The Journal of Open Source Software

refsplitr: Author name disambiguation, author georeferencing, and mapping of coauthorship networks with Web of Science data

Auriel M.V. Fournier[1], **Matthew E. Boone**[1], **Forrest R. Stevens**[2], **and Emilio M. Bruna**[3, 4]

1 Porzana Solutions, Marquette Heights, IL, 61554, USA **2** Department of Geography & Geosciences, University of Louisville, Louisville, KY, 40292, USA **3** Center for Latin American Studies, University of Florida, Gainesville, FL, 32611-5530, USA **4** Department of Wildlife Ecology & Conservation, University of Florida, Gainesville, FL, 32611-4430, USA

DOI:

Software

- Review
- Repository
- Archive

Editor: Kyle Niemeyer

Reviewers:

- @kyleniemeyer

Submitted: 14 January 2020
Published: 22 January 2020

Summary

The Science of Science (SciSci) is an emerging, trans-disciplinary approach for using large and disparate data-sets to study the emergence, dissemination, and impact of scientific research (Fortunato et al., 2018). Bibliometric databases such as the Web of Science are rich sources of data for SciSci studies (Sugimoto & Larivière, 2018). In recent years the type and scope of questions addressed with data gathered from these databases has expanded tremendously (Fortunato et al., 2018). This is due in part to their expanding coverage and greater accessibility, but also because advances in computational power make it possible to analyze data-sets comprising millions of bibliographic records (e.g., Larivière, Ni, Gingras, Cronin, & Sugimoto, 2013; Smith, Weinberger, Bruna, & Allesina, 2014).

The rapidly increasing size of bibliometric data-sets available to researchers has exacerbated two major and persistent challenges in SciSci research. The first of these is **Author Name Disambiguation**. Correctly identifying the authors of a research product is fundamental to bibliometric research, as is the ability to correctly attribute to a given author all of their scholarly output. However, this seemingly straightforward task is often extremely complicated, even when using the nominally high-quality data extracted from bibliometric databases (reviewed in Smalheiser & Torvik, 2009). The most obvious case is when different authors have identical names, which can be quite common in some countries (Strotmann & Zhao, 2012). However, confusion might also arise as a result of journal conventions or individual preferences for abbreviating names. For instance, one might conclude "J. C. Smith", "Jennifer C. Smith", and "J. Smith" are different authors, when in fact they are the same person. In contrast, papers by "E. Martinez" could have been written by different authors with the same last name but whose first names start with the same letter (e.g., "Enrique", "Eduardo"). Failure to disambiguate author names can seriously undermine the conclusions of some SciSci studies, but manually verifying author identity quickly becomes impractical as the number of authors or papers in a dataset increases.

The second challenge to working with large bibliometric data-sets is correctly **parsing author addresses**. The structure of author affiliations is complex and idiosyncratic, and journals differ in the information they require authors to provide and the way in which they present it. Authors may also represent affiliations in different ways on different articles. For instance, the affiliations might be written in different ways in different journals (e.g., "Dept. of Biology", "Department of Biology", "Departamento de Biologia"). The same is true of institution names ("UC Davis", "University of California-Davis","University of California") or the country

in which they are based ("USA", "United States", "United States of America"). Researchers at academic institutions might include the one or more Centers, Institutes, Colleges, Departments, or Programs in their address, and researchers working for the same institution could be based at units in geographically disparate locations (e.g., University of Florida researchers could be based at the main campus in Gainesville or one of dozens of facilities across the state, including 12 Research and Education Centers, 5 field stations, and 67 County Extension Offices). Finally, affiliations are recorded in a single field of the reference bibliographic records, despite comprising very different types of information (e.g., city, postal code, institution). In concert, these factors can make it challenging to conduct analyses for which author affiliation or location is of particular interest.

Package `refsplitr` helps users of the R statistical computing environment (R Core Team, 2018) address these challenges. It imports and organizes the output from Web of Science searches, disambiguates the names of authors, suggests author names that would benefit from additional review to verify the disambiguation, parses author addresses, and georeferences author institutions. It then generates maps indicating the locations of authors and georeferenced coauthorship networks. Finally, the processed data-sets can be exported in tidy formats for analysis with user-written code or, after some additional formatting, packages such as `revtools` (Westgate, 2018) or `bibliometrix` (Aria & Cuccurullo, 2017).

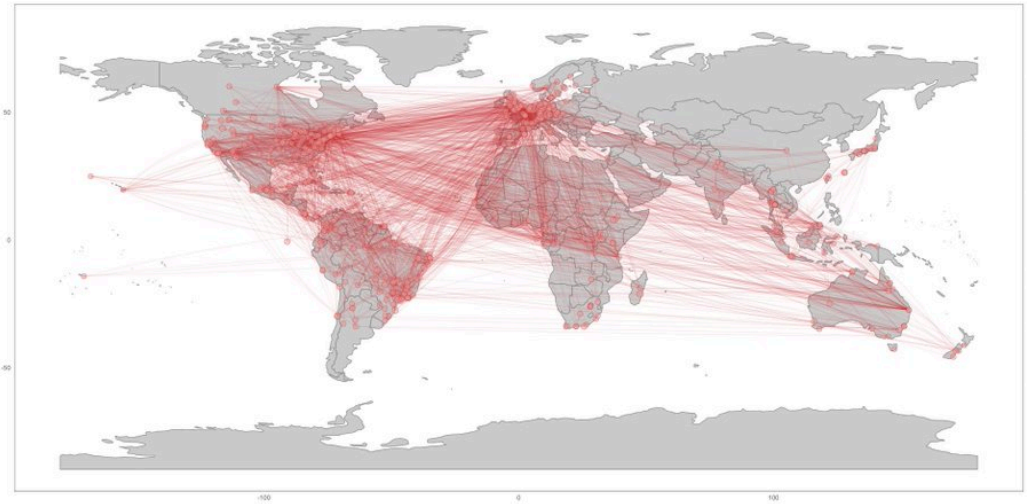

Figure 1: Map of georeferenced article coauthorships generated with `refsplitr`.

Acknowledgements

Support for the development of refsplitr was provided by grants to E. M. Bruna from the University of Florida Center for Latin American Studies and the University of Florida Informatics Institute. We are extremely grateful to Bianca Kramer and Najko Jahn for the feedback they provided during the review of `refsplitr` by rOpenSci.

References

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

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., et al. (2018). Science of science. *Science*, *359*(6379), eaao0185. doi:10.1126/science.aao0185

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature News*, *504*(7479), 211. doi:10.1038/504211a

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual review of information science and technology*, *43*(1), 1–43. doi:10.1002/aris.2009.1440430113

Smith, M. J., Weinberger, C., Bruna, E. M., & Allesina, S. (2014). The scientific impact of nations: Journal placement and citation performance. *PloS one*, *9*(10), e109195. doi:10.1371/journal.pone.0109195

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63*(9), 1820–1833. doi:10.1002/asi.22695

Sugimoto, C. R., & Larivière, V. (2018). *Measuring research: What everyone needs to know*. Oxford University Press.

Westgate, M. J. (2018). Revtools: Bibliographic data visualization for evidence synthesis in r. *bioRxiv*, 262881. doi:10.1101/262881