Higher-Order Bibliographic Services based on bibliographic networks

Vladimir Batagelj
IMFM
Ljubljana, Slovenia
IAM and FAMNIT, UP
Koper, Slovenia
vladimir.batagelj@fmf.uni-lj.si

Jan Pisanski
Faculty of Arts, UL
Ljubljana, Slovenia
jan.pisanski@ff.uni-lj.si

Tomaž Pisanski
FAMNIT, UP
Koper, Slovenia
IMFM
Ljubljana, Slovenia
tomaz.pisanski@upr.si

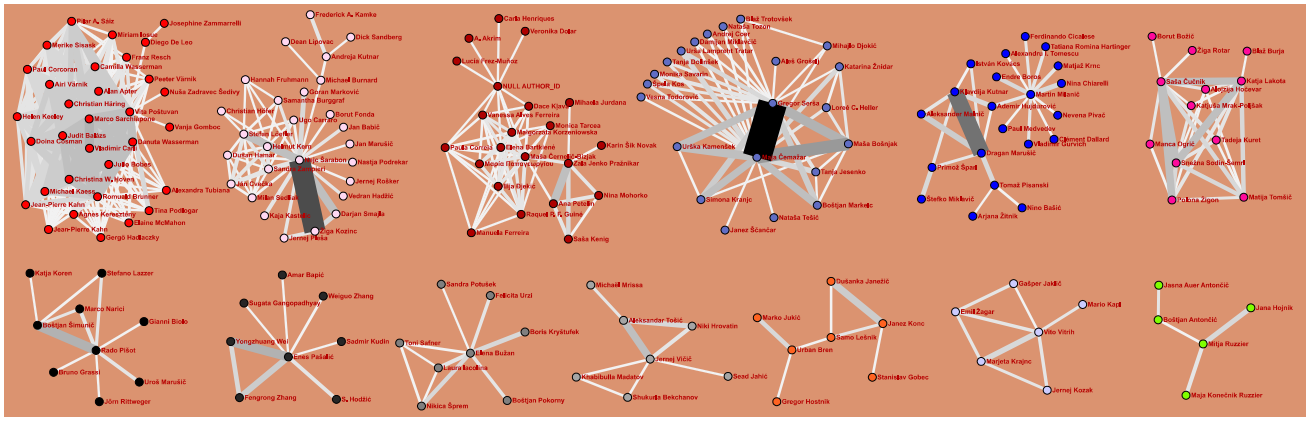

Figure 1: The largest co-author groups at level 10 at the University of Primorska until 2024.

Abstract

Bibliographic databases only provide basic services to users, but they could provide much richer information for specific user needs. The main reason for the delay in developing such higher-order bibliographic services is the limited access to data in proprietary databases. We expect the new open bibliographic databases like OpenAlex will encourage faster development of these services. We describe an approach based on a collection of bibliographic networks as a foundation to support the development of higher-order bibliographic services.

Keywords

bibliographic database, open access, network analysis, higher-order bibliographic service, prototype, OpenAlex

1 Introduction

From special bibliographies (BibTEX, EndNote) and bibliographic databases, it is possible to obtain data about works (papers, books, reports, etc.) on selected topics. A typical work description contains the following data: authors; title; publisher/journal; publication year and pages. In some sources, additional data are available including languages, classification of documents, keywords, authors' institution/country affiliation, lists of references, and the abstract. This data can be transformed into a collection of compatible two-mode networks on selected topics [5]: 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.

When constructing any of these networks, the first task is to specify the nodes and which relations are linking them. In short, the network boundary problem [16] has to be solved. This includes deciding whether a network is one-mode or two-mode and which node properties are important for the intended analyses. For specifying links, this amounts to answering a series of questions:

(1) Are the links directed?
(2) Are there different types of links (relations) to include?
(3) Can a pair of nodes be linked with multiple links?
(4) What are the weights on the links?
(5) Is the network static, or is it changing through time?

Another problem that often occurs when defining the set of nodes is the identification of nodes. The unit corresponding to a node can have different names (synonymy), or the same name can denote different units (homonymy or ambiguity). For example in the BibTEX bibliography from the Computational Geometry Database [14] the same author appears under 7 different names: R.S. Drysdale, Robert L. Drysdale, Robert L. Scot Drysdale, R.L. Drysdale, S. Drysdale, R. Drysdale, and R.L.S. Drysdale. Insider information is needed to decide that Otfried Schwarzkopf and Otfried Cheong are the same person. At the other extreme, there are at least 57 different mathematicians with the name Wang, and Li in the MathSciNet Database [20]. Its editors have tried hard, from 1985, to resolve the identification of the author's problem during the data-entry phase. The significant growth of contributions by Chinese scientists and their full name similarity in

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.12

Roman transcriptions adds additional complexity to the problem. In the future, the problem could be eliminated by implementing initiatives such as using ORCID or resolving the identification problem in bibliographic databases (Scopus, OpenAlex).

2 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. . To support this goal we have to use high-quality data often obtained by combining data from different databases.

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

3 OpenAlex

The basic type of unit in a bibliographic database is the work. A user searching the database gets a list of works satisfying the query. Usually, some operations with such lists (inspection, filtering, merging, intersection, statistics, etc.) are supported. Only basic services are provided to users.

Some web services also supporting some other types of units (authors, institutions, research fields, conferences, etc.) were developed such as Google Scholar [19], Scholar GPS [12], and DBLP – computer science bibliography [10].

Our approach is based on OpenAlex [18, 9] but this information can be obtained from most bibliographic databases [13, 11]. 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.

The simplest use of OpenAlex is through its web interface (service) https://openalex.org/ or using a direct URL request in the browser URL line. For example

- Author's name: search the OpenAlex web service
- Known author ID: URL https://openalex.org/A5001676164
- Work with DOI: URL https://api.openalex.org/works/https://doi.org/10.1007/s11192-012-0940-1
- Known work ID: URL https://openalex.org/W2083084326
- Name of the institution: search the Openalex Web service
- Known institution ID: URL https://openalex.org/institutions/I4210106342

This way, the OpenAlex web interface provides basic inspections of the selected unit. For example, by including a link with our OpenAlex author ID on our web page we get a report on our publications. Similarly, we get the report on the publication activity of the selected institution.

3.1 API

An application programming interface (API) is a way for two or more computer programs or components to communicate with each other. It is a type of software interface, offering a service to other pieces of software [21]. In our case, API enables us to use the database data from our programs. An R package supporting the use of OpenAlex is `openalexR` [1].

The OpenAlex API is available at https://api.openalex.org. Its response is returned in JSON format. Here is an R code using the OpenAlex API for the IMFM institution search

```
setwd(wdir <- "C:/work/OpenAlex/API")
library(httr); library(jsonlite)
res <- GET("https://api.openalex.org/institutions",
    query = list(search="imfm"))
str(res)
cont <- fromJSON(rawToChar(res$content))
names(cont); str(cont)
```

The response data are available in the variable `cont`. Similarly, the API can be used also from other programming languages.

The OpenAlex query can be composed of different components. Using **search** we can search for a given search text across titles, abstracts, and full-text. Using a **filter** we can limit our search to units satisfying given conditions. Using **select** we can select data fields that will appear in results. 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. The **basic paging** (up to 10 000 units) is based on two parameters `page` and `per_page`). The **cursor paging** is a bit more complicated than basic paging, but it allows us to access as many records as we like.

4 A collection of bibliographic networks

We developed an R package `OpenAlex2Pajek` to support the creation of bibliographic networks from OpenAlex [4]. 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`.

We split the process of creating the collection of bibliographic networks into two parts:

- determining the set W of relevant works using the **saturation approach** [7, page 506],
- creation of the network collection for the works from W.

The set W is determined iteratively using the function `OpenAlex-2PajekCite` and the collection is finally created using the function `OpenAlex2PajekAll`.

The function `coAuthorship` creates a weighted temporal network describing the co-authorship between world countries in selected time intervals. The weight of an edge is the number of works co-authored by authors from the linked countries.

In an analysis of weighted networks, the 1-neighbor skeleton is often used to get an overall insight into the network's basic structure. In the 1-neighbor skeleton, only its strongest link is kept for each node. The resulting directed network is forest-like. Non-trivial connected components in 1-neighbor skeletons are (usually) directed trees with a pair of nodes linked in both directions with the largest weight in the tree – these two arcs are usually replaced by an edge (undirected link). In Figure 2 the 1-neighbor skeletons for years 1990, 1995, 2000, 2010, 2015, and 2020 are presented. We see that the number of isolated nodes (countries not collaborating with other countries) is decreasing. In all analyzed years the US has a leading (hub) position. In the years 1990, 1995, 2000, and 2010 the edge in the main component links US and GB but in the years 2015 and 2020 GB is replaced by CN. In 1990, stronger secondary hubs were GB, FR, RU, JP, and DE. In the following years, some other countries SE, ES, AU, CN, BR, ZA, and IN (BRICS) became secondary hubs attracting previously non collaborating countries or geographically or linguistically close countries.

Most of the ingredients of basic reports are counters, sorted lists, (weighted) degrees and their distributions obtained from an adequate network. Sometimes also the time is considered producing time series.

An important property of a collection of bibliographic networks is that some of them are compatible – they share a common set (most often the set of works W). This allows us to use network multiplication (defined by the product of network matrices) to compute the corresponding derived network connecting the remaining two sets [5]. For example, in the derived network $\mathbf{AK} = \mathbf{WA}^T \cdot \mathbf{WK}$ its entry $AK[a,k]$ tells us in how many works the author a used the keyword k. Similarly, in the derived network $\mathbf{ACiK} = \mathbf{WA}^T \cdot \mathbf{Cite} \cdot \mathbf{WK}$ its entry $ACiK[a,k]$ tells us how many times the author a cited works described by the keyword k.

A 2-mode network is always compatible with its transpose (on both sets). The corresponding derived networks are called projections – the row projection $\mathrm{row}(\mathbf{WA}) = \mathbf{WA} \cdot \mathbf{WA}^T$ and the column projection $\mathrm{col}(\mathbf{WA}) = \mathbf{WA}^T \cdot \mathbf{WA}$. Both projections are ordinary weighted 1-mode networks that can be analyzed using standard network analysis methods.

For the authorship network **WA** its column projection $\mathbf{Co} = \mathbf{WA}^T \cdot \mathbf{WA}$ is the co-authorship network. Its entry $Co[a,b]$ counts the number of works that authors a and b co-authored. It turns out that a work with k co-authors contributes k^2 links to the co-authorship network – works with a large number of co-authors are overrepresented in it. To treat all authors equally the fractional approach is used [3]. In Figure 1 the largest co-authorship groups at level 10 at the University of Primorska are presented – connected components of the link cut at level 10 in the network **Co**. Each pair of linked authors co-authored at least 10 works

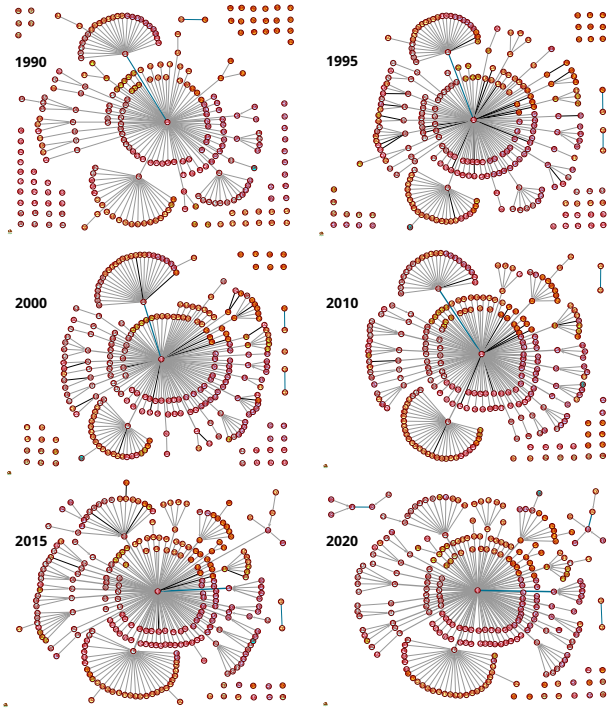

Figure 2: 1-neighbors skeletons of world co-authorship for selected years.

in the bibliography of works with at least one co-author from University of Primorska.

In bibliometric analysis, the citation network **Cite** has a very important role. It collects "votes" about the relevance of previous works for a given work. It is often used for solving the network boundary problem, and also for identifying the most relevant works in the collected bibliography [2, 6]. The derived network $\mathbf{ACiA} = \mathbf{WA}^T \cdot \mathbf{Cite} \cdot \mathbf{WA}$ describes the citations between authors – its entry $ACiA[a,b]$ counts the number of times author a cited author b. The co-citation network is defined as the column projection of the citation network $\mathbf{coCi} = \mathrm{col}(\mathbf{Ci}) = \mathbf{Ci}^T \cdot \mathbf{Ci}$ and the bibliographic coupling network is defined as the row projection of the citation network $\mathbf{biCo} = \mathrm{row}(\mathbf{Ci}) = \mathbf{Ci} \cdot \mathbf{Ci}^T$.

The idea of derived networks can be extended to temporal bibliographic networks [8]. Using derived networks we enlarge the source for different statistics. Additional insight can be gained by analyzing the structure of networks and identifying important subnetworks in them [6].

In the following, we present an overview of typical report ingredients [7, 15]. Because of limited available space, we decided to put examples on Github/bavla.

5 Report ingredients

5.1 Statistics

Because the analyzed networks are often large a complete presentation is not an option. To describe them we use different statistical descriptors.

- sizes of sets (number of nodes, number of links); structural network properties (number of components, size of the largest component, etc.)
- top units – ordered lists of units with the largest values of selected property (degre, weighted degree, link weight,

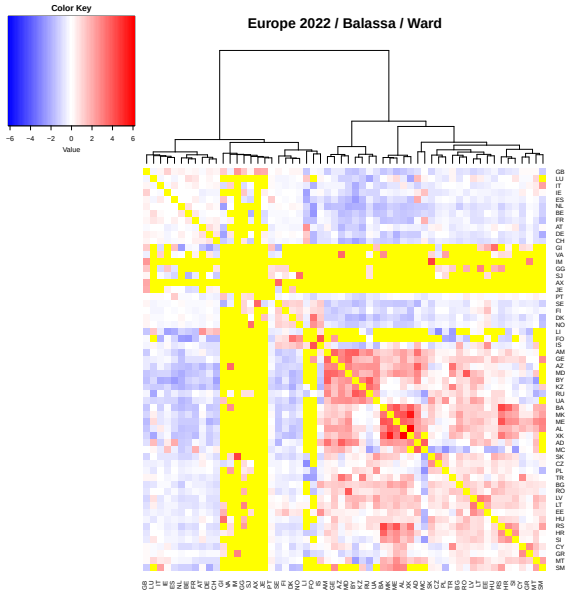

Figure 3: Balassa EU co-authorship for the year 2022.

- distribution of selected property
- time series describing temporal changes of selected properties
- scatter plots showing a possible relationship between two selected properties

Often bibliometric properties of units follow laws such as Zipf (or power) law, Bradford law, Lotka law, lognormal distribution, Hirsch index, etc.

5.2 Network analysis

Derived networks are weighted. To get readable results of reasonable size we usually search for important subnetworks, often a kind of skeleton – from a given network less important elements are removed. There are different types of skeletons (spanning forest, k closest neighbors, cuts, cores, islands, etc. [6]).

A traditional graph-based visualization is used if the obtained result network is not dense. For denser networks, the matrix display is much more readable. In a matrix display, the permutation of nodes (usually obtained by clustering) can create patterns that reveal the network's internal structure.

Figure 3 presents a matrix display of Balassa co-authorship indices between European countries in 2022 (yellow cell – no link, red/blue cell – above/below expectation) [17].

5.3 Special algorithms

Some properties can require special computational procedures and direct access to the bibliographic data. In such cases, open access to the bibliographic database is of crucial importance.

5.4 Reports

The results of analyses can be combined and presented to users in different forms:

- Booklet report (in PDF).
- (Service generated) web pages.
- Dashboards.
- Dataset (JSON, CSV, etc.).

6 Conclusions

We have presented an approach to support higher-order bibliographic services based on networks. Open access to high-quality bibliographic data is crucial for the faster development of such services. The new bibliographic database OpenAlex seems to be a step in the right direction. It needs the support of science policy and also of individual scientists (checking the correctness of their data).

Acknowledgements

The computational work reported in this paper was performed using a collection of R functions OpenAlex2Pajek and the program Pajek for analysis of large networks. Code, data, and figures are available on 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

[1] Massimo Aria, Trang Le, Corrado Cuccurullo, Alessandra Belfiore, and June Choe. 2024. openalexR: an R-tool for collecting bibliometric data from OpenAlex. *The R Journal*, 15, 4, 167–180.

[2] Vladimir Batagelj. 2003. Efficient algorithms for citation network analysis. *arXiv preprint cs/0309023*.

[3] Vladimir Batagelj. 2020. On fractional approach to analysis of linked networks. *Scientometrics*, 123, 2, 621–633. DOI: 10.1007/s11192-020-03383-y.

[4] Vladimir Batagelj. 2024. OpenAlex2Pajek. version 4, June 18. (2024). https://github.com/bavla/OpenAlex/tree/main/code.

[5] Vladimir Batagelj and Monika Cerinšek. 2013. On bibliographic networks. *Scientometrics*, 96, 3, 845–864. DOI: 10.1007/s11192-012-0940-1.

[6] Vladimir Batagelj, Patrick Doreian, Anuška Ferligoj, and Nataša Kejžar. 2014. *Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution. Wiley Series in Computational and Quantitative Social Science*. Wiley, Chichester. ISBN: 978-1-118-91537-0; 978-0-470-71452-2. DOI: 10.1002/9781118915370.

[7] Vladimir Batagelj, Anuška Ferligoj, and Flaminio Squazzoni. 2017. The emergence of a field: a network analysis of research on peer review. *Scientometrics*, 113, 1, 503–532. DOI: 10.1007/s11192-017-2522-8.

[8] Vladimir Batagelj and Daria Maltseva. 2020. Temporal bibliographic networks. *J. Informetr.*, 14, 1, Article No. 101006. DOI: {10.1016/j.joi.2020.101006}.

[9] Dalmeet Singh Chawla. 2022. Massive open index of scholarly papers launches. *Nature*.

[10] DBLP – computer science bibliography. 2024. (2024). https://dblp.org/.

[11] Lorena Delgado-Quirós and José Luis Ortega. 2024. Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, 5, 1, 31–49.

[12] Scholar GPS. 2024. (2024). https://scholargps.com/.

[13] Chenyue Jiao, Kai Li, and Zhichao Fang. 2023. How are exclusively data journals indexed in major scholarly databases? an examination of four databases. *Scientific Data*, 10, 1, 737.

[14] Bill Jones. 2002. Computational geometry database. (2002). ftp://ftp.cs.usask.ca/pub/geometry/.

[15] Daria Maltseva and Vladimir Batagelj. 2019. Social network analysis as a field of invasions: Bibliographic approach to study SNA development. *Scientometrics*, 121, 2, 1085–1128. DOI: 10.1007/s11192-019-03193-x.

[16] Peter V. Marsden. 1990. Network data and measurement. *Annu. Rev. Sociol.*, 16, 435–463. DOI: 10.1146/annurev.so.16.080190.002251.

[17] Nataliya Matveeva, Vladimir Batagelj, and Anuška Ferligoj. 2023. Scientific collaboration of post-soviet countries: the effects of different network normalizations. *Scientometrics*, 128, 8, 4219–4242.

[18] Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

[19] Google Scholar. 2024. (2024). https://scholar.google.com/.

[20] Bert TePaske-King and Norman Richert. 2001. The identification of authors in the mathematical reviews database. *Issues Sci. Technol. Librariansh.*, 31. DOI: 10.5062/f4kh0k9m.

[21] Wikipedia. 2024. API. August 22. (2024). https://en.wikipedia.org/wiki/API.