, but also from a wide variety of fields in the social sciences, natural sciences, and technology. Our results characterize the specialty structures of this domain at a time of growing emphasis on big social data, and we discuss the implications for indicating interdisciplinarity.
\end{abstract}

1. Introduction

According to Small and Griffith (1974), “science is a mosaic of specialties and not a unified whole, either socially or intellectually.” Mapping the specialty structure of scientific publications has received much attention in the examination of disciplinary and emerging interdisciplinary fields (Cobo et al., 2011; Tang et al., 2017; Edelmann et al., 2020). In particular, understanding the specialty structure and its dynamic evolution can shed light on the development of a scientific field, the interrelationships among subfields, and the process of knowledge integration (Small, 1973; Moody, 2004). Graph partitioning and community detection methods have been widely adopted to systematically explore specialty structures and their relationships to one another. This trend dates back to the 1970s when early work (Small & Griffith, 1974) used the computer-based technique to identify clusters as specialties of scientific literature. The recent development of online bibliographic databases and computational algorithms allows us to extract community structures from networks based on a large number of publications. This development offers exciting new perspectives and opportunities for understanding the intellectual connections and divisions within the dynamic institution of science.

\textsuperscript{\ast} Corresponding author.

E-mail address: yangliufan@sodas.ku.dk (Y. Fan).

https://doi.org/10.1016/j.joi.2022.101310
Received 3 November 2021; Received in revised form 28 April 2022; Accepted 7 June 2022
1751-1577/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license
(http://creativecommons.org/licenses/by/4.0/)
In this paper, we extract the specialty structures of an increasingly important domain of science by employing original bibliometric data. Specifically, we map the developing contours of (inter)disciplinary communication in research using social media data, drawing on a carefully delineated dataset of 12,732 research articles from the Web of Science (WoS) between 2005 and 2019. We identify and compare intellectual patterns in this domain from three types of citation-based networks – the bibliographic coupling network, as well as the internal and external co-citation networks, i.e., the within-population core citing papers and the complete set of citing papers across academia. Uniquely relevant to existing research, this approach allows us to illuminate the research domain in a holistic and comprehensive way, providing a solid and nuanced view of its specialty structures. By doing so, we contribute to a deeper understanding of this domain, since previous studies were mostly confined to reviewing research on particular social media platforms or big social data (Gupta & Dhami, 2015; Zhang & Leung, 2015; Williams et al., 2017; Zyoud et al., 2018; Esfahani et al., 2019).

Given the importance of specialty structures in science mapping, it is not surprising to see increased attention to cluster analysis of bibliometric networks based on relatedness measures. Such clustering is commonly based on overlapping received citations, i.e., co-citation (CC) (Small, 1973), or references, i.e., bibliographic coupling (BC) (Kessler, 1963). Scholars (Boyack & Klavans, 2010; Tang et al., 2017) have carried out cluster analyses using data from large citation indexing databases such as Web of Science. In order to study the co-citation patterns of the source articles, it is important to conduct pair-wise matching of the source articles’ received citations. Most co-citation-based network studies (Chi & Young, 2013), however, fail to collect received citations, focusing only on the co-citation patterns of the cited references. This is partly due to a lack of data availability in most bibliometric databases. Yet Microsoft Academic Graph (MAG) database (Sinha et al., 2015; Wang et al., 2019) enables researchers to trace pair-wise citation relationships between publications in a large knowledge graph. The availability of citation relationships in this new database allows us to conduct additional investigations into document coupling of the source articles.

Besides, in order to measure the differently sized clusters in the bibliometric networks, scholars (Chi & Young, 2013; Dawson et al., 2014; Tang et al., 2017) use network topological quantities, including clustering coefficient and modularity. They also add attributes to the nodes, including predefined disciplinary categories such as Web-of-Science Subject Categories (WC) or content analysis-based categories such as research methods. Most cluster analyses in bibliometrics, however, are limited to the empirical descriptions of the categories covered by a cluster while failing to establish the statistical significance of the observation. In recent bibliometric studies, the null model has been adopted, facilitating a more rigorous statistical description of bibliometric networks (Li et al., 2019). In this study, we adopt a null model to assess the statistical significance of the empirically observed patterns of clusters by randomly shuffling node attributes while maintaining edges. By this means, we try to determine whether detected clusters are associated with particular node attributes (see Methods).

We define our dataset, i.e., social media data-based research, as a knowledge domain, which we believe is a particularly relevant case for enhancing cluster analyses to extract specialty structures. Unlike relatively coherent specialty fields such as communication research or digital humanities, research that utilizes social media data does not constitute a formally recognized discipline, set of disciplines, nor indeed a well-rounded interdisciplinary field. The collected dataset is diverse and interdisciplinary in scope. Indeed, it spans all the main areas of science, including computer science and communication, as well as emerging interdisciplinary fields such as computational social science and digital humanities (Kitchin, 2014; Edelmann et al., 2020). Drawing on the conceptualization in the literature (Nascimento & Marteleto, 2008; Sugimoto & Weingart, 2015), we consider relevant research as forming a "domain" of knowledge. Relative to a more coherent "field" or "discipline" based on tighter discourse communities (as Sugimoto and Weingart suggest), a domain of knowledge is more heterogeneous, consisting of more diverse informational practices. Still, while integration is weaker, the domain's research products and subjects are governed by shared forms of attention and convergent practices that emerge at the meeting-point of various disciplines and fields (as in Nascimento and Marteleto's model). The resultant specialty structures of the domain are likely to be variously interdisciplinary. Yet, their exact composition at the community level is hard to predict and forms the main object of study in the present article.

Our study presents a novel clustering-based network approach to map out the developing specialty structures of the domain of social media data-based research. This domain provides a particularly suitable setting to apply our approach, since it is characterized by the statistical heterogeneity of citation patterns – the degree and weight distributions across multiscale – that makes extracting the interdisciplinary specialty structures difficult. This application allows us to discuss the implications of the (inter)disciplinary composition of clusters in bibliometric networks for indicating interdisciplinarity. Although our substantive focus is on a specific knowledge domain, our approach can be directly applied to a broader context, such as other (inter)disciplinary research fields and domains, and to the wider institution of science, where the identification of clusters and within-cluster heterogeneity are of interest.

2. Data

In this section, we present our search query along with the procedure through which we retrieve data from bibliometric databases in Sect. 2.1, the predefined disciplinary categories for articles in Sect. 2.2, and several tests for data quality to validate our dataset in Sect. 2.3.

2.1. Data collection

In this study, we combined bibliometric data from two large databases: Web of Science (WoS) and Microsoft Academic Graph (MAG). We did so in order to retrieve papers that meet the definition of our domain of study – i.e., research that makes active use of social media-derived data – as well as to get relevant bibliographic information on the complete set of papers for network construction. As shown in Table 1, we combined three separate queries in ways that together ensure a comprehensive operationalization of the
Table 1
Search strategy for social media data-based papers in WoS.

<table>
<thead>
<tr>
<th>Queries (TS refers to search in title, keywords, and abstract fields)</th>
<th>Number of Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1 TS=&quot;(social media data)&quot;</td>
<td>987</td>
</tr>
<tr>
<td>#2 TS=&quot;(social media&quot; OR 'social network' site&quot; OR 'online social network&quot; OR facebook OR twitter OR instagram OR blog OR microblog OR YouTube OR Flickr) AND TS=(tweet OR hashtag OR weblink OR posting OR commenting OR 'news feed' OR user profile OR geotag OR YouTube video OR 'digital data' OR 'computational social' OR 'digital social' OR 'social web data') NOT TS=(survey OR questionnaire OR interview)</td>
<td>11,377</td>
</tr>
<tr>
<td>#3 TS=&quot;facebook OR twitter OR instagram OR blog OR microblog OR YouTube OR Flickr) AND TS=(dataset)</td>
<td>1659</td>
</tr>
<tr>
<td>Combine all queries #1 OR #2 OR #3</td>
<td>12,732</td>
</tr>
</tbody>
</table>

Fig. 1. Number of social media data publications by year. Here the gray vertical bars represent the number of publications in each year. It can be seen that the number of relevant papers has been significantly increasing from 2005 to 2019, which indicates the rapid growth of this domain.

domain of social media data-based research. Importantly, building on our analytical notion of this emerging research domain, we have taken utmost care to ensure that the sample represents research that deploys and utilizes social media data, as opposed to merely mentioning or abstractly discussing it.

Specifically, we began by searching for English journal articles in WoS mentioning the term "social media data" in the title, abstract, or keywords in SCI-EXPANDED, SSCI, AHCI, and ESCI between 2005 and 2019. The year 2005 was chosen as the starting point since relevant literature published before then was scant. Second, because such data often appear through terms denoting the specific social media contents or outputs, we also searched for the most prevalent social media platforms jointly with data contents, such as 'tweet' (including variants and plural forms), to increase our coverage. Note that some conventional data-related terms such as 'survey,' "questionnaire," and "interview" were explicitly filtered out in this second query. In this way, we increased the total recall while considering the expense of reducing the precision. Third, to further improve the width of our sample, we included papers that mention these platforms combined with the term "dataset." Our search returned 12,732 articles in total (see Fig. 1), including 11,951 articles with Digital Object Identifier (DOI).

We first downloaded the basic bibliographic data, including title, author(s), source, year, abstracts, keywords, Web-of-Science Subject Categories (WC), and references from the WoS database, forming our core dataset. We then conducted pair-wise matching of DOIs of the core dataset in the Microsoft Academic Graph (MAG) database; this allowed us to form an extended dataset, which incorporates title, author(s), references, and, in particular, all the core articles' received citations, i.e., all the articles that cite them. Given that the MAG assigns each author a unique ID, author name disambiguation was facilitated by the MAG AuthorId. The top level of MAG Field of Study (MF) of the authors’ first publications was used as the proxy for their disciplinary backgrounds. Overall in our dataset, 11,736 (98%) papers and 29,458 authors are identified in the MAG.

2.2. Predefined disciplinary categories

To study the specialty structure of the dataset, we need to assign additional attributes to the articles, i.e., a predefined disciplinary category. Currently, there are two central discipline- and field-taxonomy systems for research literature, grounded in journal-based
and paper-based classification, respectively. Web-of-Science Subject Categories (WC) is journal-based, assigning the journals it indexes into categories by machine and manual corrections using journal content and citation information (Boyack et al., 2005; Leydesdorff & Bormann, 2016). By contrast, the Microsoft Academic Graph Field of Study (MF) is article-based; it is generated by a natural language processing (NLP) technique and hierarchically organized into six levels, a procedure that involves various limitations (Hug & Brändle, 2017). Although WC’s assignment is not perfect, it is the most widely adopted and, in particular, it spans all disciplines (Rinia et al., 2001). Moreover, scholars (Rafols & Leydesdorff, 2009) claim that the relationships among the WC are statistically reliable, maintaining that the classification can be applied to statistical mapping at the global level despite its occasional imprecision in the assignment of individual journals to subject categories. Thus, the WC is adopted as the primary operationalization of "discipline" and "field" in this research. The MF is employed as the substitute when the WC is not available, in our case, for the approximation of author disciplinary backgrounds.

2.3. Data validity check

To evaluate the relevance and validity of retrieved data, we randomly sampled 100 papers in our dataset, manually read the abstracts of the papers, and found 89 that met our eligibility criteria. Specifically, our criteria required the papers to have carried out empirical research practices utilizing data generated from social media sources. In turn, working in the opposite direction, we randomly discovered 100 journal papers from different data sources that met our criteria. After matching the DOIs of these papers, we found 89 of them in our dataset1. We take this to imply that, while our sampling strategy suffers from minor false positive and false negative errors, the resulting dataset can be taken as a valid approximation of the research domain under study. This holds, we suggest, despite the fact that individual articles have not been reviewed by human coders to confirm their relevance.

Further, although citation data from the MAG seems to be more available, systematic, and detailed, previous research (Visser et al., 2021) has revealed that the WoS has better coverage and a higher quality of citation links in terms of completeness and accuracy. In this regard, we conducted a pair-wise comparison of within-dataset citations between the two versions (WoS-based, MAG-based) of our dataset. It showed that the overlap is considerable, despite that 5% of citation links in MAG cannot be obtained from WoS and 13% in WoS cannot be found in MAG. A manual examination of 20 randomly selected citation links in the above two cases showed that the two databases had been incorrect not identifying these links. Therefore, in the end, we combined the two datasets to produce a more comprehensive set of citation links, incorporating the references from both WoS and MAG, and the received citations from the MAG dataset. The total number of 11,951 papers from the core dataset, 29,458 authors from the extended dataset, and 543,809 citation links from the merged dataset were retained in our analysis.

3. Methods

In this section, we first employ the concept of "interdisciplinarity" and introduce our approach to operationalization. We then go on to present our clustering-based network approach in detail, illustrating the applications of algorithms to particular problems in understanding the topological features of networks. Our approach comprises the following steps: network construction, network filtering, community detection, and null model tests.

3.1. Interdisciplinarity measures

The question of how to measure interdisciplinarity is widely discussed in the literature. Many have used diversity indicators (Stirling, 2007; Zhang et al., 2016; Leydesdorff & Ivanova, 2020) and citation-based indicators (Carley & Porter, 2011; Leydesdorff et al., 2018). Others (Rafols, 2020; Wang & Schneider, 2020) have pointed out that regardless of the variety of current measures, the uncertainty remains because of the instability and inconsistency of existing indicators and the fluidity of the concept of "interdisciplinarity". Notwithstanding the complexities, we can still measure interdisciplinarity under particular contexts and understandings. In our operationalization, we drew on Rafols and Meyer’s (2010) idea that "interdisciplinarity" can be measured through the process of knowledge integration by exploring disciplinary diversity from predefined categories of a bibliometric set and the coherence of the network generated. To explore the interdisciplinary features of this domain, we specifically extract the heterogeneous specialty structures using a network approach. In this way, we consider interdisciplinarity through knowledge integration as a systemic process in which different research communities engage in a conversation.

3.2. Network construction

Overall in this study, we used two types of document coupling techniques for network construction: bibliographic coupling (BC) and co-citation (CC). Bibliographic coupling connects two documents by matching their references, while co-citation represents a link between two documents when they are cited together by other documents (Kessler, 1963; Small, 1973; Garfield, 2001). We also analyzed co-authorship networks, although we consider these supplementary to the present analysis (see Supplementary Information

---

1 To further test the coverage of our dataset, we randomly sampled the names of ten scholars from our dataset who presented publications in computer science and another ten scholars in communication. We searched for all of their publications and identified research outputs that met our definition of this domain. Results showed that more than a third of these publications were included in our datasets, 20/52 in computer science and 15/40 in communication, respectively. What our dataset had missed primarily were conference papers, book chapters, and low-tier journal articles.
<table>
<thead>
<tr>
<th>Citation-based mapping approaches</th>
<th>Clustering example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal co-citation</td>
<td><img src="https://example.com/diagram" alt="Diagram" /></td>
</tr>
<tr>
<td></td>
<td>A B C D E F G H I</td>
</tr>
<tr>
<td>External co-citation</td>
<td><img src="https://example.com/diagram" alt="Diagram" /></td>
</tr>
<tr>
<td></td>
<td>A B C D E F G H I</td>
</tr>
</tbody>
</table>

Fig. 2. Example of how internal and external co-citation approaches partition the same set of documents (inspired by Boyack and Klavans, 2010). The gray box represents the documents within a dataset, i.e., documents A-I. Documents N-R are those documents outside the set but cite documents within the set. Solid arrows represent citations. Ovals in each panel illustrate how the documents might be clustered by each approach (depending on the clustering algorithm). In this example, documents within the dataset form our core domain, and others equate to adjacent research fields. In the case of internal co-citation, when only within-set links are used, the co-citation cluster will form a single cluster (F, G). If the process is expanded to external co-citation analysis, with more documents in the adjacent fields that cite our core domain, the co-citation cluster will constitute three clusters (A, F, G), (C, D, H), and (E, I).

Sect. S1). Using a simple time-slice approach inspired by techniques from social data science (Lehmann, 2019), we are able to describe key changes in the networks over time. A sliding window of five years was applied to each type of network in which we aggregate all links occurring within that time. Starting from 2005, we extended the window of five years to account for the continuing effects and obtained a total of three time slices (i.e., 2005-2009, 2005-2014, and 2005-2019) representing the evolution of the networks. In order to capture the dynamics of the clustering structure, we slid the window in five-year increments and obtained three more five-year non-overlapping time slices (i.e., 2005-2009, 2010-2014, and 2015-2019). Note that the temporal sub-networks only apply to the co-citation network (not the bibliographic coupling network). Since there has been a steep increase in the number of publications over time (see Fig. 1), thereby overwhelming the structure of the networks in the first period, we used this dynamic network construction to investigate how articles that are published in T1 (2005-2009) are co-cited across three periods: from T1 to T2 (2010-2014) to T3 (2015-2019).

In what follows, we outline also a more novel distinction between the internal co-citation and the external co-citation networks, referring to the former as the within-core-set (within-population) citing papers and the latter as the full or complete set of citing papers (see Fig. 2). The main reason is that drawing on the global topology of these two networks, along with the bibliographic coupling network, reveals the intellectual structure of this domain from different perspectives. These include the knowledge source (BC) and audience (CC) from both the core domain (internal CC) and adjacent fields (external CC).

More formally, we illustrate with the following setting. Consider a paper $p_0$ from our dataset $P = \{ p_0, p_1, \ldots, p_n \}$. We identify all documents in its reference, forming a set $D = \{ d_1, d_2, \ldots, d_l \}$, and all papers that cite $p_0$, forming a set $C = \{ c_1, c_2, \ldots, c_k \}$. For co-citation networks, we first identify the within-set citing papers $C_{\text{int}}$, representing papers in both C and P, and further within-set co-cited papers $C_{\text{adj}}$ representing the set of papers in P that are cited by papers in $C_{\text{int}}$. Likewise, we identify the complete set of papers in C, as well as the co-cited papers CC representing the papers in P that are co-cited by C. We then connect $p_0$ and the papers in $C_{\text{int}}$ and CC to build the within-set internal and the out-of-corpus external co-citation links; in these, $t$, the time when the link is formed, is based on the publication year of the papers in C, and disciplinary backgrounds are based on the WC. For the bibliographic coupling network, we connect two papers, $p_0$ and $p_1$. We do so if their bibliographies $D_0$ and $D_1$ have at least one common document, in which $t$ is the publication year of the paper published later, and the disciplinary backgrounds are from the WC.

### 3.3. Network filtering

For all the networks, both overall and time-sliced, link strength is defined as the number of times that nodes are connected (i.e., weight $w$). After building the weighted version of the networks, we applied thresholding methods to reduce the number of edges in our...
networks that are otherwise too dense for further analyses. In order to reduce the number of edges while extracting the truly relevant connections in the citation-based networks, we used the disparity filter algorithm (Serrano et al., 2009). The backbone structure preserved in the networks will be those edges whose weights satisfy the relation:

\[ a_{ij} = 1 - (k - 1) \int_0^{p} (1 - x)^{k-2}dx < \alpha \]  

(1)

where \( k \) is the degree of the node and \( p_{ij} \) is the normalized weight. This filtering algorithm has been proven to be able to preserve weights at all scales that are locally statistically significant with respect to the null model. Robustness checks have confirmed that the disparity filter outperformed the global threshold in our empirical networks (see Supplementary Information Sect. S2).

To find the proper reduction level, we compared the filtered subgraphs using the disparity filter for different values of the significance level \( \alpha \), in terms of different topological quantities, including the fraction of total weights (%Weight), nodes (%Nodes), edges (%Edges) preserved, average clustering coefficient (ACC), and the weight and degree distribution. There are no standard criteria for deciding the optimal \( \alpha \) (Serrano et al., 2009). We identified the optimal \( \alpha \) for each network in the sense of maintaining the majority of nodes and ACC of the original network while reducing a large proportion of edges. Fig. 3 shows that ACC for subgraphs is relatively stable for different reduction levels. Figs. 4 and 5 display that the disparity filter reveals a clearer heavy-tailed degree distribution as \( \alpha \) decreases; \( \alpha \) in the range [0.1, 0.3] is able to maintain a power-law weight distribution. We do not show the distribution when \( \alpha \) is below 0.1, in which the filter is too restrictive that the number of nodes and edges decreases significantly. After comparing the topology at different reduction levels, we conclude the values of \( \alpha \) for filtering our empirical networks, in which \( \alpha = 0.3 \) for both internal and external co-citation networks and \( \alpha = 0.1 \) for bibliographic coupling networks. The reduction can keep the majority of nodes, the same ACC of the original network, and stable degree and weight distribution while retaining a minimal number of edges (see Supplementary Sect. S3 Table S2 for network statistics).

3.4. Community detection

After filtering, we used the Infomap algorithm for community detection based on minimizing the map equation (Rosvall et al., 2009),

\[ L(M) = q_n H(Q) + \sum_{i=1}^{m} p'_i H(P^i) \]  

(2)

where \( L(M) \) is the code length, \( M \) is a given partition, \( H(Q) \) is the average length of codewords in the index codebook, and \( H(P^i) \) is the average length of codewords in the module codebook \( i \). The two terms are weighted by \( q_n \), i.e., the probability of switching modules, and \( p'_i \), i.e., the fraction of time spent in module \( i \) plus the probability of exiting it. The map equation is flow-based and information-theoretic, which specifies the theoretical limit for describing the trajectory of a random walker based on an arbitrary partition (Rosvall et al., 2009; Edler et al., 2017). This community detection algorithm is shown to outperform other methods in a comparative analysis of clustering scientific publications (Šubelj et al., 2016).

Both the weighted and unweighted filtered networks were submitted to the Infomap algorithm, and the results were evaluated using network modularity (Newman & Girvan, 2003),

\[ Q = \sum_{i} \left( e_{ii} - \frac{(\sum_i e_{ij})^2}{\sum_j e_{jj}} \right) \]  

(3)

where \( e_{ij} \) is the fraction of edges connecting to nodes in the same community, and \( \sum_i e_{ij} \) is the fraction of the edges connecting to community \( i \). A high \( Q \) value indicates a highly modular network structure – divisions of network nodes into densely connected subgroups (Newman & Girvan, 2003). Fig. 6 shows that all partitions exhibit a highly modular structure with modularity values around 0.6. Note that the Infomap algorithm shows some degree of instability that may result in a slight change in partitions and corresponding modularity. The modularity for unweighted networks is generally higher than that of the weighted networks. Therefore, we used the unweighted version of the filtered networks for community detection to reduce the effects of different citation behaviors among disciplines and obtain high network modularity. More details about the structural statistics for the partitions are laid out in Supplementary Table S3.

3.5. Null model tests

To establish the statistical significance of our observations, we use the following null models to compare with the observed indicators. In the literature of network science (Barabási, 2016), the null model is identical to the observed data in some key structural properties but is otherwise random. Conventionally, the degree-preserving random rewiring network is employed as the configuration null model to test if a specific aspect of interest is random. It keeps fixed the degree sequence of the networks while randomizing all other structures. It can tell us whether the empirical properties are meaningful or just a common consequence of the particular degree sequence.

In this study, our specific focus is to assess the statistical significance of specialties identified in a set of citation-based networks, i.e., the disciplinary categories covered by clusters. We use a null model variant, in which we randomly permute the categorical node attributes (i.e., disciplinary categories). Our randomization preserves the network structure, the corresponding community structure, and the overall distribution of node attributes while removing the systematic clustering of the categories. This reference
Fig. 3. Illustration of the fraction of weights (%W_T), nodes (%N_T), average clustering coefficient (ACC) as a function of the fraction of edges (%E_T) in the graph. (a) Internal co-citation network. (b) External co-citation network. (c) Bibliographic coupling network.
Fig. 4. Degree distribution of the filtered subgraphs for the empirical networks. (a) Internal co-citation network. (b) External co-citation network. (c) Bibliographic coupling network.
Fig. 5. Weight distribution of the filtered subgraphs for the empirical networks. (a) Internal co-citation network. (b) External co-citation network. (c) Bibliographic coupling network.
model provides an expected distribution for each category across clusters. Comparing the observed value with the 1000 randomized realizations of the network can tell us the extent to which a pattern occurs more than expected. To quantify the comparison, we estimate the Z-score given by:

$$Z = (X_D - \mu(X_R))/\sigma(X_R)$$

(4)

where $X_D$ is an indicator in the real data, and $X_R$ is the same indicator defined for the null model networks, with $\mu(X_R)$, $\sigma(X_R)$ indicating the mean and standard deviation for the null model networks, respectively. The Z-score is expected to be zero if the null hypothesis is true (i.e., the properties can be expected as a consequence of some constraints). All of the analyses presented in this study used packages and functions of Python, including NetworkX for network analysis, Matplotlib and Plotly for visualization, and Infomap for community detection.

4. Results

4.1. The disciplinary diversity in the social media data research domain

Before conducting the network analyses, we first provide an initial overview of the disciplinary diversity in this domain in terms of descriptive statistics. A total of 162 different disciplinary (and field) categories are found for papers in the dataset based on the WC. This indicates a great variety of disciplines (and fields) in this domain. Fig. 7 shows the distribution of publications over time across the ten most prevalent categories. It is easily seen that amidst the overall growth observed in Fig. 7, disciplines and fields are very far from evenly distributed. The use of social media-derived data is most active in the fields of computer science, communication, psychology, business, multidisciplinary sciences, and health care sciences and services, at least when relying on the (WC-based) publication and journal level. The author-based review and main co-authorship patterns of this domain are discussed in the Supplementary Information Sect. S1. Broadly speaking, the patterns corroborate the above pattern of main active fields.

4.2. Global structures and overall cohesion for citation-based networks

The three types of overall citation-based networks are built on and further serve to cluster different subsets derived from the same dataset (see Supplementary Table S2 for network statistics). As one might expect, overall the degree distribution in these networks is fat-tailed, with fewer heavily connected nodes and most of the nodes having only a few links. Likewise, community size distribution is fat-tailed, with a few giant clusters and a large number of small clusters. The clustering coefficient, the percentage of nodes remaining in the largest component, and the average path length have been used commonly in cohesion analysis – a form of analysis that evaluates the extent to which authors, papers, or journals are interconnected in the context of bibliometrics (Tang et al., 2017). In all our networks, the percentage of nodes in the largest connected component is significantly high (> 90%), indicating highly interconnected network structures. According to the Watts-Strogatz archetypical network model (Barabási, 2016), a short average path length is coupled with a high clustering coefficient in the small-world network model. In this regard, the average clustering is significantly large, and the giant components have an average shortest path length in the range of 3.79-4.06. The small-world property was tested using the configuration null model. Results demonstrate that both of the measures are significantly larger than those of the null model ($p < .0001$) – notably, the longer than expected average path length implies the property of high density within clusters. This is supported by the structural statistics (see Supplementary Table S3), which shows the large internal degree $K$,
the small external expansion $E$, and the low flake $F$. Hence, the networks exhibit very dense clusters connected by a small number of links.

4.3. Disciplinary compositions in the citation-based networks

We turn now to the central part of our analysis, which focuses on the disciplinary composition in the citation-based networks for our domain. As noted, all networks exhibit a remarkably diverse nature in terms of disciplines and fields present. Fig. 8 illustrates the discipline distribution across the top ten clusters in co-citation and bibliographic coupling networks2. More details are included in the Supplementary Information Sect. S4, Table S5-7, where we also show a complete list of disciplines in each cluster.

Three large communities or clusters can be identified at the core of the internal co-citation network (Fig. 8a). Notably, computer science is prominent in all clusters. Closer inspection of the three largest clusters reveals that they span various WC-based fields but exhibit a different central theme of interest. Cluster 1 is the largest and considerably interdisciplinary cluster that connects a broad range of fields, including some that are intrinsically interdisciplinary. For instance, the most notable number of papers is from communication (Z-score is 15.213, $p < .0001$), which in the WC sense spans journals belonging to many disciplines in the space between the humanities and the social sciences (Montero-Díaz et al., 2018). The cluster also includes a significant number of journal articles in psychology (6.264, $p < .0001$), business (10.338, $p < .0001$), hospitality (5.891, $p < .0001$), and political science (5.006, $p < .0001$). The number of papers in computer science (-1.899, $p > .5$) is sizable but not statistically significant for this first cluster.

Cluster 2 in the internal co-citation network (Fig. 8a) contains papers predominantly from health care sciences and services (13.892, $p < .0001$), public, environmental, and occupational health (11.895, $p < .0001$), and substance abuse (11.888, $p < .0001$); while cluster 3 comprises papers mostly from geography (17.254, $p < .0001$), environmental sciences (8.417, $p < .0001$), geosciences (7.971, $p < .0001$), and computer science (6.697, $p < .0001$). Given the network’s modular structure, these clusters with different disciplinary compositions seem to form around distinct thematic communities, consisting of several thematically related (inter)disciplinary fields in conversation. Given the statistically significant (WC-based) research fields identified in these thematic clusters, we may label these thematic clusters as the interdisciplinary socio-cultural sciences (1), health sciences (2), and geo-informatics (3), respectively. This indicates that when articles in this domain are being co-cited by later ones in the same domain, the community structure is rather discipline-spanning, yet thematically bounded.

The external co-citation network (Fig. 8b) shows a somewhat different configuration, indicating that the way research in this domain is picked up by the wider academic world (external) differs importantly from the way the domain itself is organized (internal). Each cluster in the external co-citation network has a generally predominant field of study, and there are more large-size clusters. Together, then, these amount to a distinctly more discipline- and field-based community structure. For example, health-related WC-based fields, including health care sciences and services (17.143, $p < .0001$), public, environmental, and occupational health (14.159, $p < .0001$), and substance abuse (5.521, $p < .0001$), are in cluster 1; communication (18.922, $p < .0001$) and political science (21.143, $p$
Fig. 8. The disciplinary composition of ten clusters in co-citation and bibliographic coupling networks based on WC. The graphs to the left with the gray bars show the number of papers across clusters; the graphs to the right with the colorful stacked bars show the proportion of fields in each cluster; to the extreme right is the legend that describes the WC-defined field indicated by the colored bands in each bar.
<.0001) in cluster 2; business (25.545, p < .0001) and hospitality (25.566, p < .0001) in cluster 3; psychology (25.007, p < .0001) in cluster 5; and computer science (15.778, p < .0001) in cluster 7.

Similar to the external co-citation network, the bibliographic coupling network (Fig. 8c) exhibits a mainly discipline- and field-based cluster structure. Computer science papers are markedly present in most clusters, such as 2, 3, 5, 6 (p < .0001). Cluster 1 is primarily comprised of work in business (42, p < .0001) and hospitality (9.827, p < .0001). Clusters 4 and 7 are based on socio-cultural sciences with interests in psychology (31.348, p < .0001) and communication (18.515, p < .0001), respectively. As noted before, the bibliographic coupling network originates in a different selection of documents from that of the co-citation network, and these two approaches show different relations. Thus, the patterns in the bibliographic coupling network demonstrate a different, largely discipline- and field-based intellectual structure in terms of the knowledge sources used by (cited by) scholars who deploy social media data for their research (i.e., who are active in the domain we study).

To compare the patterns found in the external co-citation network with those in the internal one, we make an alluvial diagram (Rosvall & Bergstrom, 2010) to illustrate the differences, with internal co-citation clusters (IC) on the left side and external co-citation clusters (EC) on the right (Fig. 9). The diagram illustrates the community-structural change from the IC to the EC through mergers or divergences in the stream (i.e., clustered papers) linking the blocks (see Supplementary Table S5-6). From the diagram, it can be seen that the concentrated thematic-based community structure, mentioned previously, is becoming more dispersed and field-based as we move outside the target domain (i.e., research actively using social media data). For instance, the health sciences thematic cluster (IC2) is split and becomes a subset of the health-based field cluster (EC1) and many others; and the interdisciplinary socio-cultural sciences thematic cluster (IC1) is divided into several subsets, encompassing the discipline- and field-based clusters communication-political science (EC2), business-hospitality (EC3), psychology (EC5), and so forth. As noted, the patterns suggest that as adjacent scholars formally recognize research from the domain of social media data-based research (by way of citations), this domain largely disperses into an existing discipline- and field-based community structure.

4.4. Evolution of citation-based networks in the domain

4.4.1. Cumulative network

As the number of publications in this domain increased, the citation-based networks, as well as the communities identified, evolved. We reveal the trends in our data by tracking how the networks grow over each five-year time-slice. Results show that the percentage of nodes in the giant component is generally large, and the clustering coefficient and modularity are high, indicating also
a cohesive yet shifting modular structure in this domain over time. Detailed topological quantities are laid out in Supplementary Table S4.

Fig. 10 is a time-series graph that describes the percentage of publications in the citation-based networks across the predefined disciplinary categories. In the cumulative networks, the fields consistently most active in this domain have been communication, computer science, psychology, business, and those related to health. However, several things are noteworthy about the variability of the distribution of disciplines and fields over time. First, the proportion of communication papers in the domain’s structure has declined remarkably, gradually dropping since 2009. Meanwhile, the proportion of computer science papers has increased over time, accounting for the most significant share since 2014. Second, multidisciplinary sciences first appeared in 2014 and have since increased. Likewise, other natural sciences and technology-oriented disciplines, such as mathematics & computational biology, have also increased since 2014. Linguistics likewise has increased, probably reflecting this field’s uptake of NLP-based approaches. Overall, Fig. 10 reveals the changes in and growth of more diverse interests in this domain. Further, the number and variety of disciplines in the clusters in all the networks have risen remarkably over time (see Supplementary Table S8-10). Taken together, the specialty composition of networks and clusters is increasingly interdisciplinary over time, with core fields such as communication and, in particular, computer science as (relative) focal points throughout, and with a wide variety of fields in the socio-cultural sciences, the natural sciences, and technology gradually joining in.

4.4.2. Dynamic change in network clustering structure

We further investigate the evolution of clustering structures in this domain by looking at the more dynamical aspects of our networks. In the external co-citation network (Fig. 11), we study how the articles published in T1 are co-cited across three periods, from T1 (2004-2009) to T2 (2010-2014) to T3 (2015-2019). Using an alluvial diagram (Fig. 11b), we analyze changes in the network structure, focusing on structural change at the cluster level. Note that the diagram does not present the growth of the overall network; instead, it shows how the clusters in T1 merge and split over time. In other words, the diagram indicates the flow of ideas in papers
published in the first period and how this flow changes over time. In doing so, it reveals some important changes in this domain, suggesting how, as generally tends to be the case in science, new ideas are built upon older ones.

As shown in Fig. 11a, nodes are more densely connected beginning with T2, especially for those nodes in T2C1 (a cluster composed of a diverse range of fields; see Supplementary Table S11 for a complete list of disciplines in each cluster). Meanwhile, a few nodes in T2 were no longer active (i.e., no longer being co-cited) in T3. This indicates that, overall, most of the publications in T1 are increasingly recognized over time as new research keeps referring back to (i.e., co-citing) them.

In what follows, we report the WS-based fields represented as nodes in the clusters. To begin with, clusters in the first period are generally discipline- and field-based. For instance, T1C5 consists primarily of health-related fields, including substance abuse, nursing, and public, environmental, & occupational health; and T1C1 includes mainly business. From the second period, clusters become more interdisciplinary, spanning an increasingly wide range of fields in the socio-cultural sciences, natural sciences, and technology, e.g., T2C2 and T3C2. Specifically, from T1 to T2, most of the clusters in T1 merged into T2C1, including T1C2 (computer science), T1C3
(ethics-education & educational research), T1C7 (communication), and T1C8 (computer science). In contrast, some other clusters are still distinct, such as T1C1 (business), which fully transformed to T2C2 (interdisciplinary sciences). For some clusters, however, the transformation pattern is more complex and challenging to interpret. This includes articles in T1C5 (health-related), which mainly formed T2C4 (health-related) and partly merged into T2C1 (no dominant fields). From T2 to T3, clusters merged and split into fewer and even more integrated clusters. For instance, papers in T2C1 mainly formed T3C1 (communication) and partly split into T3C4 (no dominant fields) and T3C2 (interdisciplinary sciences), while papers in T2C3 (health-engineering) and T2C4 (health-related) jointly formed T3C2 (health care science, engineering & economics). Overall, this transformation pattern indicates progressive knowledge integration in this domain, implying that earlier work in the domain has been gradually bundled to become part of fewer and larger interdisciplinary clusters of conversations. Compared with patterns in the overall external co-citation network, papers published in T1 over time, and considered as they are here, reveal that earlier core work does get picked up gradually in more interdisciplinary ways by the wider academic world. This is the case even as the global pattern is generally discipline- and field-based, owing most likely to the general size increase and diversification of the domain.

5. Discussion and conclusion

In this study, we set out to investigate the interdisciplinary specialty structures of a knowledge domain using an original dataset, as well as to enhance bibliometric cluster analysis for such investigations. We outlined a detailed approach for processing bibliometric data to extract clusters in a set of citation-based networks. We illustrated our approach through the case of an important emerging domain, i.e., research that actively utilizes (rather than merely mentioning or abstractly discussing) social media-derived data for making social-scientific and related claims. This domain presents a particularly relevant case for improving bibliometric cluster analysis to investigate specialty structures of (inter)disciplinary research fields and domains. Besides, this domain itself has been instrumental in our understanding of intellectual patterns at a time of increasing emphasis on big social data. The methods employed in this study originate in the fields of bibliometrics and network science. The data was retrieved and combined from two large databases, Web of Science (WoS) and Microsoft Academic Graph (MAG), enhancing the data availability for robust bibliometric analyses. A novel aspect of this study entails juxtaposing three types of citation-based networks – internal and external co-citation networks, as well as bibliographic coupling networks. All three reveal different aspects of this domain’s intellectual structure in terms of how knowledge is sourced and picked up in the wider academic community. Our results, based on a comparison between empirical patterns and null model observations, inform us of the disciplinary composition of subfields identified in this domain; and they suggest an increasingly interdisciplinary trend over the years as a wide variety of fields in the socio-cultural sciences, the natural sciences, and technology gradually join in.

We found a largely cohesive structure over time in the citation-based networks and associated clusters, implying that research generally progresses within relatively distinct research clusters in this domain (even if they themselves are also evolving). Concerning the primary question of the subfields identified in the domain, our study supports several important observations. In the internal co-citation network, we identified three large and mainly thematically bounded research subfields – interdisciplinary socio-cultural sciences, health sciences, and geo-informatics. All three designate the principal hub of research interests recognized by this domain itself. Another important finding was that the main clusters cover a wide range of WC-based fields, which to some extent indicates the nature of interdisciplinary by showing that later work in this domain cross-referenced and integrated bodies of knowledge from various disciplines and fields.

In this regard, we note with interest that the external co-citation network provides a more significant number of discipline- and field-based clusters relative to the internal network. This implies that in its external impact on the wider academic world, the interdisciplinary work in social media data-based research tends to get picked up and used according to more discipline-based communication channels. From the core domain perspective, then, this domain can be assumed to have split into a few large, interdisciplinary, and thematically bounded clusters; yet, when viewed from the wider fields of science, it is seen as possessing a more field-based intellectual structure. One unanticipated finding was that the bibliographic coupling clusters demonstrate a different disciplinary configuration than the co-citation clusters, even as both approaches indicate cognitive similarity in this domain. It shows the centrality of computer science as the source-based coupling (bibliographic coupling) of core domain work at the same time that this work is then channeled into a wider diversity of disciplinary audiences (external co-citation). The fact that different networks exhibit different specialty structures suggests that caution needs to be exercised when only one type of network is used for such investigations.

Previous research has proposed that "interdisciplinarity" can be measured through diversity and coherence (Rafols & Meyer, 2010). Thus, based on this approach, this emerging domain – a dataset covering many WoS-based research fields (N=1622) with coherent bibliometric networks – is increasingly interdisciplinary over time. In addition, we determine the origins of diverse bodies of knowledge under research sub-communities (specialties) in a knowledge domain, thereby providing a systemic description of the interdisciplinary research landscape. Our approach considers the cognitive heterogeneity, i.e., connecting scientific research across different subfields, allowing us to provide more fine-grained views on interdisciplinary communication rather than making aggregated or averaged measures. We therefore think that this may be better indicating interdisciplinarity features in the broader sense of knowledge integration, i.e., as a systemic process in which different scientific fields become related in distinct research communities.

Moreover, the present study provides a useful overview of the academic practice of using social media-derived data in various forms and scales, and in different fields, including both disciplinary and interdisciplinary research. One implication is that new and intensified conversations among a wide variety of fields in socio-cultural sciences, natural sciences, and technology are opening up in this domain, not least in the social sciences and computer science. Thus, while we show only relatively coarse and global patterns,
our finding to some extent corresponds to and underwrites a number of more epistemological concerns in the context of social data abundance (e.g., Kitchin, 2014), by exhibiting the density of new collaboration between different disciplines and methodologies.

It is important to emphasize that our results should be interpreted with caution, given a number of limitations and uncertainties. The major limitation in this study is the use of the predefined disciplinary and field-wide categories provided by Web of Science. As we study a developing domain through WC-defined fields, we cannot keep up with the most recently acknowledged ones, such as digital humanities and computational social science, since these fields are not (yet) recognized at the level of WC. This holds true even as we know these fields have coalesced in the period we are studying (Kitchin, 2014; Tang et al., 2017; Edelmann et al., 2020). Moreover, interpreting thematic communities from WC-defined fields is generally hard; for instance, previous research points to a tendency to overestimate the category of communication, which spans a broad range of disciplines in the social sciences and humanities (Leydesdorff & Probst, 2009). Another, although arguably lesser, limitation of this study is that our sampling strategy, along with the merged dataset, is somewhat limited in scope. We adopted a relatively conservative sampling strategy, covering only highly prestigious research outputs that met our definition of this domain; that is, journal articles covered by the Web of Science database, thus excluding conference proceedings, monographs, and other academic outputs. It is worth noting that the selection of queries for extracting relevant papers is not an easy task. It is known as a general limitation of the bibliographic approach (Kajikawa & Takeda, 2009). We collected relevant papers using a set of queries and tested their relevance. Future work could further improve our queries by adding some synonyms such as "data set". Furthermore, we only used the DOI as the standard document id and the linkage between the data sources. Our analysis might underestimate the true overlap between the two data sources. Another potential issue with data sources is the author name ambiguity problem in the MAG author IDs that we used for the author-level analysis (Panagopoulos et al., 2019). Finally, the clustering algorithm produces a large number of clusters; however, due to space limitations, we are unable to provide a thorough overview of all the clusters that may be of interest, thereby risking that we focus unduly on top-level patterns while overlooking tendencies at lower levels. These results must be interpreted with caution especially for indicating global patterns such as interdisciplinarity.

Notwithstanding these limitations and uncertainties, we believe our study offers important insights into the investigation of interdisciplinary specialty structures from bibliometric data, as seen from a global topological perspective, and the intellectual patterns of social media data-based research as an important emerging domain of science. We also acknowledge and affirm that follow-up work should move beyond our focus on (inter-)disciplinary relations and community compositions, to look more closely, e.g., at the knowledge content (data sources, methods), questions of institutional, gender-based, and other status hierarchies, as well as author- and institution-based collaboration patterns.

Availability of data and materials

The WoS dataset is proprietary and cannot be shared due to the terms of use by Clarivate Analytics. The search strings for retrieving the dataset from WoS are provided in the Data Section. The networks and node attributes data and the code for conducting analyses in this paper are made publicly available in a Github repository (YangliuF95/Sm_network_analysis, 2021).

Additional material

Additional file 1: Supplementary Information

CRediT authorship contribution statement

Yangliu Fan: Conceptualization, Data curation, Software, Investigation, Visualization, Writing – original draft. Sune Lehmann: Conceptualization, Methodology, Software, Writing – review & editing. Anders Blok: Conceptualization, Investigation, Validation, Writing – review & editing.

Acknowledgments

We would like to thank Mathias Wullum Nielsen and Morten Axel Pedersen for generous suggestions. We thank Copenhagen Center for Social Data Science (SODAS) for financial support. We thank the two anonymous reviewers as well as the editor for their comments on this article.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.joi.2022.101310.

References

Visit, Š Small, Leydesdorff, Kessler, Y. W., information H., H. M., YangliuF95/sm_network_analysis
https://leidenmadtrics.nl/articles/on-measuring-interdisciplinarity-from-indicators-to-indicating
Leung, the Bergstrom, Tavasoli, & evi Article and Song, N. A.
From "interdisciplinarity: The technology of networking people" in Academic: analyzing the publication output of a university. Scientometrics, 113(3), 1551–1571.


Leydesdorff, L., & Ivana, I. (2020). The measurement of “interdisciplinarity” and “synergy” in scientific and extra-scientific collaborations. 10.1002/asi.24416.


YangliuF95/sm_network_analysis https://github.com/YangliuF95/sm_network_analysis.

