Higher-Order Bibliographic Services

based on bibliographic networks

Vladimir Batagelj, Jan Pisanski, Tomaž Pisanski
UP FAMNIT IAM, IMFM, UL FF

Information Society 2024
27th international multiconference
SiKDD 2024 – Slovenian KDD Conference

Ljubljana, 7. October 2024

Outline

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Current version of slides (October 6, 2024 at 21 : 03): slides PDF
https://github.com/bavla/OpenAlex

Collections of bibliographic networks

From special bibliographies (BibTEX, EndNote) and bibliographic databases (WoS, Scopus, DBLP, OpenAlex, etc.), it is possible to obtain data about works (papers, books, reports, etc.) on selected topics.

This data can be transformed into a collection of compatible two-mode networks on selected topics [1]: works × authors; works × keywords; works × countries, and other pairs of characteristics describing works.

Besides these networks, we can also get the partition of works by their publication years, the partition of works by journals or publishers, the vector of the number of pages, and, in some cases, the (one-mode works × works) citation network.

Higher-Order Bibliographic Services

The data collected in different bibliographic databases can be used to provide higher-order bibliographic and bibliometric services such as what to read (contact/visit)? – a list of relevant articles/books (authors, institutions) on selected topic; where to publish? – a list of journals suitable for the publication of an article, automatic suggestion of keywords; reviewer selection – a list of reviewers suitable for a submitted article; possible partners for research collaboration; a career application – a candidate's activity report draft; etc.) for different types of users (students, researchers, teachers, decision-makers, funding agencies, research institutions, database managers, etc.

For the development of higher-order bibliographic services, open bibliographic databases such as OpenAlex are particularly welcome, as the developed services can remain open.

OpenAlex

Our approach is based on OpenAlex [2, 3] but this information can be obtained from most bibliographic databases [4, 5]. OpenAlex is a fully open catalog (launched in January 2022) of the global research system [2]. It's named after the ancient Library of Alexandria and made by the nonprofit OurResearch.

OpenAlex indexes more than twice as many scholarly works as the leading proprietary products and the entirety of the knowledge graph and its source code are openly licensed and freely available through data snapshots, an easy-to-use API, and a nascent user interface.

OpenAlex is based on 7 types of units (entities): **W**(ork), **A**(uthor), **S**(ource), **I**(nstitution), **C**(oncept), **P**(ublisher), or **F**(under) (and some additional ones such as topics, keywords, countries, continents, languages, etc.). Each unit gets its OpenAlex ID – we assume that the identification problem is solved by the database.

OpenAlex API

The OpenAlex API is available at https://api.openalex.org. Its response is returned in JSON format.
The OpenAlex query can be composed of different components (search, filter, select). The query can be further controlled by some parameters. For example

```
wd <- GET("https://api.openalex.org/works",
  query = list(
    search="handball",
    filter="publication_year:2015",
    select="id,title",
    page="2", per_page="200"))
names(wd)
wc <- fromJSON(rawToChar(wd$content)); names(wc)
names(wc$meta); wc$meta$count; str(wc$results)
```

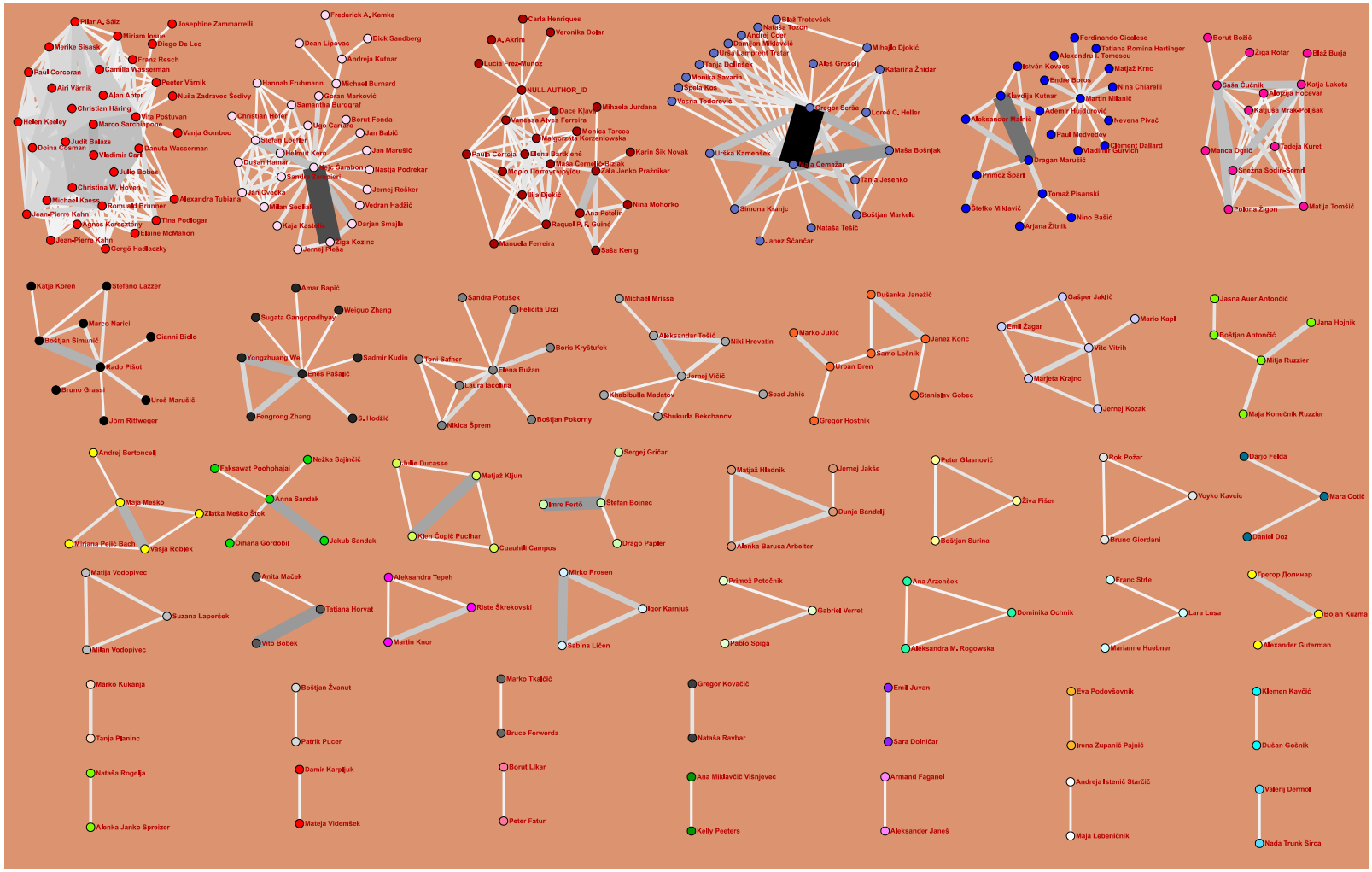
returns the second page (with up to 200 entries) on works on handball published in the year 2015. Only information about works ID and title is returned.
The OpenAlex API uses paging – the list data are provided by pages.

OpenAlex2Pajek

We developed an R package `OpenAlex2Pajek` to support the creation of bibliographic networks from OpenAlex [6]. We get a collection of bibliographic networks (citation network **Cite**, authorship network **WA**, sources network **WJ**, keywords network **WK**, countries network **WC**), some partitions and vectors (properties of nodes) (publication year, type of publication, language of publication, cited by count, countries distinct count, referenced works, and additionally two files containing names of works `xyzW.nam` and names of authors `xyzA.nam`. Most acquired networks are 2-mode – they link units of two different types; an ordinary or 1-mode network links units of the same type.

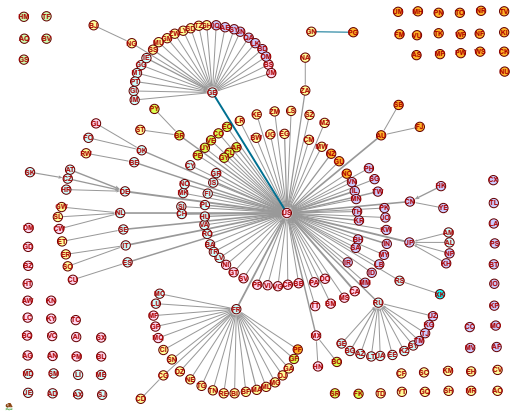
Currently, `OpenAlex2Pajek` contains three main functions `OpenAlex2PajekCite`, `OpenAlex2PajekAll`, and `coAuthorship`.

The largest co-author groups at level 10 at the University of Primorska until 2024

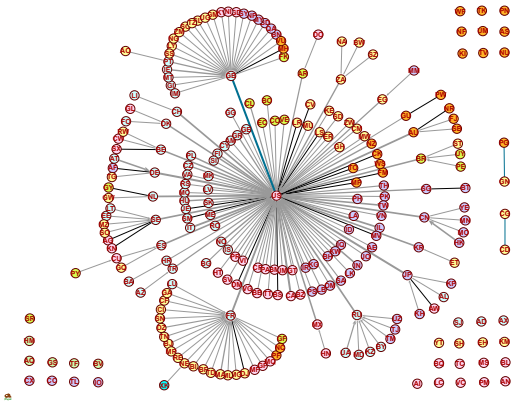


$$\mathbf{Co} = \mathbf{WA}^T * \mathbf{WA}$$

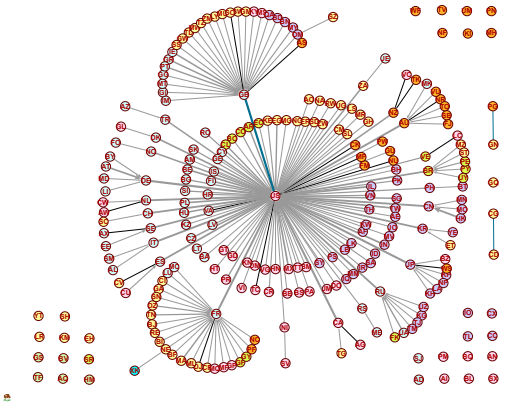
1-neighbors skeletons of world co-authorship for selected years



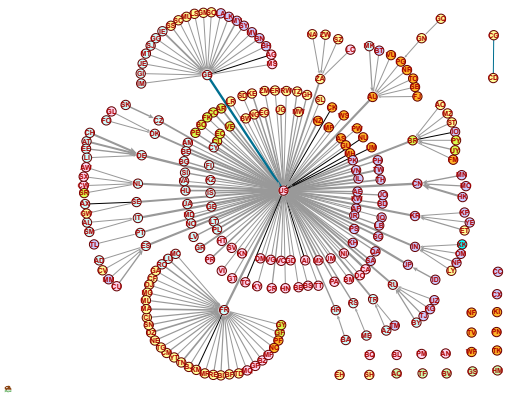
1990



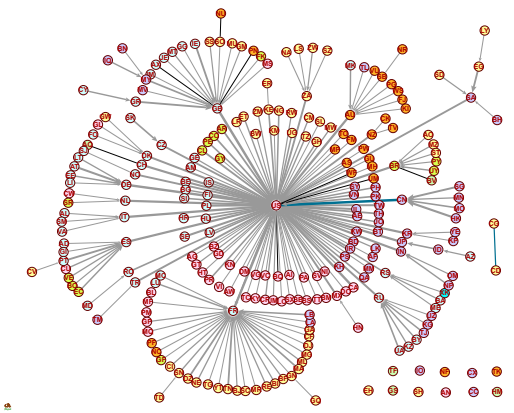
1995



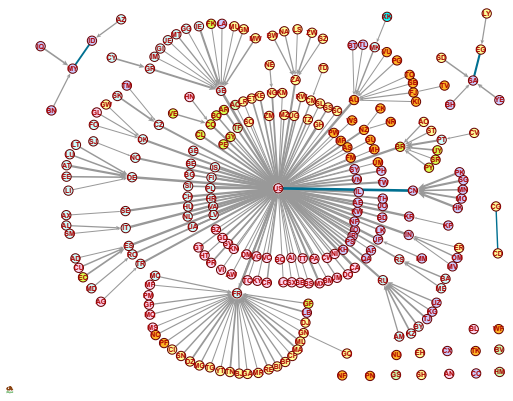
2000



2010



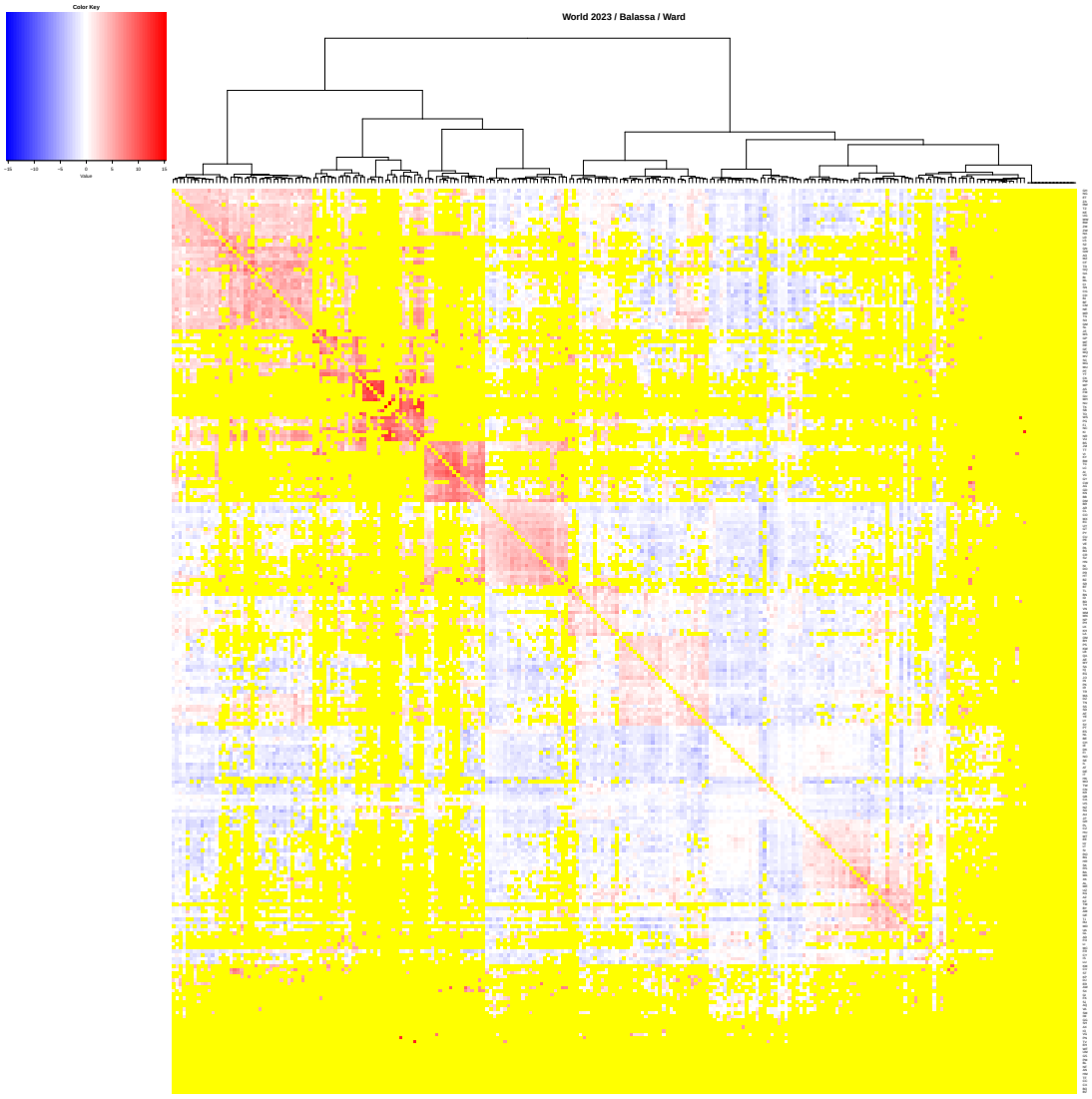
2015



2020

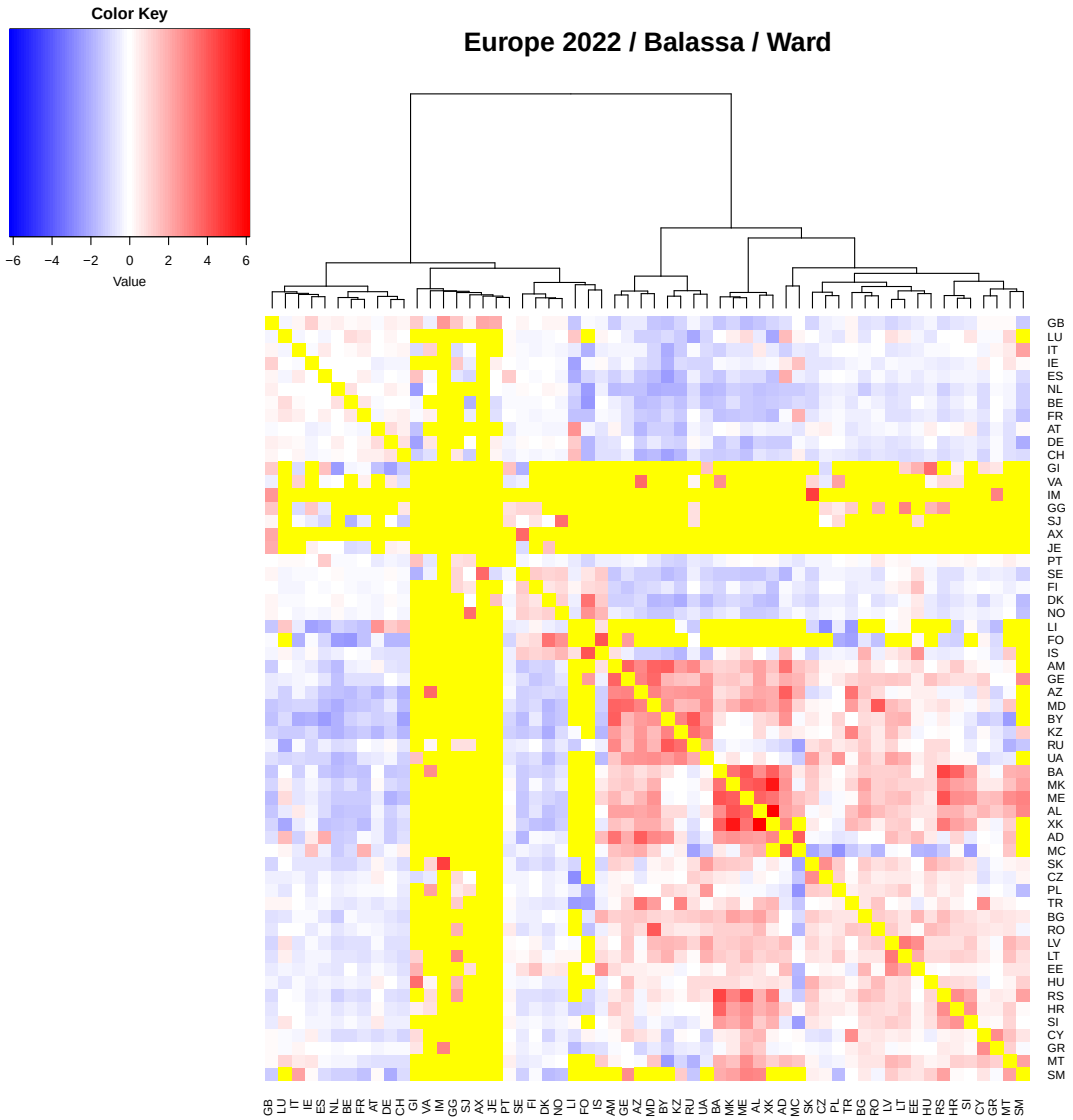
Balassa world co-authorship for the year 2023

yellow cell – no link, red/blue cell – above/below expectation



Balassa EU co-authorship for the year 2022

yellow cell – no link, red/blue cell – above/below expectation



Conclusions

OpenAlex is a rich source of bibliographic data that opens up new possibilities for bibliographic analyses and services.

Acknowledgments

The computational work reported in this paper was performed using a collection of R functions `OpenAlex`, R program `OpenAlex2Pajek`, and the program Pajek for analysis of large networks. The code and data are available at Github/Bavla/OpenAlex.

VB's work is partly supported by the Slovenian Research Agency ARIS (research program P1-0294, research program CogniCom (0013103) at the University of Primorska, and research projects J1-2481, J5-2557, and J5-4596), and prepared within the framework of the COST action CA21163 (HiTEc). JP's work is partly supported by ARIS (research program P5-0361 and research projects J1-2551 and J5-4596). TP's work is partly supported by ARIS (research program P1-0294 and research projects N1-0140, J1-2481, J5-4596).

References I

1. Batagelj, V. & Cerinšek, M. On bibliographic networks. *Scientometrics* **96,** 845–864. ISSN: 0138-9130 (2013).
2. Priem, J., Piwowar, H. & Orr, R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833* (2022).
3. Chawla, D. S. Massive open index of scholarly papers launches. *Nature* (2022).
4. Jiao, C., Li, K. & Fang, Z. How are exclusively data journals indexed in major scholarly databases? An examination of four databases. *Scientific Data* **10,** 737 (2023).
5. Delgado-Quirós, L. & Ortega, J. L. Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies* **5,** 31–49 (2024).
6. Batagelj, V. *OpenAlex2Pajek.* version 4, June 18. 2024. https://github.com/bavla/OpenAlex/tree/main/code.
7. Matveeva, N., Batagelj, V. & Ferligoj, A. Scientific collaboration of post-Soviet countries: the effects of different network normalizations. *Scientometrics* **128,** 4219–4242 (2023).