RESEARCH ARTICLE

Reproducible science of science at scale: *pySciSci*

Alexander J. Gates[1] and **Albert-László Barabási**[2,3]

[1]School of Data Science, University of Virginia, Charlottesville, VA
[2]Network Science Institute, Northeastern University, Boston, MA
[3]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA

Keywords: bibliometric data, citation networks, code, research assessment, science of science, scientometrics

ABSTRACT

Science of science (SciSci) is a growing field encompassing diverse interdisciplinary research programs that study the processes underlying science. The field has benefited greatly from access to massive digital databases containing the products of scientific discourse—including publications, journals, patents, books, conference proceedings, and grants. The subsequent proliferation of mathematical models and computational techniques for quantifying the dynamics of innovation and success in science has made it difficult to disentangle universal scientific processes from those dependent on specific databases, data-processing decisions, field practices, etc. Here we present *pySciSci*, a freely available and easily adaptable package for the analysis of large-scale bibliometric data. The *pySciSci* package standardizes access to many of the most common data sets in SciSci and provides efficient implementations of common and advanced analytical techniques.

1. INTRODUCTION

Science of science (SciSci) as a discipline has grown rapidly over the last century, reflecting an increasing interest in quantitatively modeling the processes underlying science—from the novelty of scientific discoveries to the interconnectivity of scientists. The increasing prevalence of SciSci research is due in large part to the availability of large-scale bibliometric data capturing the products of scientific discourse, including publications, patents, and funding. Jointly with the analysis of scientific processes, such bibliometric data are used to map the evolution of specific fields, evaluate scientific performance and eminence, and support government policy and funding decisions (Fortunato, Bergstrom et al., 2018; Wang & Barabási, 2021; Wu, Kittur et al., 2022). However, bibliometric data are distributed across diverse databases, each with its own criteria for inclusion, and varied processes to assure the data's quality and accuracy (Csiszar, 2017). The manifold uses and applications for bibliometric data, combined with the call for reproducible and replicable science, has prompted the need for flexible analysis that is reliably reproduced across multiple data sets.

Here, we introduce *pySciSci*, an open-source Python package for the analysis of larges-cale bibliometric data. The *pySciSci* package provides

- standardized preprocessing and access to many of the most common data sets in SciSci;
- an extensive library of quantitative measures fundamental to SciSci; and
- advanced methods for mapping bibliometric networks.

an open access journal

Citation: Gates, A. J., & Barabási, A.-L. (2023). Reproducible science of science at scale: *pySciSci*. *Quantitative Science Studies*. Advance publication. https://doi.org/10.1162/qss_a_00260

DOI: https://doi.org/10.1162/qss_a_00260

Peer Review: https://www.webofscience.com/api/gateway/wos/peer-review/10.1162/qss_a_00260

Received: 7 September 2022
Accepted: 8 March 2023

Corresponding Author:
Alexander J. Gates
agates@virginia.edu

Handling Editor:
Ludo Waltman

The MIT Press

The *pySciSci* package is intended for researchers of SciSci working from complete bibliometric databases or those who wish to integrate large-scale bibliometric data into other existing projects. By creating a standardized and adaptable programmatic base for the study of bibliometric data, we intend to help democratize SciSci, support diverse research efforts based on bibliometric data sets, and address calls for open access and reproducibility in the SciSci literature and community (Light, Polley, & Börner, 2014).

To the best of our knowledge, our package constitutes one of the most comprehensive collections of methods and data sources in scientometrics and bibliometrics. It complements and extends the capabilities of the *Bibliometrix* (Aria & Cuccurullo, 2017), *BiblioTools* (Grauwin & Jensen, 2011), and *Citan* (Gagolewski, 2011) libraries to multiple databases and more advanced metrics. Although two of the most popular bibliometric programs, *VOSviewer* (van Eck & Waltman, 2010) and *CiteSpace* (Chen, 2006), are designed to provide graphical network maps of science, neither program is open sourced and modifiable. Several programs are much more specialized than *pySciSci*, and focus on implementations of method families for specific tasks (Moral-Muñoz, Herrera-Viedma et al., 2020); for example, *CRXexplorer* analyzes a publication's distribution of reference years (Marx, Bornmann et al., 2014), and the open-source Python package *ScientoPy* offers tools specifically for topical trend analysis (Ruiz-Rosero, Ramırez-González, & Viveros-Delgado, 2019). Our package also complements the *CADRE* (Mabry, Yan et al., 2020) environment built to host bibliometric data sets. Ultimately, our goal is not to supplant these other efforts to provide access to SciSci research but to facilitate a unified and generalizable open-source environment across different databases and methods of analysis.

2. THE *pySciSci* PACKAGE

The *pySciSci* package is built around Python Pandas data frames (McKinney, 2010), providing the simplicity of Python with the increased speed of SQL relational databases. *pySciSci* provides a standardized interface for working with several of the major data sets in the Science of Science, including the Microsoft Academic Graph (MAG), the Web of Science (WOS), the American Physics Society (APS), PubMed, the DBLP Computer Science Bibliography (DBLP), and OpenAlex (Priem, Piwowar, & Orr, 2022). Each data set is referenced in *pySciSci* as a customized variant of the *BibDataBase* class, which handles all data loading and preprocessing. For an example of loading and preprocessing each database, we include a Getting Started jupyter notebook in the *examples* directory. The storage and processing frameworks are highly generalizable, and can be extended to other databases not mentioned above (e.g., United States Patent Office, Scopus, Lens).

The *pySciSci* pipeline starts by preprocessing raw data into a standardized tabular format (Figure 1). The package creates several relational data tables based on a balance between commonly associated data fields and memory footprint. Bibliometric records are split into five types of entities: publication, author, affiliation (institution), journal/venue, and field of study. The primary unit of analysis in *pySciSci* is the publication—a catch all phrase encompassing scientific articles, preprints, patents, books, conference papers, and other bibliometric products disseminated as a single entry in a database. The publication objects are stored in their own data table, *publication*. As the year of publication is the most commonly used publication property, the mapping of publications to year is also replicated in its own Python dictionary, *pub2year*, for quick reference. Depending on the specific database, the author names, affiliation names, and journal names may be available and are stored in their own data tables: *author*, *affiliation*, and *journal* respectively. In some databases, data fields represent expert

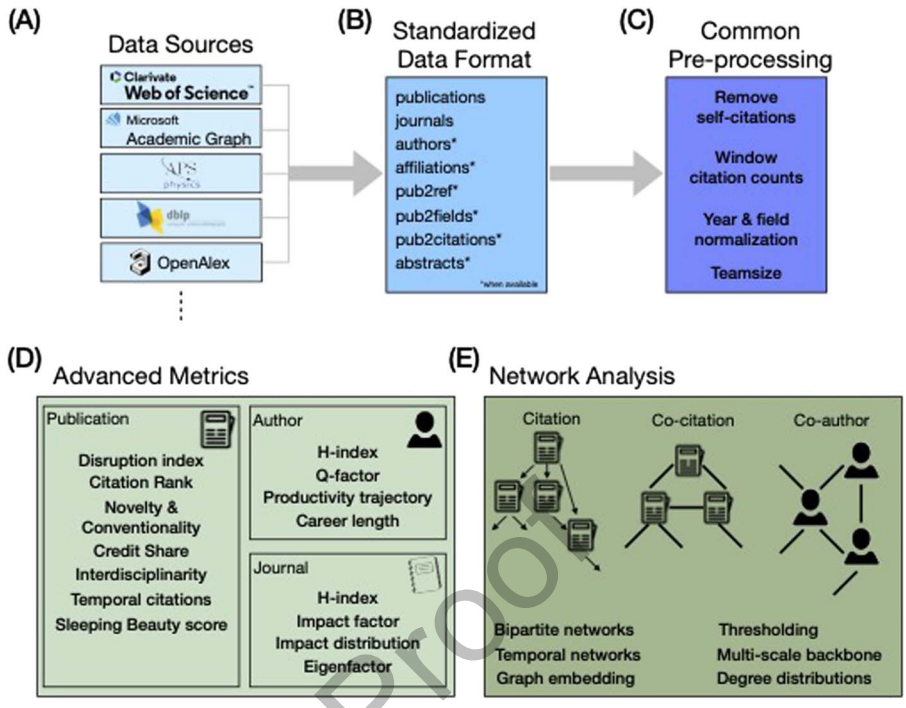

Figure 1. Data processing overview. The *pySciSci* package preprocesses many of the common bibliometric data sources (A) into a standardized set of relational tables (B). The package also cleans and precomputes measures that are frequent building blocks for more advanced computations (C). The *pySciSci* package provides efficient implementations for many advanced metrics focusing on publications, authors, or journals (D), as well as advanced network analysis (E).

curated entries, and in other databases, data fields may be algorithmically inferred by the database curators; see the specific database references for details. Finally, three relational tables are built to link between the entities: *pub2ref* captures reference and citation relationships between publications; *publicationauthoraffilliation* links between publications, their authors, and the author affiliations; and *pub2field* links publications to their field of study. The preprocessing step which builds the data tables only needs to be run once for each database.

After extracting the data tables, the *pySciSci* package precomputes several of the most common and useful bibliometric properties that form the backbone of many more advanced methods. For example, if the author information is available, the team size (number of authors) is found for all publications (Wuchty, Jones, & Uzzi, 2007). When the reference/citation information is available, the number of citations within a user defined window (default 10 years) is also precomputed (Wang, 2013). Finally, when both the author and reference/citation information is available, the *pySciSci* package will archive a copy of the reference/citation relationships in which self-citations are removed, *pub2ref_noself*.

To facilitate data movement and lower memory overhead when the complete tables are not required, the *pySciSci* preprocessing step chunks the data tables into smaller tables. When loading a table into memory, the user can quickly load the full table by referencing the table name as a database property or specify multiple filters to load only a subset of the data. The *pySciSci* also supports dask dataframes (Rocklin, 2015), which add parallelization and block scheduling, allowing large dataframes to be processed without loading the full dataframe into memory.

Due to variations in data coverage between databases, the available package functionality will vary between data sets. For example, the DBLP database does not provide citation relationships between publications, and the APS database does not disambiguate author careers. The *pySciSci* package supports methods to link bibliometric entities between databases and the framework easily facilitates augmenting a database with additional data sources, allowing for enriched analysis (Gates, Gysi et al., 2021).

Our distribution of *pySciSci* is accompanied by a growing library of jupyter notebooks that illustrate its basic functionalities and usage. We also encourage the SciSci community to contribute their own implementations, data, use cases, or attempts to reproduce key results from the Science of Science.

3. PUBLICATIONS AND CITATIONS

The coverage of bibliometric databases varies, with some focusing only on a narrow subset of publications defined by journal or field, and others attempting to encompass all peer-reviewed scientific communication. As shown in Figure 2A, the number of publications and temporal coverage vary dramatically between four common databases. This variability reflects important decisions about data quality and generalizability that a researcher must make; for example, DBLP provides user-curated author careers in computer science, but does not contain citation information, whereas MAG contains a wide range of document types from all of science, with algorithmically inferred fields and author career information. With few exceptions, these databases focus on English-language publications, offering only sparse coverage of publications in other languages. The *pySciSci* package facilitates restricting each database to specific document types, fields, or years, allowing researchers more control over the publications and authors under study.

Citation analysis is the examination of the frequency, patterns, and networks of citation relationships between publications. Some citation measures have become commonplace, with many implementations available; others are precomputed by major database portals based on proprietary algorithms, and still others require complex processing and computational steps that have largely inhibited their general usage (Bollen, Van de Sompel et al., 2009). The *pySciSci* package facilitates the analysis of total citation counts for publications, as well as citation time series, fixed time window citation analysis, citation count normalization by year and field, and fractional citation contribution based on team size. Due to the package's modular design, the choice of citation count and normalization is made before calculating specific metrics. The package also includes a simplified interface for fitting models to citation time series, such as in the prediction of the long-term citation counts to a publication (Wang, Song, & Barabási, 2013), or in the assignment of the sleeping beauty score (Ke, Ferrara et al., 2015). Exemplar code illustrating citation metrics can be found in the examples folder.

Due to the prevalence of citation metrics as measures of scientific prominence, techniques for "gaming the system" have flourished that inflate an author's citation metrics for reasons other than scientific impact. For instance, it has been found that men tend to cite themselves more often than women, contributing to widening gender imbalances in scientific impact (King, Bergstrom et al., 2017). Consequently, one of the primary preprocessing steps for contemporary citation analysis is the removal of self-citations occurring between publications by the same author. All analysis facilitated in the *pySciSci* package can be run either with or without the self-citations when authors are available in the database.

The comparison of citation counts between different disciplines and fields is complicated by differing citation norms and community sizes (Radicchi, Fortunato, & Castellano, 2008). Therefore, it is common to normalize citation counts by field or year averages to create a

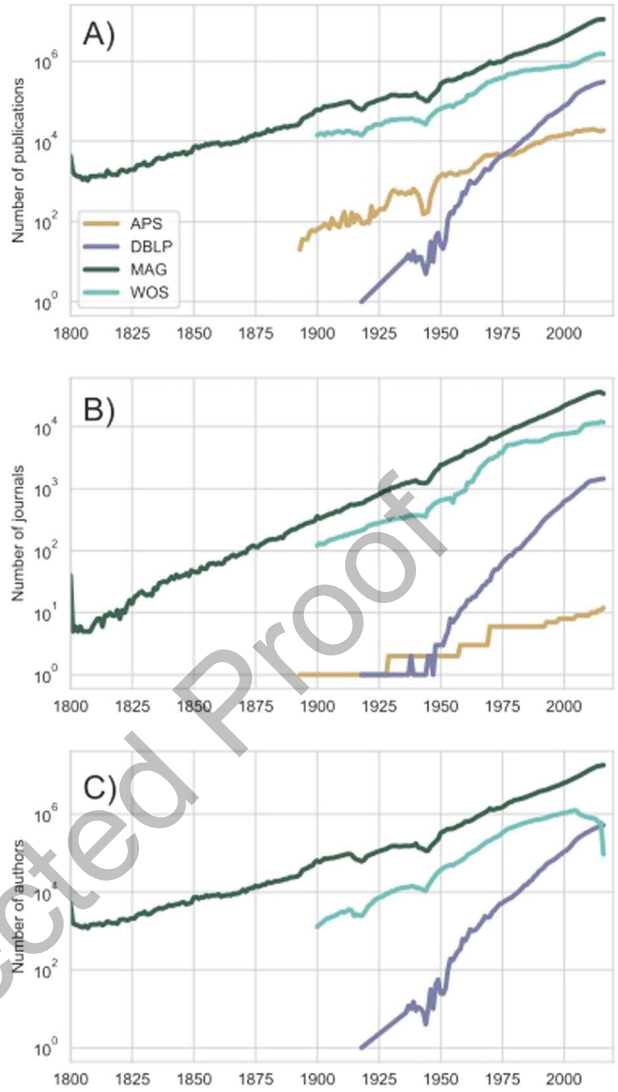

Figure 2. Growth of science across databases. The database size as measured by the number of A) publications, B) journals, and C) authors varies over several orders of magnitude between the APS (gold), DBLP (purple), MAG (dark green), and WOS (teal). As the APS does not include disambiguated author careers, it does not appear in C).

common reference point, or to rank publications to identify "top publications" in the top 1% or 5% of publications from a field. The *citation_rank* function facilitates the ranking of publications by different citation metrics and groups. We also provide extended normalization measures that account for a publication's interdisciplinarity by controlling for citation patterns in the immediate cocitation neighborhood.

The diversity of disciplines or journals reflected in a publication's reference and citation relationships has been used to quantify the publication's interdisciplinarity or novelty (Gates, Ke et al., 2019; Porter & Rafols, 2009; Stirling, 2007; Uzzi, Mukherjee et al., 2013). The *pySciSci* package provides several measures of interdisciplinarity, including the Rao-Stirling diversity index, the Gini coefficient, Simpson's diversity index, and entropy measures, which can be computed using the distribution of publication references or publication citations. For example, consider the publication shown in Figure 3A, with five references in three disciplines: physics,

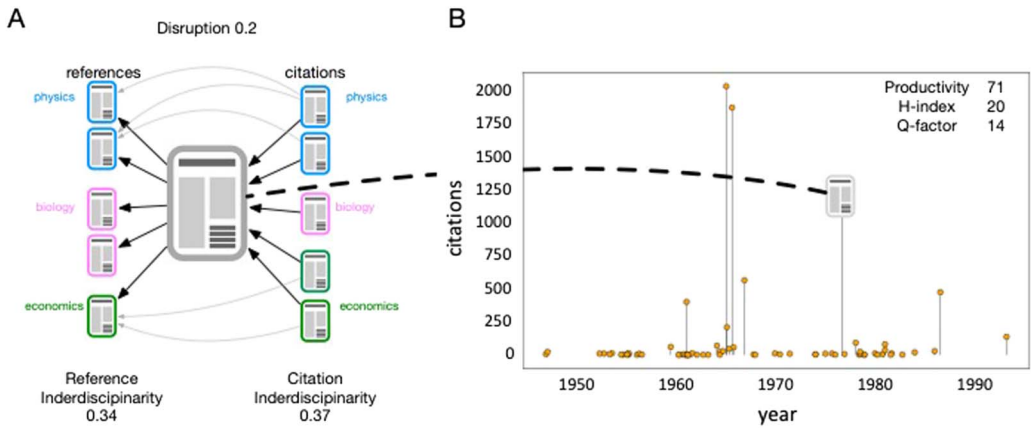

Figure 3. Advanced career and publication metrics. (A) The pySciSci package captures the several advanced characterizations of publication's influence (references) and impact (citations) including the disruption index, and Rao-Stirling Interdisciplinarity. (B) The package also facilitates the analysis of full author careers and summarizing metrics such as total productivity, *h*-index, and *Q*-factor.

biology, and economics. The Rao-Stirling reference interdisciplinarity, calculated using the *raostriling_interdisciplinarity* function, reflects the diversity of the disciplines referenced by the publication (left, 0.34), and the Rao-Stirling citation interdisciplinarity reflects the diversity of the disciplines citing the publication (right, 0.37). The *pySciSci* package also facilitates the computation of publication novelty and conventionality as measured by atypical combinations of journals in the reference list using the *novelty_conventionality* function (Uzzi et al., 2013). Other measures based on the local citation graph capture the disruptive influence of a publication as measured by the frequency with which the publication is cited alongside its own references (Funk & Owen-Smith, 2017; Park, Leahey, & Funk, 2023; Wu, Wang, & Evans, 2019), calculated by the *disruption_index* function. For example, in Figure 3A, four of the citing publications also cite three of the references, resulting in a disruption index of 0.2. Exemplar code for the analysis of publication interdisciplinarity can be found in the examples folder.

4. PUBLICATION GROUPS: AUTHORS, JOURNALS, FIELDS AND AFFILIATIONS

The next unit of analysis aggregates publications into groups by common author, journal, discipline/field, or affiliation. For example, the infamous journal impact factor considers the group of all publications from the same journal over a fixed time window (typically 2, 3, or 5 years), and is found by averaging their citation counts (Bordons, Fernández, & Gómez, 2002). The *pySciSci* package implements over 12 citation metrics for groups of publications, which can be easily applied to journal, author, discipline/field, or affiliation aggregations when available in the database. Combined with the different normalization decisions for citation counts, the *pySciSci* package implements nearly 200 different measures for scientific impact.

At the heart of scientific discoveries are the scientists themselves. Consequently, the sociology of science has analyzed scientific careers in terms of individual incentives, productivity, competition, collaboration, and success. The *pySciSci* package facilitates author career analysis through both aggregate career statistics and temporal career trajectories. We implement more than 10 metrics for author citation analysis, including the *h*-index (Hirsch, 2005), *author_hindex*, and *Q*-factor (Sinatra, Wang et al., 2016), *author_qfactor*. The package also includes a simplified interface for fitting models to author career trajectories, such as identifying topic switches (Zeng, Shen et al., 2019), the assessment of yearly productivity patterns (Way, Morgan et al., 2017), or the hot-hand effect (Liu, Wang et al., 2018).

For example, consider the representation of Derek de Solla Price's publication career as represented in the MAG, shown in Figure 3B. It captures the citations recieved by 71 articles and books published over 50 years (even though Dr. de Solla Price died in 1983, articles can be reprinted or published posthumously). Using this career trajectory, we find that Dr. de Solla Price has an *h*-index of 20 and a *Q*-factor of 14. Exemplar code for the analysis of author careers can be found in the examples folder.

Greater scrutiny is being given to the prevalence of systematic bias in science (Saini, 2019), supported by observations that, for example, female authors have fewer publications than their male colleagues (Larivière, Ni et al., 2013; Xie & Shauman, 1998). Although most databases do not include author biographical information (gender, race, age, position, sexual orientation, etc.), the *pySciSci* package facilities linking user provided biographical information to author careers. Implementations are then available for advanced measures of inequality, including the measurement of categorical bias in reference lists (Dworkin, Linn et al., 2020), or career lengths (Huang, Gates et al., 2020).

In addition, the movement of scientists between institutions and countries requires longitudinal data capturing the changes in affiliation throughout a career. When the affiliations are disambiguated, the *pySciSci* package allows for collaboration and mobility networks between affiliations. These affiliations can be aggregated to the city, state, and country level, allowing for large-scale analysis of global patterns in scientific production and impact.

5. NETWORK ANALYSIS

Scientific discoveries and careers do not exist in isolation; rather, science evolves as a conversation between scientists, empowered by links between authors, publications, institutions, and other entities. Consequently, many key results from SciSci consider publications, authors, or fields as embedded in a complex web of interrelationships. The *pySciSci* package provides a flexible interface for working with networked bibliometric data. First, the bibliometric relationships are processed to extract the edge list representation of the network. The package then maps these edge lists to an adjacency matrix, treated internally as a scipy sparse matrix—a memory-efficient and highly flexible network representation. All network relationships can be further unraveled over time by considering snapshots of the network for each year. *pySciSci* facilities basic network measures, including the number of connected components, extraction of the largest connected component, threshold filtering, disparity filter (Serrano, Boguná, & Vespignani, 2009), and analysis of degree distributions (Barabási, 2016). The scipy sparse adjacency matrix can also be directly imported into many of the most common packages for more advanced network analysis and visualization.

One of the most common bibliometric networks is the coauthorship network, in which nodes represent authors and two authors are linked if they coauthored a publication (Barabási, Jeong et al., 2002; Gold, Gates et al., 2022; Newman, 2004). Coauthorship networks are used to capture general patterns of collaboration including how many different people an author publishes with, how often an author's collaborators are also each-other's collaborators (network clustering), what the typical networked-based distance between authors is (average path length), and how patterns of collaboration vary between fields and over time. Given a subset of publication and author relations, the *coauthorship_network* function can build both the static and temporal coauthorship networks. Exemplar code for the analysis of coauthorship networks can be found in the examples folder.

The scientific community's perception of which publications are most related to each other is reflected in the publication cocitation network (Boyack & Klavans, 2010; Gates, Ke et al.,

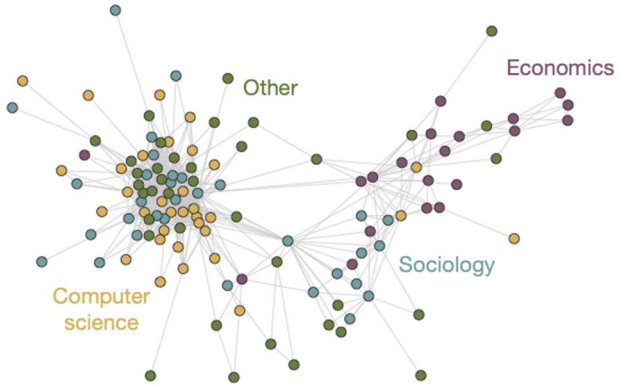

Figure 4. Cocitation network. The interdisciplinary impact of a publication is illustrated through the cocitation network between citing articles. Here nodes are publications that cited Stirling (2007). Two nodes are linked if some other publication cited both. Node color reflects the publication's discipline: (yellow) computer science, (magenta) economics, (blue) sociology, and (green) other. The three prominent clusters reflect the fact that Stirling (2007) impacted three distinct communities of researchers.

2019). Here, nodes represent publications and two publications are linked if they are both cited by another publication. For example, consider the cocitation network shown in Figure 4, in which nodes come from the set of publications that cite Stirling (2007). The cocitation network shows three distinct clusters of publications, each of which is enriched by a subset of related fields (*Computer Science*, *Economics*, *Sociology*), and a fourth that features *Other* publications. Indeed, modularity maximization using the Louvain heuristic (Blondel, Guillaume et al., 2008; Newman, 2006) identifies four communities. The similarity of the publication fields and the detected communities can be assessed using the element-centric similarity (Gates & Ahn, 2019; Gates, Wood et al., 2019), a measure between 0 and 1, where 1 captures that the two network communities are identical, and 0 reflects two network communities that group publications very differently. The element-centric similarity between the publication fields and the detected communities is 0.33, reflecting a modest level of agreement. Decomposing the error terms into contributions from publications in different fields, we find the the majority of the error arises from the *Other* publications (0.26), whereas publications in *Economics* and *Computer Science* are more faithfully recovered (0.45 and 0.35 respectively). This cocitation network analysis demonstrates how the diversity measure introduced in Stirling (2007) has impacted three distinct scientific communities. Exemplar code for the analysis of cocitation networks can be found in the examples folder.

Citation networks form the basis for collective measures of scientific impact. For example, the collective assignment of credit to a publication's authors can be measured by the frequency with which an author's other publications are cocited alongside the focus publication (Shen & Barabási, 2014). The *pySciSci* package algorithmically calculates the collective credit allocation temporally for each year since the article's publication.

Advance in statistical learning methods for graph embedding allow networks to be represented in high-dimensional metric spaces (Goyal & Ferrara, 2018). Such graph embedding methodologies provide compressed representations of the original network that can be used, for example, to predict new connections based on node similarities (Martinez, Berzal, & Cubero, 2016). The *pySciSci* package provides implementations of the *node2vec* (Grover & Leskovec, 2016) graph embedding method and its extension for authors, *persona2vec* (Yoon, Yang et al., 2021), which produces effective representations of scientific journals (Peng, Ke

et al., 2021) and author mobility (Murray, Yoon et al., 2020). Exemplar code for graph embedding can be found in the examples folder.

6. SUMMARY AND DISCUSSION

Here we introduced the open-source *pySciSci* Python package for bibliometric data analysis. Due to its modular structure, the *pySciSci* framework is highly generalizable and can easily accommodate many available data sets beyond the four mentioned here. The package also provides efficient implementations of common and advanced SciSci methods, facilitating reproducible analysis across multiple data sets. Most importantly, it is our hope that this package stimulates other researchers to add their own methods and facilitates large-scale collaborations throughout the Science of Science community.

ACKNOWLEDGMENTS

Special thanks to Jisung Yoon for contributing the implementation of *graph2vec* and Yong-Yeol Ahn for feedback on package structure. Thanks to two anonymous reviewers for wonderful suggestions that improved the presentation and functionality of the package, and thanks to the wonderful research community at the Center for Complex Network Research for helpful discussions.

AUTHOR CONTRIBUTIONS

Alexander J. Gates: Conceptualization, Data curation, Funding acquisition, Methodology, Software, Visualization, Writing—Original draft, Writing—Review & editing. Albert-László Barabási: Conceptualization, Funding acquisition, Writing—Review & editing.

COMPETING INTERESTS

A.-L.B. is co-scientific founder of and is supported by Scipher Medicine, Inc., which applies network medicine strategies to biomarker development and personalized drug selection, and is the founder of Naring Inc. that applies data science to health and nutrition.

FUNDING INFORMATION

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-19-1-0354. A.-L.B. is also supported by the Templeton Foundation, contract #62452, the European Union's Horizon 2020 research and innovation programme under grant agreement No 810115 – DYNASNET, by The Eric and Wendy Schmidt Fund for Strategic Innovation, Grant G-22-63228, and NSF grant SES-2219575.

CODE AND DATA AVAILABILITY

The *pySciSci* package and all code used to generate the figures in this publication can be found on github: https://github.com/SciSciCollective/pyscisci. The data sets referenced are freely or commercially available as described in the package documentation.

REFERENCES

Aria, M., & Cuccurullo, C. (2017). *Bibliometrix*: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Barabási, A.-L. (2016). *Network science*. Cambridge University Press.

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*(3–4), 590–614. https://doi.org/10.1016/S0378-4371(02)00736-7

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

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLOS ONE*, *4*(6), e6022. https://doi.org/10.1371/journal.pone.0006022, PubMed: 19562078

Bordons, M., Fernández, M., & Gómez, I. (2002). Advantages and limitations in the use of impact factor measures for the assessment of research performance. *Scientometrics*, *53*(2), 195–206. https://doi.org/10.1023/A:1014800407876

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404. https://doi.org/10.1002/asi.21419

Chen, C. (2006). Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377. https://doi.org/10.1002/asi.20317

Csiszar, A. (2017). The catalogue that made metrics, and changed science. *Nature*, *551*(7679), 163–165. https://doi.org/10.1038/551163a, PubMed: 29120444

Dworkin, J. D., Linn, K. A., Teich, E. G., Zurn, P., Shinohara, R. T., & Bassett, D. S. (2020). The extent and drivers of gender imbalance in neuroscience reference lists. *Nature Neuroscience*, *23*, 918–926. https://doi.org/10.1038/s41593-020-0658-y, PubMed: 32561883

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., ... Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), eaao0185. https://doi.org/10.1126/science.aao0185, PubMed: 29496846

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science*, *63*(3), 791–817. https://doi.org/10.1287/mnsc.2015.2366

Gagolewski, M. (2011). Bibliometric impact assessment with R and the CITAN package. *Journal of Informetrics*, *5*(4), 678–692. https://doi.org/10.1016/j.joi.2011.06.006

Gates, A. J., & Ahn, Y.-Y. (2019). CluSim: A Python package for calculating clustering similarity. *Journal of Open Source Software*, *4*(35), 1264. https://doi.org/10.21105/joss.01264

Gates, A. J., Gysi, D. M., Kellis, M., & Barabási, A.-L. (2021). A wealth of discovery built on the Human Genome Project—By the numbers. *Nature*, *590*(7845), 212–215. https://doi.org/10.1038/d41586-021-00314-6, PubMed: 33568828

Gates, A. J., Ke, Q., Varol, O., & Barabási, A.-L. (2019). Nature's reach: Narrow work has broad impact. *Nature*, *575*, 32–34. https://doi.org/10.1038/d41586-019-03308-7, PubMed: 31695218

Gates, A. J., Wood, I. B., Hetrick, W. P., & Ahn, Y.-Y. (2019). Element-centric clustering comparison unifies overlaps and hierarchy. *Scientific Reports*, *9*(1), 8574. https://doi.org/10.1038/s41598-019-44892-y, PubMed: 31189888

Gold, J. R., Gates, A. J., Haque, S. A., Melson, M. C., Nelson, L. K., & Zippel, K. (2022). The NSF ADVANCE Network of Organizations. *ADVANCE Journal*, *3*(1). https://doi.org/10.5399/osu/ADVJRNL.3.1.3

Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, *151*, 78–94. https://doi.org/10.1016/j.knosys.2018.03.022

Grauwin, S., & Jensen, P. (2011). Mapping scientific institutions. *Scientometrics*, *89*(3), 943–954. https://doi.org/10.1007/s11192-011-0482-y

Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 855–864). https://doi.org/10.1145/2939672.2939754, PubMed: 27853626

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102, PubMed: 16275915

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, *117*(9), 4609–4616. https://doi.org/10.1073/pnas.1914221117, PubMed: 32071248

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, *112*(24), 7426–7431. https://doi.org/10.1073/pnas.1424329112, PubMed: 26015563

King, M. M., Bergstrom, C. T., Correll, S. J., Jacquet, J., & West, J. D. (2017). Men set their own cites high: Gender and self-citation across fields and over time. *Socius*, *3*. https://doi.org/10.1177/2378023117738903

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, *504*, 211–213. https://doi.org/10.1038/504211a, PubMed: 24350369

Light, R. P., Polley, D. E., & Börner, K. (2014). Open data and open code for big science of science studies. *Scientometrics*, *101*(2), 1535–1551. https://doi.org/10.1007/s11192-014-1238-2

Liu, L., Wang, Y., Sinatra, R., Giles, C. L., Song, C., & Wang, D. (2018). Hot streaks in artistic, cultural, and scientific careers. *Nature*, *559*(7714), 396–399. https://doi.org/10.1038/s41586-018-0315-8, PubMed: 29995850

Mabry, P. L., Yan, X., Pentchev, V., Van Rennes, R., McGavin, S. H., & Wittenberg, J. V. (2020). CADRE: A collaborative, cloud-based solution for big bibliographic data research in academic libraries. *Frontiers in Big Data*, *3*, 556282. https://doi.org/10.3389/fdata.2020.556282, PubMed: 33693415

Martinez, V., Berzal, F., & Cubero, J.-C. (2016). A survey of link prediction in complex networks. *ACM Computing Surveys*, *49*(4), 1–33. https://doi.org/10.1145/3012704

Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology*, *65*(4), 751–764. https://doi.org/10.1002/asi.23089

McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *Profesional de la Información*, *29*(1). https://doi.org/10.3145/epi.2020.ene.03

Murray, D., Yoon, J., Kojaku, S., Costas, R., Jung, W.-S., ... Ahn, Y.-Y. (2020). Unsupervised embedding of trajectories captures the latent structure of mobility. *arXiv preprint*, arXiv:2012.02785. https://doi.org/10.48550/arXiv.2012.02785

Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, *101*(Suppl. 1), 5200–5205. https://doi.org/10.1073/pnas.0307545100, PubMed: 14745042

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*,

103(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103, PubMed: 16723398
Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, *613*(7942), 138–144. https://doi.org/10.1038/s41586-022-05543-x, PubMed: 36600070
Peng, H., Ke, Q., Budak, C., Romero, D. M., & Ahn, Y.-Y. (2021). Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Science Advances*, *7*, eabb9004. https://doi.org/10.1126/sciadv.abb9004, PubMed: 33893092
Porter, A., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719–745. https://doi.org/10.1007/s11192-008-2197-2
Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint*, arXiv:2205.01833. https://doi.org/10.48550/arXiv.2205.01833
Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272. https://doi.org/10.1073/pnas.0806977105, PubMed: 18978030
Rocklin, M. (2015). Dask: Parallel computation with blocked algorithms and task scheduling. In *Proceedings of the 14th Python in Science Conference* (pp. 130–136). https://doi.org/10.25080/Majora-7b98e3ed-013
Ruiz-Rosero, J., Ramırez-González, G., & Viveros-Delgado, J. (2019). Software survey: Scientopy, a scientometric tool for topics trend analysis in scientific publications. *Scientometrics*, *121*(2), 1165–1188. https://doi.org/10.1007/s11192-019-03213-w
Saini, A. (2019). *Superior: The return of race science*. Beacon Press.
Serrano, M. Á., Boguná, M., & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, *106*(16), 6483–6488. https://doi.org/10.1073/pnas.0808904106, PubMed: 19357301
Shen, H.-W., & Barabási, A.-L. (2014). Collective credit allocation in science. *Proceedings of the National Academy of Sciences*, *111*(34), 12325–12330. https://doi.org/10.1073/pnas.1401992111, PubMed: 25114238
Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science*, *354*(6312), aaf5239. https://doi.org/10.1126/science.aaf5239, PubMed: 27811240
Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, *4*(15), 707–719. https://doi.org/10.1098/rsif.2007.0213, PubMed: 17327202
Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468–472. https://doi.org/10.1126/science.1240474, PubMed: 24159044
van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3, PubMed: 20585380
Wang, D., & Barabási, A.-L. (2021). *The science of science*. Cambridge University Press. https://doi.org/10.1017/9781108610834
Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127–132. https://doi.org/10.1126/science.1237825, PubMed: 24092745
Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, *94*(3), 851–872. https://doi.org/10.1007/s11192-012-0775-9
Way, S. F., Morgan, A. C., Clauset, A., & Larremore, D. B. (2017). The misleading narrative of the canonical faculty productivity trajectory. *Proceedings of the National Academy of Sciences*, *114*(44), E9216–E9223. https://doi.org/10.1073/pnas.1702121114, PubMed: 29042510
Wu, L., Kittur, A., Youn, H., Milojević, S., Leahey, E., ... Ahn, Y.-Y. (2022). Metrics and mechanisms: Measuring the unmeasurable in the science of science. *Journal of Informetrics*, *16*(2), 101290. https://doi.org/10.1016/j.joi.2022.101290
Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378–382. https://doi.org/10.1038/s41586-019-0941-9, PubMed: 30760923
Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039. https://doi.org/10.1126/science.1136099, PubMed: 17431139
Xie, Y., & Shauman, K. A. (1998). Sex differences in research productivity: New evidence about an old puzzle. *American Sociological Review*, *63*, 847–870. https://doi.org/10.2307/2657505
Yoon, J., Yang, K.-C., Jung, W.-S., & Ahn, Y.-Y. (2021). Persona2vec: A flexible multirole representations learning framework for graphs. *PeerJ Computer Science*, *7*, e439. https://doi.org/10.7717/peerj-cs.439, PubMed: 33834106
Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., ... Havlin, S. (2019). Increasing trend of scientists to switch between topics. *Nature Communications*, *10*(1), 3439. https://doi.org/10.1038/s41467-019-11401-8, PubMed: 31366884