A systematic metadata harvesting workflow for analysing scientific networks

Bilal Hayat Butt[1], Muhammad Rafi[2], and Muhammad Sabih[3]

[1]**Department of Computer Science, D.H.A. Suffa University, Karachi, Pakistan**

[2]**Department of Computer Science, National University of Computer and Emerging Sciences, Karachi, Pakistan**

[3]**Department of Electrical Engineering, D.H.A. Suffa University, Karachi, Pakistan**

Corresponding author:

Bilal Butt[1]

Email address: bilal.hayat@dsu.edu.pk; bilal.dsu@gmail.com

ABSTRACT

One of the disciplines behind the science of science is the study of scientific networks. This work focuses on scientific networks as a social network having different nodes and connections. Nodes can be represented by authors, articles or journals while connections by citation, co-citation or co-authorship. One of the challenges in creating scientific networks is the lack of publicly available comprehensive data set. It limits the variety of analyses on the same set of nodes of different scientific networks. To supplement such analyses we have worked on publicly available citation metadata from Crossref and OpenCitatons. Using this data a workflow is developed to create scientific networks. Analysis of these networks gives insights into academic research and scholarship. Different techniques of social network analysis have been applied in the literature to study these networks. It includes centrality analysis, community detection, and clustering coefficient. We have used metadata of Scientometrics journal, as a case study, to present our workflow. We did a sample run of the proposed workflow to identify prominent authors using centrality analysis. This work is not a bibliometric study of any field rather it presents replicable Python scripts to perform network analysis. With an increase in the popularity of open access and open metadata, we hypothesise that this workflow shall provide an avenue for understanding scientific scholarship in multiple dimensions.

INTRODUCTION

Scientific networks provide useful information in understanding the dynamics of science (Price, 1965). With the advent of numerous bibliographic data sources (Waltman and Larivière, 2020), it is now possible to analyse different scientific networks. The proposed study focuses on article citation network, author citation network, and co-authorship network. Usually, studies that focus on co-authorship do not require information about the citation. However, having citation links would enable a more complete and holistic view of the possible relations among authors (Zingg et al., 2020). To achieve this objective comprehensive access to citation metadata is required. This can be accomplished using publicly available citation metadata available via Crossref (Hendricks et al., 2020). However, applying network analysis on it requires a series of steps that may not be intuitive. Proposed study furnish details of these steps so that it is easy to supplement it with different analyses.

Social network analysis techniques are applied to study scientific networks. It includes citation networks of article or author, and author collaboration network. Usually, these networks are build using different data sources. However, our workflow can create all these networks using OpenCitations data and Crossref. The workflow presented in this article is part of a study on the influence of scholarly research artefacts. To this end, we primarily limit our research goal to have a systematic workflow for analysing scientific networks. In this work, we aim to utilise open metadata (Peroni et al., 2015), made available using Crossref. Also, we utilise open source Python libraries for network analysis, namely, NetworkX (Hagberg et al., 2008) and SNAP (Leskovec and Sosič, 2016). Python is used based on its popularity with researchers as per survey results by AlNoamany and Borghi (2018). Although graphical software has

47 an ease of use, we prefer to provide workflow as a set of Python scripts to facilitate advance analysis.
48 Details of batch execution of workflow scripts are available on GitHub for researchers with programming
49 background (Butt and Faizi, 2020). This article outlines details of a case study for analysing collaboration
50 network of Scientometrics journal metadata. All steps are documented for the replication of this study.
51 This work shall lay the groundwork for analysing scientific networks using metadata of different journals,
52 set of journals or a subject category. One such analysis is the identification of prominent authors (gurus).
53   Identifying prominent authors of any field is one of the primary focus for young researchers. Likewise,
54 other researchers tend to follow research published by gurus of the field. Defining a guru of the field is
55 not an easy task, and the definition of guru will be very subjective. To this end, we focus on the definition
56 of guru using the centrality measures of social network analysis. Details of different centrality measures
57 are depicted in Figure 1 (Newman, 2010). The following description was inspired by (Milojević, 2014).
58 Simply put, any author with a high citation count may be considered the guru. It can be achieved using
59 degree centrality. Another way of identifying a highly cited individual is to calculate whose paper is
60 in top percentile within the domain. However, we currently limit such definitions to degree centrality
61 of articles. It is not always the case that all highly cited authors are equally influential. Those who are
62 cited by other influential authors may also be termed as influential even though they may or may not
63 have high citation count. Likewise, any author frequently collaborating with influential authors would
64 also influence that field. This recursive definition of influence is well captured by eigenvector centrality.
65 Another centrality measure, namely betweenness centrality would define an author as prominent if author
66 collaborates with different groups. Centrality measures of closeness and farness is the extent to which an
67 author is on average close to or far from other authors within the network, respectively.
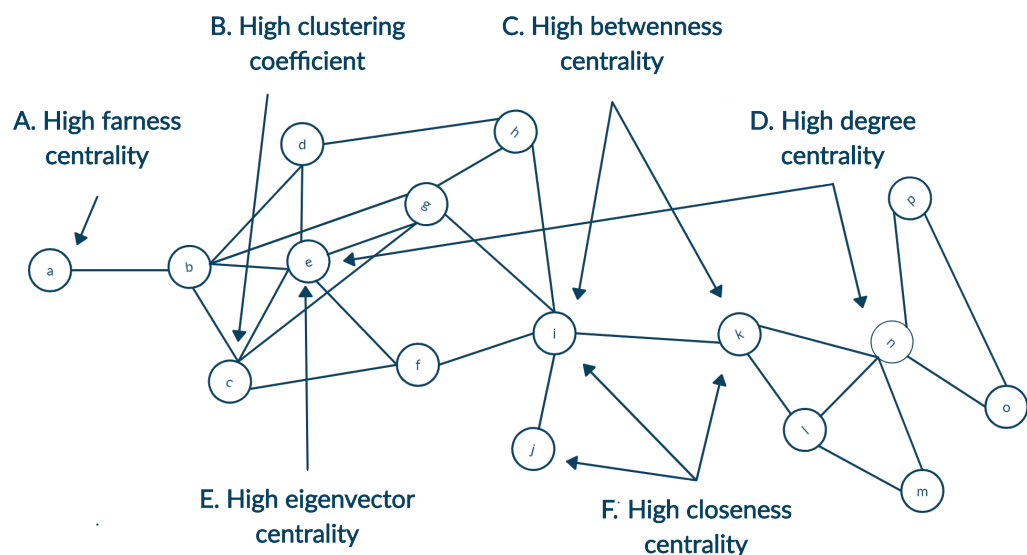
Figure 1. A toy network showing nodes with high centrality for different measures.
A. shows high farness centrality since the node (a) has the maximum average distance to other nodes.
B. shows high clustering coefficient since neighbours of the node (c) are all connected as well.
C. shows high betweenness centrality since the highest number of shortest paths will go from the node (i)
and (k) since they are bridging two parts of the network.
D. shows high degree centrality as both the nodes (e) and (n) have highest count of neighbours.
E. shows high eigenvector centrality since node (e) is connected to many neighbours with a relatively
higher degree.
F. shows high closeness centrality as the average distance from nodes (i), (j) and (k) are minimum to other
nodes.

68   In the case of analysing the citation network with a limited snapshot of data, this could be supplemented
69 by creating the ego-centered network (Newman, 2003). Citation index allow fetching the metadata of

70   articles in two directions. First, articles cited by the original article can be accessed as references. Second,
71   articles that cites the original article can be accessed as citations. For each node, its connecting nodes
72   inclusion enhance the breadth of the network. Node is known as the ego and its neighbours are termed as
73   alters. In Figure 1 alters of ego node (n), namely, node (k), (l), (m), (o) and (p) will form its ego network.
74   The remainder of this paper is structured as follows. After giving some preliminary definitions, an
75   overview of research is presented on scientific networks using centrality analysis. Next, we provide details
76   about the pipeline architecture in methodology before moving to the steps of acquiring and analysing data.
77   Steps include extracting details of the citation index which is downloaded from the web and loaded in
78   memory. Next, we discuss how the citation metadata are fetched and filtered. In the last step, we explain
79   the analysis of different scientific networks. Further, a case study is discussed for applying centrality
80   analysis on collaboration network. We conclude with details of how this work can be further enhanced.

PRELIMINARIES

82   Some terminologies are introduced below that are necessary for understanding this study. Also, some
83   acronyms used in this article are listed.

84   **SNA** Social network analysis is concerned with techniques and measures applied to real-world networks.

85   **Centrality** A property of the node which shows the relative rank of a node compared to other nodes.

86   **Degree centrality** Nodes having a high number of connections are more central.

87   **Betweenness centrality** Nodes on the majority of the shortest path connecting other pair of nodes.

88   **Closeness centrality** Average distance to other nodes in the network is minimum.

89   **Farness centrality** Reciprocal of closeness centrality.

90   **Clustering coefficient** Well knitted alliances get a higher score.

91   **Eigenvector centrality** Influence of one node transfers to other connected nodes.

92   **PageRank** Variant of eigenvector centrality along with provision for randomness.

93   **Katz centrality** Variant of eigenvector centrality for directed graphs without cycles.

94   **Guru** Any important node in the scientific network as defined by centrality measures.

95   **Scientific network** A social network with nodes as author/article and edges as co-author/citation.

96   **Citation network** A type of scientific network with article nodes connected by cited-to and citing edge.

97   **Collaboration network** A type of scientific network with author nodes connected by co-author edge.

98   **Co-citation network** A type of scientific network with nodes connected if both nodes get cited together.

99   **Bibliographic coupling** A type of scientific network with nodes connected if both cites the same node.

100 **Ego network** Include all nodes connected to and by any specific (ego) node.

101 **JSON** JavaScript Object Notation is s format to represent complex and unstructured data.

102 **Crossref** Using `api.crossref.org` for publicly available citation data (Hendricks et al., 2020).

103 **COCI** The OpenCitations Index of Crossref open DOI-to-DOI citations (Heibi et al., 2019).

104 **WoS** Clarivate Analytics' Web of Science `www.webofknowledge.com` (Birkle et al., 2020)

105 **JCR** Journal Citation Reports `incites.clarivate.com`

106 **Scopus** Elsevier's Scopus `www.elsevier.com/solutions/scopus` (Baas et al., 2020)

107 **DBLP** Digital Bibliography and Library Project `www.dblp.org`

108 **ACM** ACM Digital Library `dl.acm.org`

109 **APS** American Physical Society `www.aps.org`

110 **MAG** Microsoft Academic Graph `academic.microsoft.com` (Wang et al., 2020)

111 **PubMed** PubMed Central Database `www.ncbi.nlm.nih.gov/pmc`

112 **CiteSeerX** Scientific literature digital library and search engine `citeseerx.ist.psu.edu`

113 **ACL** ACL Anthology Reference Corpus `acl-arc.comp.nus.edu.sg`

114 **SNAP** Stanford Network Analysis Project `snap.stanford.edu`

115 **SCIM** Scientometrics Journal `www.springer.com/journal/11192`

116 **API** Application Programming Interface

RELATED WORK

118 Visualising bibliographic data as a network is not new, Price (1965) introduced the work more than 50
119 years ago. Most recent studies are on co-authorship network (Abbasi et al., 2012; Milojević, 2014; Lee,
120 2019; Singh and Jolad, 2019; Waheed et al., 2019), however others have focused on citation network of
121 authors (Ding, 2011; Milojević, 2014; Xu and Pekelis, 2015; Waheed et al., 2019) or citation network of
122 journals (Waltman and Yan, 2014; Leydesdorff et al., 2018). Only a couple of studies have utilised more
123 than one scientific network for analysis (Milojević, 2014; Waheed et al., 2019). Traditionally bibliometric
124 analysis has been done using WoS and Scopus (Waltman and Larivière, 2020). A similar case has been
125 observed in studies on scientific network analysis where the data sources used are Scopus (Abbasi et al.,
126 2012; Lee, 2019) or WoS (Ding, 2011; Milojević, 2014; Waltman and Yan, 2014; Leydesdorff et al., 2018;
127 Massucci and Docampo, 2019). However, some recent studies have focused on open access data sources
128 (Singh and Jolad, 2019; Van den Besselaar and Sandström, 2019; Waheed et al., 2019). Other data sources
129 such as PubMed, CiteSeerX and ACL are not discussed in this article. They are used mostly for text
130 analysis instead of network analysis. Below we list a brief account of work done on scientific networks
131 using centrality measures. Details are summarized in Table 1 in chronological order. Some earlier studies
132 such as (Newman, 2004) are not included as we have focused on studies published in the last decade.

133 Ding (2011) proposed to analyse the author citation network with weighted PageRank. The author
134 proposed the strategy on predicting prize winners that outperforms the conventional h-index and related
135 citation count measures. Abbasi et al. (2012) discussed the use of betweenness centrality as a measure
136 of getting more collaborators compared to degree and closeness centrality. They have used temporal
137 co-authorship network in the steel research domain. Data was manually curated and downloaded from
138 Scopus.

139 Ortega (2014) analysed 500 co-authors' ego network and conclude that bibliometric indicators and
140 centrality measures are correlated. They have used clustering coefficient, degree and betweenness
141 centrality as local metrics. Some global level metrics were also analysed using the ego network. It is one
142 of the early studies using MAG.

143 Two book chapters provide hands-on details about centrality measures (Milojević, 2014) and PageRank
144 (Waltman and Yan, 2014) using WoS data. Milojević (2014) constructed the author collaboration network
145 and calculated degree, betweenness, eigenvector and closeness centrality. Waltman and Yan (2014) details
146 applying PageRank on journal citation network.

147 Xu and Pekelis (2015) used a manually curated dataset for authors of China and Taiwan in the field
148 of Chinese Language Interpreting Studies. They have applied PageRank and degree centrality to find
149 influential authors within different clusters identified using community detection.

150 Leydesdorff et al. (2018) have used betweenness centrality to measure multidisciplinary journals.
151 Authors have limited their approach with a diversity measure and evaluated it on JCR data. Usually,
152 a journal gets citation within its subject category but those journals cited/citing the other fields are
153 considered a bridge between the subject categories.

154 A case study for young researchers performance evaluation is presented by Lee (2019). The author
155 analysed the collaboration network of these researchers using statistical analysis for the frequency of
156 collaborators. The degree centrality is showed to correspond with future publication count. It is akin to Li

et al. (2019) who concludes that collaboration of young scientist with top-ranked co-authors has a high probability of future success.

Massucci and Docampo (2019) applied the PageRank algorithm on a university citation network. Working on five different subject categories, they proposed a framework which is more robust than existing university rankings. It holds a high correlation with these accepted rankings. Singh and Jolad (2019) utilised data of APS journals to form collaboration network of Indian physicist. In this co-authorship network, they have applied different centrality measures and report the overlapping top authors.

Van den Besselaar and Sandström (2019) discuss the potential use of clustering coefficient and eigenvector centrality in ego network of research students and their supervisor. Both metrics are used to gauge the independence of a researcher. They have handpicked 4 pairs of researchers. The authors suggested that there are numerous ways to capture the researcher's autonomy. However, when evaluating large data sets the clustering coefficient and eigenvector centrality can be effective.

Waheed et al. (2019) discusses the use of centrality measures on multiple scientific networks to improve article recommendation. They filter the citation network to five levels in cited-by and citing directions. Evaluating a large-scale network available at AMiner they proposed a hybrid recommendations strategy. It includes different centrality measures on author collaboration network, author citation network and article citation network.

Table 1 summarises the studies in three aspects. First, the bibliographic data source used. Second, the scientific network created. Last, details of techniques applied for analysis. Studies show that applying centrality measures is a useful analysis in bibliometrics. However, these approaches are mostly not scalable and require considerable effort to apply the same analysis on bigger networks. In some cases, the tools limit the size of network analysed, whereas in other studies the data sets are manually curated. In comparison to our work, most of the studies are limited to one type of network. The way data sets are acquired limits the analysis to expand to another type of networks (Zingg et al., 2020). We observe that very few studies have either used multiple networks or mentioned that if these can be curated with the same data source. With WoS and Scopus, it is theoretically possible to create all networks. However, with other data sources, a dump is usually uploaded with limited metadata. It restricts the authors to confine their studies.

Case studies similar to our workflow are also available on the proprietary data source of WoS (Milojević, 2014) and Scopus (Rose and Kitchin, 2019). Further, a set of graphical tools are also available as discussed by Moral-Muñoz et al. (2020) in a recent survey. Most tools do not give access for Crossref apart from (Van Eck and Waltman, 2014; Chen, 2005).

One of the recent studies (Rose and Kitchin, 2019) focuses on using Scopus data for network analysis. They have provided an API for researchers to perform useful analyses. Accessing Scopus is possible with Elsevier Developer API Key. However, it requires institutional or authenticated access. Such access is not possible, especially for developing countries (Herzog et al., 2020). Although our work is similar to Rose and Kitchin (2019) in using Python for analysing scientific network, it is different in two aspects. Firstly, we are working with OpenCitatons data using Crossref. Secondly, we have not developed an API interface that needs maintenance and support since Crossref, NetworkX and SNAP fulfil the purpose.

Chen (2005) discusses the identification of highly cited clusters of a scientific network. The pivotal points in the scientific network are captured using betweenness centrality. The author uses clinical evidence data associated with reducing risks of heart disease to illustrate the approach. They have discussed the design of the CiteSpace tool and its new feature for identifying pivotal points. They used betweenness centrality to identify pathways between thematic clusters. Nodes with high betweenness centrality are potential pivotal points in clustering the scientific network. Instead of a graphical software tool, we propose to use Python scripts. It gives more flexibility for advance analysis. For a detailed survey, we would refer the interested reader to (Moral-Muñoz et al., 2020).

Usage of open metadata are gaining popularity. On the other hand, publicly available metadata has its limitations with completeness and verification. Iorio et al. (2019) concludes that using OpenCitatons data for evaluation purpose is not enough due to the unavailability of complete data. However, more than half of the records are available in comparison to WoS and Scopus. A similar evaluation is also done by Nishioka and Färber (2019) and Martín-Martín et al. (2020). Further, there are different approaches to augment the current OpenCitatons data (Daquino et al., 2018; Heibi et al., 2019; Peroni and Shotton, 2020). Kamińska (2018) discusses a case study for using OpenCitatons data for visualising citation network. Zhu et al. (2020) has used COCI to evaluate books scholarship. With a scripted workflow, we

Study	Bibliographic Data Source	Scientific Network(s)	Social Network Analysis
Ding (2011)	WoS	Author Citation	Weighted PageRank
Abbasi et al. (2012)	Scopus	Author Collaboration	Degree, Betweenness, Closeness Centrality
Ortega (2014)	MAG	Co-Author Ego Network	Clustering Coefficient, Degree and Betweenness Centrality
Milojević (2014)	WoS	Author Collaboration and Citation, Article Citation	Degree, Betweenness, Closeness, Eigenvector Centrality
Waltman and Yan (2014)	WoS	Journal Citation Network	PageRank
Xu and Pekelis (2015)	Manual	Author Citation	PageRank and Degree Centrality
Leydesdorff et al. (2018)	WoS/JCR	Journal Citation	Betweenness Centrality
Lee (2019)	Scopus	Author Collaboration	Degree and Betweenness Centrality, Clustering Coefficient
Massucci and Docampo (2019)	WoS	Institutional Citation	PageRank
Singh and Jolad (2019)	APS Journals	Author Collaboration	Centrality, Community Detection
Van den Besselaar and Sandström (2019)	Manual	Researchers Ego Network	Clustering coefficient, eigenvector Centrality
Waheed et al. (2019)	DBLP, ACM, MAG	Author Collaboration and Citation, Article Citation, Co-citation and Bibliographic Coupling	Degree, Betweenness, Closeness, Eigenvector Centrality

Table 1. Review of studies applying social network analysis on scientific networks.

hypothesise that it would be easier for masses to adopt OpenCitatons data for bibliometric analysis.

METHODOLOGY

This section details a systematic workflow from data fetching to analysis. A series of steps are required to apply centrality analysis on the author collaboration and author citation networks. Utilising the article citation network, available as citation index, these networks get created. All scripts were executed on Windows Server machine having Quad-Core AMD Opteron(TM) Processor 6272 with 128 GB RAM installed. The initial processing of data requires heavy computation and memory once. Later, the data are converted to a compressed binary format using libraries for processing large networks. It can run on any standard laptop machine. Below, we provide details of the workflow to create scientific networks. Although the case study is limited to data of SCIM, we have made the process automated. This automation helps applying the same script for other journals with minimum changes.

Overview of the process is shown in Figure 2 and further details about each of the following steps are documented separately. Each of the steps is further distributed with three sub-steps for clarity and batch execution.

Step 1 Download the citation index available as COCI (Heibi et al., 2019).

₂₂₇ **Step 2** Download the metadata for given ISSN through Crossref.

₂₂₈ **Step 3** Fetch the ego network from COCI data for the DOIs of respective ISSN.

₂₂₉ **Step 4** Merge these data to create scientific networks.

₂₃₀ **Step 5** Apply the centrality analysis on these networks.
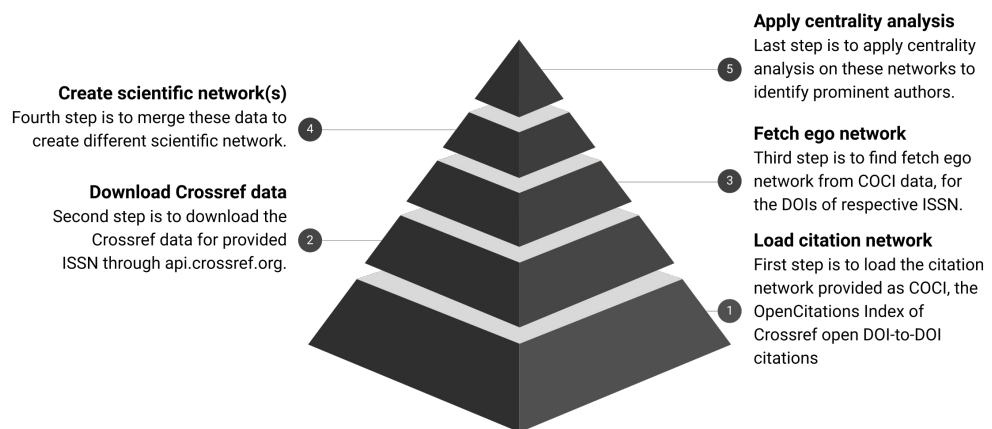
Figure 2. Workflow for analysing scientific networks. The pyramid shows the refinement of data at every step. COCI contains approximately 625 M edges. It gets reduced as a subset of nodes fetched for respective ISSN. Finally, the top of the pyramid shows the output in the form of a few nodes identified with high centrality.

₂₃₁   Python scripts are uploaded as Supplemental Files which can also be accessed on GitHub (Butt and
₂₃₂ Faizi, 2020). It gives replication and reuse of this study for other ISSN or bibliometric analyses for
₂₃₃ different network types. Details are provided below for the understanding of this study.

₂₃₄ **Load citation network**
₂₃₅ Summary of the sub-steps to load citation network is shown in Figure 3. Below we define the sub-steps to
₂₃₆ convert the COCI data to use in Python libraries for network processing. This step is computation and
₂₃₇ memory intensive but needs to be performed only once.
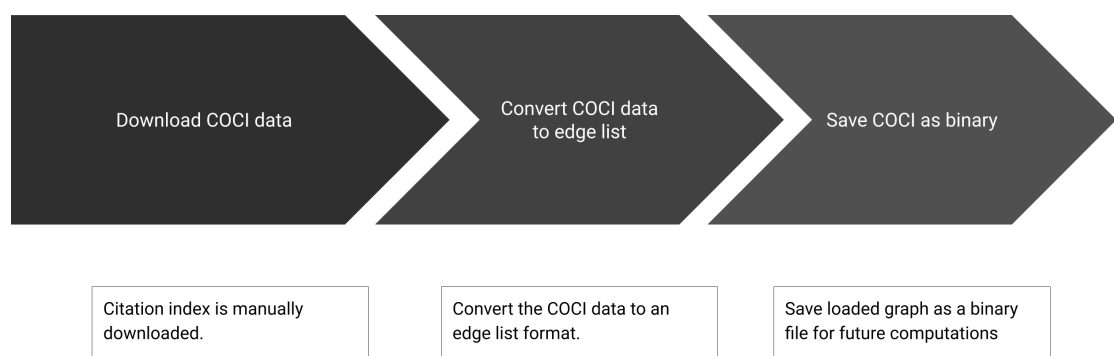
Figure 3. Step 1 of the workflow with details of creating the citation network. Sub-steps are applied sequentially.

238 **Download COCI data**   COCI is manually downloaded from (OpenCitations, 2020). The 15 GB Zip file
239 extracts to 98 GB set of files. Loading this data in-memory resulted in memory-overflow even when using
240 128 GB RAM. Therefore, in the next step, we remove the columns other than citing and cited. These two
241 columns are used to create the article citation network.

242 **Convert COCI data to edge list**   This step is needed to convert the COCI data to an edge list format. In
243 this format, two nodes on each row signify an edge. This format is supported by SNAP (Leskovec and
244 Sosič, 2016) for processing large-scale network data such as COCI. After this step, the edge list file is
245 approx 35 GB. We convert the COCI from comma-separated-values (CSV) to space-separated-values
246 having citing and cited columns. It is the only format supported by SNAP for bulk upload. Some
247 formatting corrections are done for removing extra CR/LF and quotes. It hampers the loading process of
248 SNAP. We have failed to load the same files with other libraries which are relatively more intuitive but not
249 as powerful as SNAP (Leskovec and Sosič, 2016). However, we later discuss how this data can be used
250 with other libraries. Details to save network in a format supported by most network processing libraries is
251 provided in subsequent steps.

252 **Save COCI as binary**   Loading 35 GB edge list in-memory using SNAP takes approx 5.5 hours. Since
253 the edge labels are DOI in the COCI data, therefore they are saved as strings. However, this slows down
254 further processing so strings are converted to a hash file. There are two binary files generated when
255 loading the COCI data in SNAP. First is DOIDirected.graph file which contains the directed citation
256 network of COCI with integer node labels. Second is DOIMapping.hash which maps the integer node
257 label to respective DOI. We save loaded graph as binary files for further computations. Loading binary
258 file in-memory takes a few minutes, compared to hours for loading CSV data. Downside is that additional
259 columns of COCI are currently not being utilised. DOIMapping.hash is simply a node list where node
260 number is mapped to its label (DOI). DOIDirected.graph is an edge list on node number. Using numeric
261 labels makes SNAP less intuitive but more powerful since computations are much faster when integer
262 labels are used. The mapping to string labels is possible with the node list.

263 **Fetching Crossref metadata**
264 Summary of the sub-steps to download Crossref metadata are shown in Figure 4. Below, we define the
265 sub-steps to fetch the citation metadata and converting it to list of authors and DOIs. These steps only
266 give API string to fetch data for a single journal. However, it is possible to fetch data with other filters
267 and details are available in Crossref documentation. The metadata downloaded via different filters is in a
268 similar format which makes this script reusable for a variety of tasks.
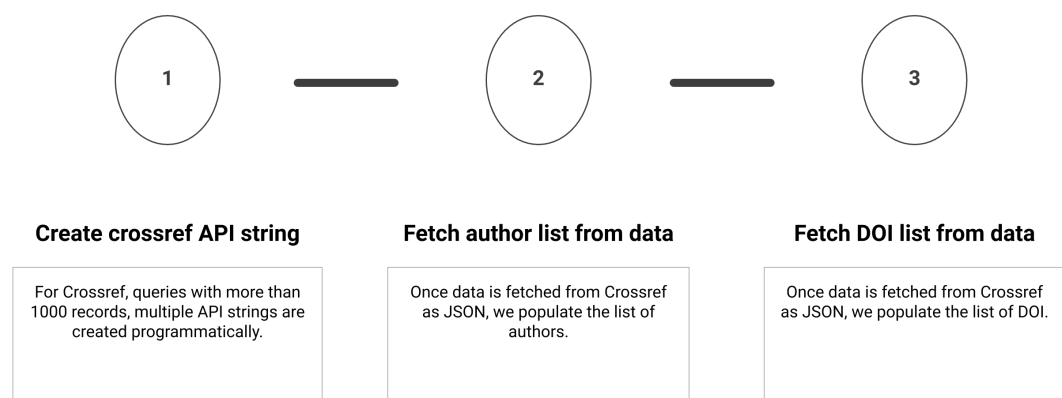
Figure 4. Step 2 of the workflow with details of fetching metadata from Crossref API. Sub-steps are
applied sequentially.

269 **Create Crossref API string**   Crossref limits a one time query to 1000 records for a single ISSN. For
270 queries with more than 1000 records, multiple API strings are programmatically created. Metadata of
271 SCIM is fetched via Crossref API which contains total 1857 records. These records are fetched by two
272 API requests and a combined JSON file is created.

273 **Fetch author(s) list from data**   Once data are fetched from Crossref as JSON we populate the list of
274 authors. We extract authors from the previous downloaded JSON. It is important to note that we do
275 not apply any technique for author name disambiguation and rely on Crossref for correct author names.
276 Although this is problematic for further analysis. Corrected data from a single source is much more
277 efficient than using local methods of cleaning. A similar approach is used by MAG (Wang et al., 2020).

278 **Fetch DOI list from data**   Once data are fetched from Crossref as JSON we populate the list of DOI.
279 DOIs are extracted from the previously downloaded JSON. Although the purpose of fetching DOI is
280 redundant, it's replica script is created to suggest that analysis with only given DOI list is also possible.
281 So the previous two sub-steps can be ignored if analysing a specific journal is not needed. If the list of
282 DOI is fetched from an external source then it can be easily incorporated in this workflow.

Creating ego network

284 Summary of the sub-steps to create ego network are shown in Figure 5. Below, we define the sub-steps to
285 create ego network. This is an optional step. Iterating this step multiple times will grow the network as
286 desired. This step is not used in the case study, however, with publicly accessible metadata it is easier to
287 scale our approach. Further, this step justifies our approach of using SNAP over other network processing
288 libraries. The process of creating the ego network is not only fast but intuitive to code due to a variety of
289 functions available in the extensive library documentation. These functions make it easier to access the
290 nodes in both directions of an edge. Also, the hash with integer labels makes the sub-graph computation
291 faster than using string labels.
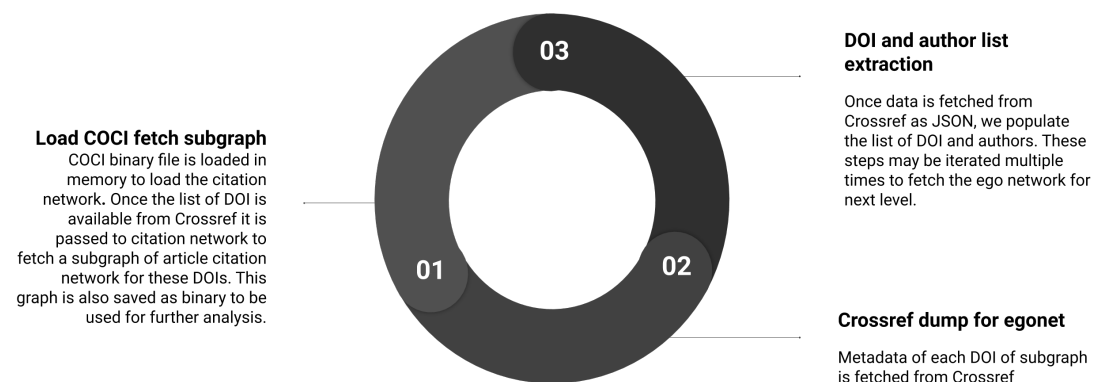
Figure 5. Step 3 of the workflow with details of creating the ego network. Sub-steps are applied
sequentially, and may be iterated to create the next level of ego network.

292 **Load COCI binary to fetch subgraph**   After loading a binary file of COCI, a subset of the graph is
293 fetched with nodes linked at one level apart. These nodes are either cited-by or cite the existing articles.
294 Processing a subgraph from 625M edges takes a few minutes on a Core i5 laptop with 16 GB RAM.

295 **Crossref dump for egonet**   Crossref data is fetched for all DOIs of article ego network created in the
296 previous step. First, all data is downloaded and then it is processed to create the network. Depending on
297 the size of the network, the number of ego levels, and connectivity bandwidth this process may continue
298 from hours to days. Once a local copy of data is available this delay can be reduced. Since we do not have
299 access to complete dump of Crossref, we could not identify that whether these same scripts can be reused.
300 We assume that there would be few changes required to access the data locally.

301 **DOI and author list extraction**   Processing of ego network for authors is similar to nodes of SCIM
302 downloaded earlier. However, the connecting nodes fetched in subgraph above are added and their
303 respective author details are downloaded.

Creating scientific network(s)

305 Summary of the sub-steps to create scientific networks are shown in Figure 6. Once all the data are
306 pre-processed this step creates different types of network. We can also add bibliographic coupling and

307 co-citation network within the list but they are ignored for two reasons. First, we did not find much
308 evidence of centrality analysis on these networks. Secondly, the processing time for creating these
309 networks for a very large citation network is relatively much longer than creating author collaboration
310 or author citation network. These networks are created by making an edge list for authors who have
311 collaborated or cited each other, respectively.
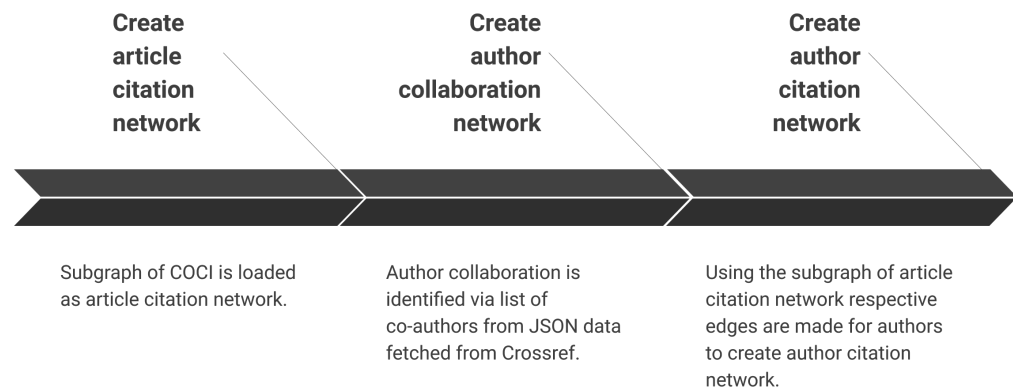
Figure 6. Step 4 of the workflow with details of creating different scientific networks. Sub-steps are applied sequentially.

312 **Create article citation network**   Once the list of DOI is available it is used to fetch subgraph of article
313 citation network. Article citation network is fetched as a subgraph from COCI. Further, it saves the same
314 graph as a binary file for further analysis. Also, the CSV file can be used with any graph processing
315 library (such as NetworkX) while binary file can be read using SNAP.

316 **Create author collaboration network**   Author collaboration is identified by the list of co-authors from
317 JSON data fetched via Crossref. This refined data are further used in the case study in the subsequent
318 section. It is important to note that the count of authors at this sub-step may vary from next sub-step of
319 creating author citation network. The list of co-authors in Crossref is available as a list of names and do
320 not include further metadata about these authors.

321 **Create author citation network**   Using the subgraph of article citation network respective edges for
322 authors are made to create author citation network. All co-authors are linked to use full counting method.
323 In order to utilise partial counting method, this script needs to be modified. However, our workflow is not
324 affected by the use of a partial or full counting method and hence we have picked simpler one for brevity
325 (Glanzel, 2003). This network shall supplement the analysis on the article citation network, as well as, the
326 collaboration network that was constructed in the previous step.

Centrality analysis

328 Summary of the sub-steps to apply centrality analysis are shown in Figure 7. Below, we define the
329 sub-steps to apply different centrality measures on the scientific networks. Based on numerous studies
330 discussed in Table 1 it is evident that centrality measures are a popular way of identifying prominent
331 authors. This is one of the common method employed in the bibliometric analysis and other methods of
332 SNA can also be applied at this step. Any tool or wrapper API may restrict the functionality at this point,
333 however, this work can be extended to use any functions in existing network processing libraries.

334 **Applying centrality measures on article citation network**   The article citation network is a Directed
335 Acyclic Graph (DAG). Two measures are presented since most centrality analyses are not meaningful
336 on DAG. Firstly, degree centrality gives highly cited articles. Secondly, influence definition in DAG is
337 captured by Katz centrality.
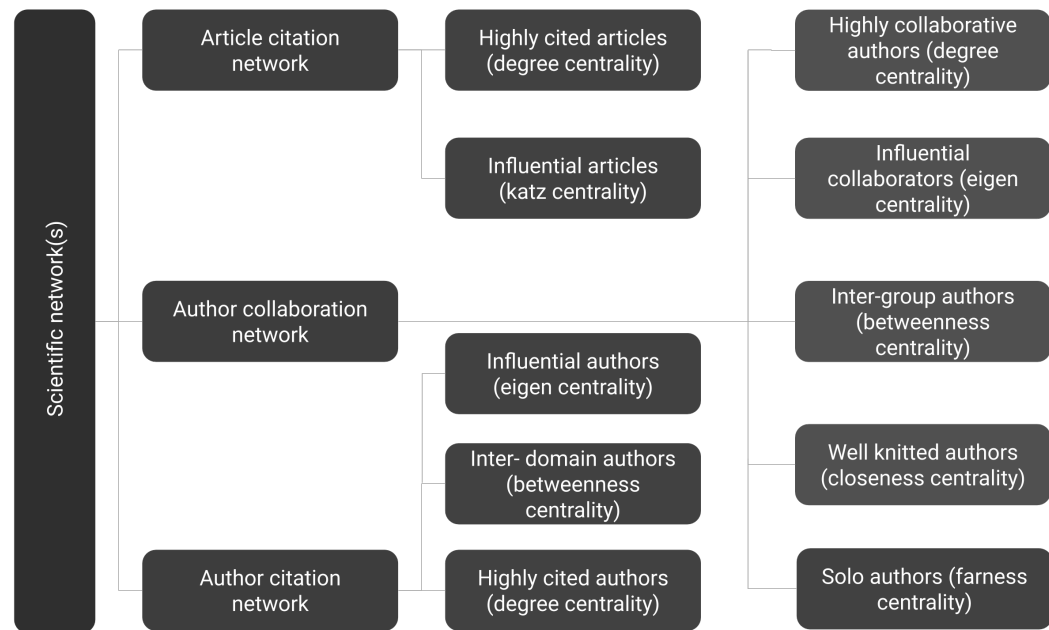
Figure 7. Step 5 of the workflow with details of centrality measures that are applied on different scientific networks. Sub-steps may be applied as required since there is no dependency within steps.

338 **Applying centrality measures on author citation network**   The author citation network has cyclic
339 nature. Three measures are presented, namely highly cited authors (degree centrality), influential authors
340 (eigenvector centrality), and authors working in multiple domains (betweenness centrality).

341 **Applying centrality measures on author collaboration network**   The author collaboration network
342 has cyclic nature and most centrality analyses are possible. Five measures are presented, namely highly
343 collaborative authors (degree centrality), influential collaborators (eigenvector centrality), authors working
344 in multiple groups (betweenness centrality), well-knitted authors (closeness centrality), and solo authors
345 (farness centrality). Ranks captured here are presented in Table 2. This work was done manually by
346 sorting individual lists on respective centrality scores and identifying their rank position.

Batch execution

348 All python scripts can be executed through a sample batch file by modifying the ISSN and date range. This
349 batch processing will also be useful for developing a front-end tool, as well as modifying the sequence as
350 per need. Details are available at (Butt and Faizi, 2020).

CASE STUDY USING SCIM

352 Milojević (2014) analysed collaboration network using WoS data of SCIM. Scores were calculated for
353 degree, betweenness, eigenvector, and closeness centrality. The outcome of their analysis was provided in
354 a table having authors that had top five ranks in either of the centrality scores. The respective rank of those
355 authors for each centrality was also listed. To verify whether or not our workflow can capture a similar
356 pattern we provide the results in a similar tabular form. For each of the centrality measure we provide
357 the rank given in (Milojević, 2014) using WoS data. It is aligned with the rank obtained in our study
358 using OpenCitatons data. We observe that the rank of authors for the degree, betweenness and closeness
359 centrality is more or less similar. Further analysis is required to inquire the reason for the difference of
360 eigenvector centrality ranks. Such an analysis is outside the scope of this study.

361   Further, we do not discuss the reasons for why a specific author has topped the list and what the
362 centrality measure signifies, and the interested reader is referred to (Milojević, 2014). Ranks in Table 2
363 were entered manually after execution of the workflow. Author names are placed in the same sequence
364 as per the original study along with their respective ranks. Table 2 has four sections for the degree,

betweenness, eigenvector and closeness centrality, respectively. Each section has two columns with the
left column showing rank from Milojević (2014) and the right column shows the rank calculated for the
same author using our workflow. It is pertinent to note that a very hand-on approach is given in Milojević
(2014). However, due to access restriction of WoS and its unaffordability for developing countries such
useful analysis is only limited to researchers of specific institutes having WoS subscription (Waltman and
Larivière, 2020). This highlights the importance of our workflow to provide access to any individual who
can download the publicly available metadata.

Collaborator Name	Degree Rank		Betweenness Rank		Eigenvector Rank		Closeness Rank	
	Prev	Curr	Prev	Curr	Prev	Curr	Prev	Curr
Glanzel, W	1	1	1	1	1	2	1	1
Rousseau, R	2	2	3	2	16	1	5	2
DeMoya-Anegon, F	3	8	12	20	26	16	6	15
Klingsporn, B	4	3	21	16	89	121	174	144
Ho, Ys	5	5	22	125	2096	3	613	575
Thijs, B	63	51	65	44	2	30	10	1710
Schubert,A	36	48	38	28	3	18	27	24
Debackere, K	6	6	16	15	4	7	4	5
Schlemmer, B	670	832	382	962	5	808	33	37
Meyer, M	43	39	2	4	14	9	2	3
Leydesdorff, L	54	35	4	3	46	5	42	44
Rafols,I	1058	387	5	23	83	239	45	49

Table 2. Comparison of ranks by previous study (Milojević, 2014) and current study. Table is divided
into 4 sections for each centrality measure. The left column in each section showing the rank from
(Milojević, 2014), and the right column showing the rank calculated by our workflow.

CONCLUSION AND FUTURE WORK

Scientific networks rely on completion of data (Zingg et al., 2020). Although the field has existed for more
than 50 years (Price, 1965) but the limitations on data access have not helped to reach its true potential.
We aim that with the availability of publicly available citation metadata (Waltman and Larivière, 2020)
and a scripted workflow to access it (Butt and Faizi, 2020) a researcher from any field will be able to
analyse the scientific networks. Its application can be vast, from identifying reviewers for a manuscript
(based on article's references) to a graduate student finding a supervisor (through collaboration network).
The time it takes to completely execute the workflow scripts is well under an hour, barring the two time
intensive steps. First, saving citation index as a binary file which needs to be done only once. Second,
downloading Crossref DOI files for individual nodes of ego-centered network can be optimised with a
local copy. The workflow provides a means for fast and interactive analysis.

Since using graphical tools is easier than executing the scripts so a future application of this study is to
create a front-end tool. A web-based portal is also under construction where the user may be able to select
the date range along with other filters, and the system will initiate the scripts at the back-end. This way
the researchers who are not familiar with programming can also benefit. It would enhance the capability
and usefulness of this workflow. Techniques for author name disambiguation and partial counting have
not been included. For effective analysis these need to be incorporated in future.

REFERENCES

Abbasi, A., Hossain, L., and Leydesdorff, L. (2012). Betweenness centrality as a driver of preferential
attachment in the evolution of research collaboration networks. *Journal of Informetrics*, 6(3):403–412.

AlNoamany, Y. and Borghi, J. A. (2018). Towards computational reproducibility: researcher perspectives on the use and sharing of software. *PeerJ Computer Science*, 4:e163.

Baas, J., Schotten, M., Plume, A., Côté, G., and Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, 1(1):377–386.

Birkle, C., Pendlebury, D. A., Schnell, J., and Adams, J. (2020). Web of science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, 1(1):363–376.

Butt, B. H. and Faizi, S. (2020). Guru workflow scripts. https://zenodo.org/record/4268321.

Chen, C. (2005). The centrality of pivotal points in the evolution of scientific networks. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, IUI '05, page 98–105, New York, NY, USA. Association for Computing Machinery.

Daquino, M., Tiddi, I., Peroni, S., and Shotton, D. (2018). Creating Open Citation Data with BCite. In *Enabling Open Semantic Science - Proceedings of the 2nd Workshop on Enabling Open Semantic Science (SemSci 2018)*, volume 2184 of *CEUR Workshop Proceedings*, pages 1–11, Monterey, CA, USA. CEUR-WS.

Ding, Y. (2011). Applying weighted pagerank to author citation networks. *Journal of the American Society for Information Science and Technology*, 62(2):236–245.

Glanzel, W. (2003). Bibliometrics as a research field: A course on theory and application of bibliometric indicators.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J., editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

Heibi, I., Peroni, S., and Shotton, D. (2019). Crowdsourcing open citations with CROCI – An analysis of the current status of open citations, and a proposal. In *Proceedings of the 17th International Conference on Scientometrics and Informetrics*. ISSI.

Heibi, I., Peroni, S., and Shotton, D. (2019). Software review: Coci, the opencitations index of crossref open doi-to-doi citations. *Scientometrics*, 121(2):1213–1228.

Hendricks, G., Tkaczyk, D., Lin, J., and Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1):414–427.

Herzog, C., Hook, D., and Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, 1(1):387–395.

Iorio, A. D., Peroni, S., and Poggi, F. (2019). Open data to evaluate academic researchers: an experiment with the Italian Scientific Habilitation. In *Proceedings of the 17th International Conference on Scientometrics and Informetrics*.

Kamińska, A. M. (2018). Plos one-a case study of quantitative and dynamic citation analysis of research papers based on the data in an open citation index (the opencitations corpus). *Bibliotheca Nostra. Śląski Kwartalnik Naukowy*, pages 168–186.

Lee, D. H. (2019). Predicting the research performance of early career scientists. *Scientometrics*, 121(3):1481–1504.

Leskovec, J. and Sosič, R. (2016). Snap: A general-purpose network analysis and graph-mining library. *ACM Trans. Intell. Syst. Technol.*, 8(1).

Leydesdorff, L., Wagner, C. S., and Bornmann, L. (2018). Betweenness and diversity in journal citation networks as measures of interdisciplinarity—a tribute to eugene garfield. *Scientometrics*, 114(2):567–592.

Li, W., Aste, T., Caccioli, F., and Livan, G. (2019). Early coauthorship with top scientists predicts success in academic careers. *Nature communications*, 10(1):1–9.

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., and López-Cózar, E. D. (2020). Google scholar, microsoft academic, scopus, dimensions, web of science, and opencitations' coci: a multidisciplinary comparison of coverage via citations. *Scientometrics*.

Massucci, F. A. and Docampo, D. (2019). Measuring the academic reputation through citation networks via pagerank. *Journal of Informetrics*, 13(1):185–201.

Milojević, S. (2014). Network analysis and indicators. In *Measuring scholarly impact*, pages 57–82. Springer.

Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., and Cobo, M. J. (2020). Software tools

447   for conducting bibliometric analysis in science: An up-to-date review. *El profesional de la información*
448   *(EPI)*, 29(1).

449   Newman, M. (2010). Measures and metrics. In *Networks: An Introduction*. Oxford University Press.

450   Newman, M. E. (2003). Ego-centered networks and the ripple effect. *Social Networks*, 25(1):83–95.

451   Newman, M. E. (2004). Who is the best connected scientist? a study of scientific coauthorship networks.
452   In *Complex networks*, pages 337–370. Springer.

453   Nishioka, C. and Färber, M. (2019). Evaluating the availability of open citation data. *BIRNDL 2019 :*
454   *Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries*,
455   2414.

456   OpenCitations (2020). COCI CSV dataset of all the citation data. last accessed on 13/03/2020.

457   Ortega, J. L. (2014). Influence of co-authorship networks in the research impact: Ego network analyses
458   from microsoft academic search. *Journal of Informetrics*, 8(3):728 – 737.

459   Peroni, S., Dutton, A., Gray, T., and Shotton, D. (2015). Setting our bibliographic references free: towards
460   open citation data. *Journal of Documentation*.

461   Peroni, S. and Shotton, D. (2020). Opencitations, an infrastructure organization for open scholarship.
462   *Quantitative Science Studies*, 1(1):428–444.

463   Price, D. J. D. S. (1965). Networks of scientific papers. *Science*, 149(3683):510–515.

464   Rose, M. E. and Kitchin, J. R. (2019). pybliometrics: Scriptable bibliometrics using a python interface to
465   scopus. *SoftwareX*, 10:100263.

466   Singh, C. K. and Jolad, S. (2019). Structure and evolution of indian physics co-authorship networks.
467   *Scientometrics*, 118(2):385–406.

468   Van den Besselaar, P. and Sandström, U. (2019). Measuring researcher independence using bibliometric
469   data: A proposal for a new performance indicator. *PloS one*, 14(3).

470   Van Eck, N. J. and Waltman, L. (2014). Visualizing bibliometric networks. In *Measuring scholarly*
471   *impact*, pages 285–320. Springer.

472   Waheed, W., Imran, M., Raza, B., Malik, A. K., and Khattak, H. A. (2019). A hybrid approach toward
473   research paper recommendation using centrality measures and author ranking. *IEEE Access*, 7:33145–
474   33158.

475   Waltman, L. and Larivière, V. (2020). Special issue on bibliographic data sources. *Quantitative Science*
476   *Studies*, 1(1):360–362.

477   Waltman, L. and Yan, E. (2014). Pagerank-related methods for analyzing citation networks. In *Measuring*
478   *scholarly impact*, pages 83–100. Springer.

479   Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., and Kanakia, A. (2020). Microsoft academic graph:
480   When experts are not enough. *Quantitative Science Studies*, 1(1):396–413.

481   Xu, Z. and Pekelis, L. B. (2015). A survey of chinese interpreting studies: Who influences who. . . and
482   why? *PeerJ Computer Science*, 1:e14.

483   Zhu, Y., Yan, E., Peroni, S., and Che, C. (2020). Nine million book items and eleven million citations: a
484   study of book-based scholarly communication using opencitations.

485   Zingg, C., Nanumyan, V., and Schweitzer, F. (2020). Citations driven by social connections? a multi-layer
486   representation of coauthorship networks. *Quantitative Science Studies*, 1(4):1493–1509.

ACKNOWLEDGEMENTS

488   Authors acknowledge lengthy discussions with Rauf Shams Malik and Azmi Umer that shaped this
489   study, as well as, the efforts of students Sufyan Faizi, Adnan Anwar, Talha Aslam, Naveed Nizar, Sahil
490   Shahbuddin and Anas ur Rehman in developing the Python scripts. Also acknowledge the invaluable
491   comments of reviewers and editor that shaped this study.