Scientometrics (2025) 130:3013–3028
https://doi.org/10.1007/s11192-025-05298-y

Finding Doppelgängers in Scopus: how to build scientists control groups using sosia

Michael E. Rose[1] · **Stefano H. Baruffaldi**[1,2]

Received: 7 December 2024 / Accepted: 27 March 2025 / Published online: 16 April 2025

Abstract
The construction of control groups of scientists is often a daunting effort. This paper presents `sosia`, an open-source Python-based software designed to efficiently query the Scopus database via RESTful API. `sosia` searches for researchers with publication profiles similar to a given researcher up to a given year based on all main standard bibliometric indicators. The user can choose flexibly a set of parameters to restrict the search to more or less narrow boundaries upfront and obtain additional similarity indicators to select a subset of authors after the search. Advanced settings also allow narrowing the search to a list of affiliations and to minimize the possible errors arising from ambiguous author profiles. One basic search can be set up in a few command lines and the average time of computation goes between 60 and 300 minutes. We discuss the functioning, characteristics, limitations and possible extension of the software.

Keywords Statistical doftware · Control group · Diff-in-diff · Scopus

JEL Classification C00 · A14

Introduction

Econometric analysis in Economics of Science and Innovation often requires control groups. These control groups need to have similar observable characteristics to a sample of researchers of interest. There are specific methodologies and tools to assist

We are grateful to the editor Wolgang Glänzel, two anonymous referees, to Alex Oettl as well as participants of the 4th Summer School on Data and Algorithms for ST&I Studies 2019 at the EPO Vienna Office and at the "The Economics and Organisation of Science conference at the TU Munich Heilbronn campus in 2022 for helpful feedback and comments. We are thankful to Alex Oettl for sharing the data used in Oettl (2012). Finally, we thank colleagues at the Max Planck Institute for Innovation and Competition for extensive testing of the software. Carolin Formella, Kubra Kocak and Iliana Radeva provided able research assistance.

✉ Michael E. Rose
michael.rose@ip.mpg.de

[1] Max Planck Institute for Innovation and Competition, Munich, Germany

[2] Politecnico di Milano, Milan, Italy

Econometricians in the matching exercise once a broad population of scientists is identified (Iacus et al., 2012; Blackwell, 2009; Iacus et al., 2009; Berta et al., 2017). However, the identification of such a population often constitutes a daunting data effort which may turn impossible for samples of scientists spanning multiple fields, institutions, or countries. The Python package `sosia`—Italian for Doppelgänger—intends to simplify and automate the search for comparable researchers in the Scopus database.

The typical use case is the study of the effect of an exogenous shock affecting a sample of researchers. While the shock can be conceived as exogenous, the Econometrician does not often have a natural control group, and the sample of treated individuals does not present balanced characteristics to the general population of researchers. The Econometrician then needs to construct a control group of researchers with similar characteristics. The standard characteristics considered are demographic and professional information, such as the age and the research field, and publication-based indicators of scientific productivity, up to the year of interest. `sosia` automates the process of identification of potential controls by efficiently querying Scopus via their RESTful API[1] and then selectively searching for researchers with similar observable characteristics to a given researcher in a given year.

The online search in Scopus offers three main advantages. First, Scopus provides author identifiers with a good degree of reliability (Baas, 2020; Moed et al., 2013). Second, the online database allows taking into account older publications as well and, therefore, to better consider the start of one's career. Third, Scopus offers an API through which we can search programmatically for scientific articles and authors. For the latter task, `sosia` relies on the Python package *pybliometrics*[2] that needs to be running and set up as well (Rose & Kitchin, 2019). However, the API comes with some constraints that make large-scale searches non-trivial.

In a nutshell, `sosia` optimizes the search for similar researchers, with a minimum number of queries. The user provides at least the Scopus author ID of the treated researcher, and the year of treatment. `sosia` then compiles a first list of candidates who publish regularly in topically similar sources (i.e., journals, conference proceedings, books, etc.) since the focal author has started to publish. Co-authors are excluded. Then, `sosia` consecutively filters these candidates based on the main discipline they are active in, the first year of publication, the number of publications, the number of coauthors, and the number of citing papers. Users can switch off these filters and provide margins for all but the first one. Finally, users can search for additional characteristics of the matches to manually assess the match quality. Among these are topical similarity scores and country of publication.

In the remainder of the paper, we discuss the backdrop against which we develop our software. It follows extensive documentation with examples of code. After that, we present the results of a search performed on the sample used in Oettl (2012) to provide an indication of the timing and number of results for a typical use case. Finally, we discuss limitations and possible future extensions.

[1] An Application Program Interface (API) through which users interact via HTTP requests.

[2] See https://pybliometrics.readthedocs.io/en/stable/.

Background

Econometric considerations

The identification of adequate controls is a common challenge in applied econometric studies. In experimental settings, the Econometrician can assign a treatment randomly to part of a target population and use the remaining part as control, which would most likely show comparable characteristics. This is often not possible in the case of natural experiments because a defined control may not exist. In these cases, the Econometrician has to construct the control group.

As a canonical example, existing literature on externalities in collaboration networks often exploit the unexpected death of researchers as an exogenous shock (Azoulay et al., 2010; Oettl, 2012; Mohnen, 2022). The premature death of some individuals can be conceived as a rare and unexpected event that is therefore unlikely to be in itself a function of any characteristic of the individuals in the network. However, the effect of this event cannot be properly estimated within the sample of affected researchers (Azoulay et al., 2010). At the same time, a control cannot be randomly selected in the population of unaffected researchers. This is because the event is not strictly randomized across the entire population of researchers at large and the treated population will likely show different characteristics from the general population. Just, for instance, the death of more prominent scientists with higher potential for future contributions may be more prominently covered in newspapers and more likely observed, leading to a higher level *and* differential trend *after* the event (had the event not occurred).

Finding a suitable control group involves systematically searching for comparable individuals from a broad population with pre-computed matching values. Most scholars have relied on upfront construction of large researcher databases, thereby expanding their samples beyond the treated group. Comparing the focal sample with a wide candidate pool increases the chance of identifying at least one researcher with similar characteristics. Data are mostly assembled manually to capture basic demographic information and reliable indicators of scientific productivity, making control group identification extremely time-consuming in terms of data construction and cleaning. Moreover, the candidate pool for a given treated scientist is expected to be homogeneous across certain dimensions (e.g., research field or location). Studies focusing on one sector, country, or period can concentrate on a single search population and manually compile the necessary information, whereas a manual approach becomes infeasible for large or multidimensional samples.

An alternative is resorting to the use of bibliographic databases that can be obtained and processed offline (e.g., Scopus, Web of Science, Microsoft Academic Graph, PubMed). This alternative approach has two limitations. First, the lack of name disambiguation would likely induce large noise in the data. For small samples, also large-scale automatic disambiguation efforts may leave unacceptable levels of noise. Second, often the data provided are truncated and do not allow observing publications before a certain year. For instance, to our knowledge, most Web of Science (WOS) versions available start in 1970. This is problematic because the year in which a researcher starts publishing is an important variable to consider that would be often unobserved in these data.

Once a database with pre-computed matching values is available, the task becomes reducing variation between treated and control researchers. Various methodologies have been developed for this purpose, including Propensity Score Matching (PSM) and Coarsened Exact Matching (CEM) (Berta et al., 2017; Iacus et al., 2012; Blackwell, 2009; Iacus

et al., 2009). CEM is often the preferred method for identifying a control group with comparable observable characteristics at the moment of the event. In CEM, researchers are classified into strata based on their characteristics, and only those within the same strata are considered valid controls. In contrast, PSM retains all observations and parametrically estimates the probability of treatment, which is then used either to weight observations or to select "neighbors."

Both methodologies assume that, in the absence of the event, the treated individual would have exhibited a similar level and, more importantly, a similar trend in the outcome variable as the selected control (or conditional on the estimated probability of treatment). The characteristics usually considered to perform the match are the age of the researcher (sometimes substituted by the year of the Ph.D. or the first year of publication as a proxy), the levels of productivity in terms of number and quality of publications, the number of coauthors, etc. The target population is normally represented by researchers within a specific research field and country. However, in general, the target population, the matching variables, and, in the case of CEM, the bandwidths to create the observation strata are choices in the hands of the researcher.

CEM is often preferred because it makes these choices transparent, with the maximum level of imbalance in observable characteristics determined ex-ante rather than ex-post (Iacus et al., 2012). Ultimately, identification relies on specific assumptions. First, the treatment must be truly exogenous. Second, given that the treatment is a relatively rare event, it is assumed that a population closely resembling the treated sample can be identified based on a finite set of variables highly correlated with expected future outcomes, serving as a valid counterfactual for the treated observations had the treatment not occurred.

Related literature and software

A seminal paper applying CEM is Azoulay et al. (2010). The authors rely on a database comprising information on more than 10,000 elite life scientists. They apply CEM to find a control group for 5,267 collaborators of 112 star-scientists who died, matching on scientific publications, career age, and collaboration patterns. They find a match for 96% of the sample. Another example is Oettl (2012), who performs an empirical analysis based on the death of 161 immunologists. The author looks for control authors for the deceased immunologists based on the year of first publication, the number of coauthors at the time of death, the number of publications and their quality, the number of citations, and the number of acknowledgments received. He finds matches for 149 authors. We use Oettl (2012) as a test run for `sosia` in Sect. "Testing on samples of previous studies".

Numerous other studies have also used comparable methodologies, for instance: to study the effect of migration flows on scientists' productivity comparing scientists differently exposed to migrations, either due to their geographical location (Ganguli, 2015) or field of specialization (Teodoridis et al., 2019); the effect of negative shocks to reputation and status (Azoulay et al., 2014); the effect of student mobility, comparing students with higher or lower probability of leaving the US, based on visa regulation (Kahn & MacGarvie, 2016); the role of various aspects in collaboration networks on someone's productivity (AlShebli et al., 2018; Yadav et al., 2023); exposure to mentorship or mentors with different experience (Klingbeil, 2019; Muschallik & Pull, 2016); the effect of funding and research programs on productivity (Graddy-Reed et al., 2018; Colatat, 2015); the impact of broader policy changes on researchers' productivity (Hird & Pfotenhauer, 2017); the effect

of the access to physical resources or information on knowledge production (Baruffaldi & Gaessler, 2021; Hussinger & Palladini, 2024). The list does not mean to be exhaustive.[3]

To date, there is no other open-source software project dedicated to the specific challenge of identifying and downloading information on control groups in publications data. Related software is "Pub-Harvester", which enables users to automatically identify co-authors of a sample of scientists and build their publications panels from the database PubMed. The software is described in detail in Azoulay et al. (2010) and Azoulay et al. (2019).[4] If a population of researchers is already available, a number of software packages, in particular for Stata and SAS, exist that can support the implementation of CEM methodologies, enabling users to divide the population in strata along with any variable available (Berta et al., 2017; Iacus et al., 2012; Blackwell, 2009; Iacus et al., 2009).

Scopus RESTful API

The Dutch publisher Elsevier launched Scopus as a commercial product in 2004. Scopus is mostly subscription-based and therefore not accessible for free. Scopus has a broad coverage of scientific fields and extensive bibliographic information (Mongeon & Paul-Hus, 2016). Scopus uses many data entities with stable identifiers, namely documents, sources (e.g. scholarly journals or books), authors, and affiliations. These entities are interconnected, so that, for a given author, all documents with citation counts, the corresponding sources, and past and current affiliations are available.

`sosia` relies heavily on Scopus' author identifiers that have a high degree of reliability (Baas, 2020; Moed et al., 2013). Baas (2020) provides details on the disambiguation process adopted and report an assessment of its quality based on the number of publications correctly allocated to an author identifier. Scopus assigns publications to the same author based on name, and additionally on field and affiliation. The algorithm prefers "split profiles" over "merge profiles" (Moed et al., 2013): if unsure, it rather casts too many profiles for the same researcher than lumping publications of many different researchers into one profile. This leads to higher precision, relative to recall. Accordingly, Baas (2020) estimates precision equal to 98.1% and recall equal to 94.4%. This has important implications for the functioning of `sosia` (see section 3.4). Users of the Scopus database can also request changes to any author profiles (not only their own).

Scopus users can query the entire database programmatically via a RESTful API (Moed et al., 2013). `sosia` accesses the API through *pybliometrics* (Rose & Kitchin, 2019) to perform systematic queries and cache downloaded results. This allows access to the database in a variety of forms, searching for articles, authors, and affiliations, based on their characteristics. In relation to our objective, the possibility of querying data on bibliographic records and authors is the main feature of interest.

However, the API has a number of constraints:

1. The maximum number of results that can be extracted is limited and query time increases considerably the larger the number of results and size of information required.
2. API keys face a weekly quota.
3. There is a maximum length of a query URL.

[3] See Liu (2023) for a recent review of empirical methods in the science of science.

[4] See http://stellman-greene.com/SCGen/.

4. It is not possible to search authors based on the years in which they publish (only searching first for specific publications).
5. It is not possible to directly search authors based on their field of research.

`sosia` furthermore relies on Scopus' assignment of fields to sources because it infers the fields from the sources a researcher publishes in. Scopus uses the All Science Journal Classification (ASJC) system, which has a total of 331 fields grouped into 25 broader fields. Wang and Waltman (2016) find that there is a considerable degree of over-assignment of journals (which are a subset of all sources), but there is a low degree of under-assignment of ASJC fields to sources. When `sosia` infers the field(s) of a researcher via the sources she publishes in, it tends to assign the researcher to too many fields. This is preferable over the alternative, where true matches would be overlooked. As a consequence, however, `sosia` might look at too many researchers but initially and in the end. `sosia` provides relative measures of topical similarity based on text and references which helps users to filter out matches from alien fields.

Documentation

Installation and requirements

`sosia` is available for installation from PyPI or directly from the GitHub repository. It installs the necessary dependencies, notably `pandas` (McKinney, 2010) and `pybliometrics` (Rose & Kitchin, 2019). `pybliometrics` provides access to Scopus via its RESTful API. Users need to obtain credentials from Scopus and configure *pybliometrics* properly before using `sosia`. Users also need to have a connection to the Scopus API, typically via their institutions.

Order of operation

The objective of `sosia` is to identify all researchers in the Scopus database that appear similar to a given author a in a given year y, based on bibliographic information. We designed `sosia` to perform the most extensive and efficient search possible, given the constraints (discussed in Sect. "Scopus RESTful API") and the powers the Scopus API offers. There are many options and parameters at the user's disposal to steer the search speed, the probability of finding a match, and the quality of the matches.

1. In the initial set-up, `sosia` downloads, if not present, information on all sources available in Scopus, with information on their internal identifier, the type of source, and the fields (ASJC-4) they belong to. It also establishes a local SQLite database to speed up the search if it does not exist already.
2. `sosia` downloads all publications of a to determine the characteristics of a as of y, including its discipline (ASJC-2) and research fields (ASJC-4).
3. *Search set*: The set of candidates

(a) `sosia` obtains the list of sources[5] that are associated to these fields. Users can decide whether to include or exclude sources that are also associated to fields not among a's fields.
(b) `sosia` then downloads publication lists for all these similar sources for all relevant years.
(c) `sosia` considers only authors that publish in these source around the time a started to publish, as well as in regular, user-defined intervals (or more often) until the comparison year.
(d) `sosia` drops co-authors of a.

4. *Filtering candidates:* In a consecutive process, `sosia` may filter authors based on the following criteria:
   - Discipline (ASJC-2 code) must match that of a (assessed by the sources in which the Original publishes).
   - Publication count passes lower threshold (number of a's publications minus margin).
   - First publication within the margins around y_0.
   - Number of coauthors is within coauthor count margin of a as of y.
   - Number of citing papers within citing papers margins of a as of y.[6]
5. *Providing information*: For the list of matches, optionally, `sosia` obtains data to provide additional indicators of similarity, including affiliation information, the languages they publish in, and an indicator of topical relatedness.

Example

While the full documentation is hosted on ReadTheDocs.io under https://sosia.readthedocs.io/en/stable/, we give a brief example of a complete search process in Source Code 1.

Source Code 1: Example search.

```
import sosia

db_path = "cache.sqlite"
log_path = "queries.log"
stefano = sosia.Original(55208373700, 2019, db_path=db_path, log_path=log_path)
stefano.define_search_sources(mode="narrow")
stefano.identify_candidates_from_sources(first_year_margin=1, frequency=2)
stefano.filter_candidates(same_discipline=True, first_year_margin=1,
pub_margin=0.2, coauth_margin=0.15,
cits_margin=200)
stefano.inform_matches()
```

[5] All types of bibliographic publications available in Scopus: journals, conference proceedings, books, etc.
[6] This differs from the total number of citations because one paper citing multiple papers of a would be counted only once. This count is much easier and less time consuming to compute than the total number of citations. At the same time, it is not a priori obvious whether double-counting citing papers is preferable.

In line 3, users initiate the main class, `sosia.Original()`. Only the Scopus Author ID and the match year are mandatory. Instead of a single Scopus Author ID, users may pass a list of Scopus Author IDs. This feature deals with split authors" (Moed Aisati, and Plume, 2013, p. 941). In case there are merged authors", i.e., profiles to which publications of multiple real authors belong, users pass a list of document identifiers using parameter `eids`, while passing a single Scopus Author ID. All characteristics will be inferred from research-type publications published before 2019. Put differently, `sosia` will base matches on characteristics of author $a = 55208373700$ as of $y = 1$ January 2019. In case the database does not exist, `sosia` will create it with the necessary tables. In case the sources-field mapping has not been downloaded, `sosia` will do so. We provide this mapping via a separate repository,[7] which is based on Scopus' own official "Source title list"[8] We will update this information in accordance with Scopus' update cycle in the spring and in fall.[9]

In line 4, `sosia` defines the set of search sources. A search source is a source (journal, conference proceeding, etc.) that is connected to the fields of research (ASJC-4) of a, and the same types of sources. If a publishes exclusively in journals, `sosia` will consider only journals. Using `mode="narrow"`, `sosia` will exclude sources with fields that are not among the fields of a; using `mode="wide"` will include sources with those fields. However, users may actually skip this step and provide their own list of sources by simply assigning a list of Scopus source IDs to the property `stefano.search_sources`.

In line 5, `sosia` extracts all authors who publish research articles regularly in these sources. Here, "regular" refers to the parameter `frequency`, which states how often a candidate needs to be observed. A value of 2 means an author needs to publish in the search sources at least every other year, starting from the same year of first publication y_0 including a left margin.[10]

In lines 8 through 10, `sosia` applies the filter criteria with user-provided margins. There are no default margins: If a margin is not specified, that filter will not apply. Here, a candidate must have the same main discipline (as of the day of download), must have started around $y_0 \pm 1$, must have the same number of publications in $y \pm 20\%$, must have the number of coauthors in $y \pm 15\%$, and must have the same number of citations in $y \pm 200$. If matches have been found, they become available through property `stefano.matches()`.

In line 11, users may want to add additional information to the matches. Currently, 14 fields are available.[12] If none are specified, `sosia` will use all the fields, although some require downloading additional data. At the end of this step, property `stefano.matches()` is a list of named tuples. Depending on the use case, this information could

[7] See https://github.com/sosia-dev/sosia-data.

[8] See https://www.elsevier.com/products/scopus/content.

[9] Users have to actively update the information using `sosia.get_field_source_information()`

[10] Technically, `sosia` defines buckets of `frequency` years such that all the years between a's first year and the comparison year are evenly distributed.[11] Then it proceeds to determine the initial search group as the intersection of authors publishing in all these buckets. Thus, the parameter `frequency` can be thought of as the requirement to publish at least every `frequency` years in the search sources.

[11] The first bucket is extended to the left by the provided `first_year_margin` If the last bucket is smaller than `frequency/2`, it is merged into the next-to-last-bucket.

[12] The first name and the surname of the match; the first year and the last year (prior to the match year) in which the match published; the actual number of coauthors, of publications and of citations in y; the research fields; the country, the name and the Scopus ID of the last available mode affiliation in y; the languages of publication; the number jointly cited references.

be used to further filter the candidates. For instance, users may want to restrict matches to specific countries, exclude researchers that publish in a language other than English, or use the first name to estimate a scientist's gender. The field `num_cited_refs` can be used to rank researchers by topical similarity.

General considerations

If no matches have been found, users have three options: Applying a wide definition of similar sources, lowering the frequency with which candidates have to publish in the search sources, or increasing the margins. Feedback on the progress, which is available in all methods through parameter `verbose=True`, can guide the user to get a feel for how to obtain matches, as the process is almost always dependent on the field and the time frame. While the first run for a focal author can be slow, subsequent runs will be much faster as all retrieved data is stored in the local SQLite database.

In previous applications of CEM, margins are more typically determined by a set of chosen percentiles of the productivity distribution in the sample of available researchers. Particularly at the extremes of the distribution, these percentiles allow for implicit margins of acceptance of matches that, computed as the ratio between the target and match authors, would yield margins above 100%.[13] As of now it is left to the user to explore the trade-off between higher quality of the match and shorter search time when using narrower margins, and higher probability to find matches when using broader margins. The user can also define the margins to be a function of the level of productivity of the target scientists. Integrating similar functionalities in `sosia` is among the possible future extensions of the software (see Sect. 5).

Two important steps are currently left to the user. First, some matches may actually turn out to be inadequate due to errors in the Scopus author profiles. One author may be in reality associated with more than one identifier or, more rarely, one identifier may confound different authors. Sometimes this will have no bearing on the quality of the match, for instance, if the additional information only matters after the year of interest. However, in other cases, it may imply that the author has to be dropped from the list of possible matches. In addition, authors that would be good matches may not be identified.

Second, additional non-bibliometric dimensions relevant to the match are obviously not considered by `sosia`. In particular, curriculum information may be relevant, for instance, to ensure that treated and controls have comparable professional situations. Some adjustment in this direction is likely required because matching on productivity does not guarantee in general matching on professional situation.[14] Previous studies have more often started from a rather homogeneous sample of researchers in terms of professional situation (e.g., all already established academic professors, prize winners, etc.), and later balance the

[13] For instance, consider the case of a first bin for publications or citations determined by the 5th or 10th percentile. This would easily correspond to a bin ranging between 0 or 1 publications (citations) to several units (say, up to 5 or 10). This corresponds to equivalent accepted margins that go up to 10 times the productivity of the target author. This is similar at the end of the distribution which is typically highly sparse (i.e. the last decile can in principle include researchers with a number of publications from a few hundred to a thousand publications)

[14] As a matter of fact, whenever focal researchers are biased in favor of established scientists in highly ranked institutions (arguably the most frequent case), match candidates will tend to be in less favorable professional situations, despite their similarity in productivity indicators.

Table 1 Differences of treated and control scientists matched by sosia of the Oettl (2012) sample

	Treated scientists		Control scientists		Difference	
	Mean	Std. dev	Mean	Std. dev	Absolute	t-statistic
Year of First Publication	1965.52	12.05	1965.40	12.23	0.13	0.08
Career Age at Treated Death	34.52	10.88	34.64	10.94	−0.13	−0.09
Publications	147.87	96.68	143.71	96.05	4.16	0.34
Citations	2143.97	2301.57	2091.24	2272.09	52.73	0.18
Coauthors	194.21	123.94	189.53	123.98	4.68	0.30

Table comparing matching characteristics of treated and control scientists akin to Oettl (2012, Table 2), who reports the matching characteristics for 149 immunologists. Our sample includes matches for 126 of these, while 3 immunologists could not be identified on Scopus

productivity levels via CEM. With `sosia,` one does rather do the opposite, narrowing down the search to scientists with similar productivity, leaving to verify other dimensions later.

Testing on samples of previous studies

As an indication of the performance, we replicate the control groups of scientists using `sosia` of a seminal paper in the economics of science, namely Oettl (2012), who studies the premature death of 149 immunologists.[15] We are able to identify Scopus Author Profiles for 143 of the 149 originally included researchers. `sosia` is able to find matches for 126 immunologists (92%). We used the following margins: 30% around the number of publications and the number of coauthors, a margin of 500 citations, and margins of 1 year (for those scientists who began their career after 1996), 3 years (for those scientists who began their career between 1950 and 1995) and 5 years around the year of first publication. We used a narrow definition of search sources, and resorted to a wide definition if no matches were found.

Most of the time `sosia` finds multiple matches, namely up to 346. Since all matches are within the margins, each match is as good as the other. One could now iteratively drop matches whose characteristics are furthest away from the original research. Alternatively, one may want to include researchers of a specific gender or in a specific country. We chose to pick the one that is topically most similar to the treated scientist in terms of jointly cited references, and if multiple share the top rank, we choose one match at random.

The search lasted multiple weeks, with little time of supervision.[16] In Table 1 we present the characteristics of the treated scientists and the control scientists along with the *t*-statistic on the difference. The table mimics Table 2 of Oettl (2012). A direct comparison reveals that sosia achieves a higher matching quality since absolute differences and the corresponding *t*-statistics are smaller (albeit on a smaller sample).

[15] In fact, the study starts with 360 immunologists, but many are dropped because either of a common name or a high career age. 12 immunologists are dropped because no match has been found. The list that was thankfully shared with us includes only the 149 immunologists with matches.

[16] In rare cases, the Scopus API might become unresponsive resulting in exceptions and termination of the progress. In these cases, user must simply restart the script.

Two notes are warranted here: First, we matched on a distinct set of characteristics, so the comparison is limited. Second, we neither verified the completeness of the Scopus profiles nor the comparability of key biographical dimensions, such as professional situation and location.

As discussed in Sect. "Example", the econometrician may want to inspect the matches before moving on to regression analyses. This pertains to the completeness and accuracy of the profile (although this is not always an issue), and the professional situations in the comparison year.

A recent example of an application of `sosia` is Widmann et al. (2022), who retrieve matches for 210 alleged perpetrators of sexual misconduct. On top of similar publication counts, career ages, coauthor counts and citation counts, the authors had three additional requirements: Matches must have been affiliated to a university that is based in the US (both pieces of information are available through `sosia`), and must be of the same gender (this information had to be sourced externally). They found 181 matches.

In a second example, Baruffaldi & Gaessler (2021) found matches for 406 of 427 researchers in departments where adverse events had damaged or destroyed one or more laboratories. Here, the treated sample consists of relatively productive researchers—laboratory heads at the time of the event—active across various countries and fields. The authors progressively expanded the matching margin, up to a maximum of 0.8, for treated researchers with unique characteristics when no initial match was found. Additionally, they conducted an extensive CV review to exclude candidates with incomplete Scopus profiles and in incompatible professional positions.

Limitations and possible future developments

There are two kinds of limitations users of `sosia` need to be aware of, namely limitations due to the quality of Scopus, and technical limitations of `sosia`.

Database-related limitations

`sosia` relies heavily on identifiers for authors (both treated authors and potential matches), as well as for sources (e.g., journals), and possibly affiliations. All of these are imprecise, though the degree of imprecision varies over time, space and disciplines.

Author identifiers Author disambiguation is essential for `sosia`, as each author ID represents a distinct researcher. Issues like split profiles, where a researcher's publications are attributed to multiple author IDs, and merged profiles, where a single author ID includes publications from multiple researchers, can lead to false matches and missed connections. Although previous studies (Baas, 2020); David and Struck, 2019; Aman, 2018) report high-quality profiles, Scopus remains more reliable than name-based, ad hoc author disambiguation (Rose, 2022).

In our experience, disambiguation is especially challenging for authors who frequently relocate, work across disciplines, or have common or multiple names. The issue of multiple names can arise from compoud surnames (as in Spain) or name changes (e.g., after marriage). This bias can lead to both under- and over-representation in profiles. Therefore, users of `sosia` should be particularly cautious when matching Spanish, Korean, or Chinese names, as these profiles may omit some publications or include

too many. We recommend that users manually check the selected Scopus profile to ensure that no relevant profiles are left unmerged.

Source information Source identifiers are also crucial because `sosia` uses them to create the initial set of candidates. Several issues arise here. First, ASJC field assignments can be problematic, as sources may be linked to too many fields or too few. In the first case, irrelevant candidates may appear in the search group; in the second, relevant matches might be missed. According to Wang and Waltman (2016), who analyzed 5,800 Scopus journals, the first issue is more common, with 21% of journals only loosely associated with their assigned ASJC fields.

Scopus updates source-field and source-type data twice a year, mainly to add new sources, but it also sometimes revises fields and types for existing sources, even those no longer indexed. These updates can be inaccurate (e.g., mislabeling a journal as conference proceedings), and Scopus unfortunately does not correct errors in de-indexed sources. These frequent changes introduce instability, as `sosia` might return different matches based on varying source-field and source-type assignments. Therefore, we advise users to update these lists carefully.

Another issue arises from multiple identifiers. A source may have multiple identifiers when it is a conference proceeding or when it is a journal that changes its name or ownership. If additionally Scopus fails to map all identifiers to the same ASJC fields, the initial search group that `sosia` identifies might be too small.

Affiliation profiles By default, `sosia` does not utilize affiliation information beyond providing additional context. However, users can restrict matches to specific affiliations if desired. The quality of affiliation data in Scopus is often considered low. Studies like Schmidt (2016), in the context of institutions, and Donner et al. (2020), for German public institutions, indicate that an affiliation profile rarely covers all relevant publications. The most common reason is missing affiliation data in publication records, often due to missing information from the authors themselves. Hottenrott et al. (2021) estimate that about 2% of research articles in Scopus lack affiliation details, with older articles and those in Social Sciences and Humanities being most affected (up to 10%).

Another issue is the accuracy of affiliation disambiguation. Hottenrott et al. (2021) find that newer publications are more prone to incorrect affiliation assignments, though Scopus frequently corrects these errors. Affiliation profiles are updated at a high rate, which means that `sosia` might initially yield fewer matches. We recommend users to rerun searches after some time to capture updates and corrections in affiliation data.

Technical limitations

Additionally, users must be aware of technical limitations, which are partly our design choices.

Speed The slow processing speed may be the most significant hurdle. Given `sosia`'s need to download and search through extensive data, users should not expect rapid performance. We have taken steps to minimize redundant searches by using simple queries and a local MySQL database. For example, the initial setup for a researcher with about 100 publications involves downloading these records along with additional details (e.g., citations, affiliations). This first-time setup, managed through "pybliometrics," may take two minutes or more. However, setting up the same researcher again reduces the time to about 30 seconds, as the data is cached locally, and using MySQL further shortens the wait to just a few seconds.

Despite these optimizations, predicting the full search duration in advance is challenging due to several variables. Longer searches can result from numerous publications, fields with many or large journals, publications with multiple authors, recent publication years, and additional data retrieval for matches. These factors are inherent and, unfortunately, users have limited control over them.

Restrictive topical fit `sosia` currently searches for match candidates who publish in the same or topically related sources as the target scientist. "Topically related" refers to sources sharing the same 4-digit ASJC codes, with an option for users to include additional fields unrelated to the target scientist's primary field. This approach is conservative, as researchers within the same area often publish in broader or more distant journals. Some studies have created control groups across very broad fields, like "life science." A narrow search may not always be necessary, depending on the study's goals, and may reduce the likelihood of finding matches.

Users can expand `sosia`'s list of considered sources by adding other source identifiers to the relevant list object before filtering authors, as outlined in Sect. "Example". Future `sosia` versions may offer automatic options to broaden the search scope.

Linear margins Using percentage or fixed margins to search for matches can lead to a stricter selection than standard CEM methodology. In CEM, bandwidths for accepted matches are typically set using percentiles of the variable distribution. This approach allows for wider margins, often several times the target scientist's variable value, especially at distribution extremes (e.g., between 1 and 10). For highly productive scientists, margin-based methods can yield fewer matches due to the skewed nature of productivity distribution.

Currently, users can adjust margins based on the target scientist's productivity level or gradually widen the margins if no matches are found. Future versions of `sosia` may offer searches based on fixed bandwidths rather than margins.

Career information In many scientific fields, an author's position in a publication often correlates with their professional role. For example, last authors are frequently laboratory heads, first authors are typically project leaders (often young PhD students or postdocs), and middle authors are more likely to be research assistants. Future versions could leverage this information to improve match quality within `sosia`, not only by finding authors with similar productivity but also by identifying those in comparable professional roles.

Conclusion

`sosia` is designed to automatically look for authors with similar characteristics to a given author in a given year in the Scopus database online. It can be used to build control groups in line with the standard methodologies in the existing leading literature in the Economics of Science.

Particularly for large samples that span different research fields and geographic locations, `sosia` can drastically reduce the time to build control groups. Ex post cleaning and information collection remains necessary to refine the match across relevant biographic dimensions and to eliminate false matches due to imprecise Scopus identifiers. Advanced options allow adjusting the computation time and partly work around the limitations of the basic settings.

`sosia` connects to Scopus through *pybliometrics*. Users need to install and configure both packages. The reliance on Scopus is by design: Alternative databases such as Web of

Science, Semantic Scholar, or OpenAlex do not offer similar web APIs. Key requirements include author entities, a concept of topics to which sources are associated, and the ability to search for authors (or their works) by sources.

`sosia` is an open-source project. Future versions will incorporate improvements to some of the issues discussed. Contributions are welcome.

Author contributions MR: Conceptualization, Software, Validation, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. SB: Conceptualization, Software, Validation, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing

Funding Open Access funding enabled and organized by Projekt DEAL. No funding has been obtained for this study.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

AlShebli, B. K., Rahwan, T., & Woon, W. L. (2018). The preeminence of ethnic diversity in scientific collaboration. *Nature Communications, 9*(1), 5163. https://doi.org/10.1038/s41467-018-07634-8

Aman, V. (2018). Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates. *Scientometrics, 117*(2), 705–720. https://doi.org/10.1007/s11192-018-2895-3

Azoulay, P., Fons-Rosen, C., & Zivin, J. S. (2019). Does science advance one funeral at a time? *American Economic Review, 109*(8), 2889–2920. https://doi.org/10.1257/aer.20161574

Azoulay, P., Graff Zivin, J. S., & Wang, J. (2010). Superstar extinction. *Quarterly Journal of Economics, 125*(2), 549–589. https://doi.org/10.1162/qjec.2010.125.2.549

Azoulay, P., Stuart, T., & Wang, Y. (2014). Matthew: Effect or fable? *Management Science, 60*(1), 92–109. https://doi.org/10.1287/mnsc.2013.1755

Baas, J., et al. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies, 1*(1), 377–386. https://doi.org/10.1162/qss_a_00019

Baruffaldi, S., & Gaessler, F. (2021). The returns to physical capital in knowledge production: Evidence from lab disasters. *Max Planck Institute for Innovation & Competition Research Paper*. https://doi.org/10.2139/ssrn.3912401

Berta, P., Bossi, M., & Verzillo, S. (2017). %CEM: A SAS macro to perform coarsened exact matching. *Journal of Statistical Computation and Simulation, 87*(2), 227–238. https://doi.org/10.1080/00949655.2016.1203433

Blackwell, M., et al. (2009). Cem: Coarsened Exact Matching in Stata. *The Stata Journal: Promoting Communications on Statistics and Stata, 9*(4), 524–546. https://doi.org/10.1177/1536867X0900900402

Colatat, P. (2015). An organizational perspective to funding science: Collaborator novelty at DARPA. *Research Policy, 44*(4), 874–887. https://doi.org/10.1016/j.respol.2015.01.005

David, C., & Struck, B. (2019). Reliability of Scopus author identifiers (AUIDs) for research evaluation purposes at different scales. In: *Proceedings of the 17th International Conference of the International Society for Scientometrics and Informetrics (ISSI)*. Vol. II, pp. 1276–1287.

Donner, P., Rimmert, C., & van Eck, N. J. (2020). Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. *Quantitative Science Studies, 1*(1), 150–170. https://doi.org/10.1162/qss_a_00013

Ganguli, I. (2015). Immigration and ideas: What did Russian scientists bring to the United States? *Journal of Labor Economics, 33*, S257–S288. https://doi.org/10.1086/679741

Graddy-Reed, A., Lanahan, L., & Ross, N. M. (2018). The effect of R &D investment on graduate student productivity: Evidence from the life sciences. *Journal of Policy Analysis and Management, 37*(4), 809–834. https://doi.org/10.1002/pam.22083

Hird, M. D., & Pfotenhauer, S. M. (2017). How complex international partnerships shape domestic research clusters: Difference-in-difference network formation and research re-orientation in the MIT Portugal program. *Research Policy, 46*(3), 557–572. https://doi.org/10.1016/j.respol.2016.10.008

Hottenrott, H., Rose, M. E., & Lawson, C. (2021). The rise of multiple institutional affiliations in academia. *Journal of the Association for Information Science and Technology, 72*(8), 1039–1058. https://doi.org/10.1002/asi.24472

Hussinger, K., & Palladini, L. (2024). Information accessibility and knowledge creation: The impact of Google's withdrawal from China on scientific research. *Industry and Innovation, 31*(6), 753–783. https://doi.org/10.1080/13662716.2023.2298293

Iacus, S., King, G., & Porro, G. (2009). Cem: Software for coarsened exact matching. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v030.i09

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20*(1), 1–24. https://doi.org/10.1093/pan/mpr013

Kahn, S., & MacGarvie, M. J. (2016). How important is U.S. location for research in science? *Review of Economics and Statistics, 98*(2), 397–414. https://doi.org/10.1162/REST_a_00490

Klingbeil, C., Semrau, T., Ebers, M., & Wilhelm, H. (2019). Logics, leaders, lab coats: A multi-level study on how institutional logics are linked to entrepreneurial intentions in academia. *Journal of Management Studies, 56*(5), 929–965. https://doi.org/10.1111/joms.12416

Liu, L., Jones, B. F., Uzzi, B., & Wang, D. (2023). Data, measurement and empirical methods in the science of science. *Nature Human Behaviour, 7*(7), 1046–1058. https://doi.org/10.1038/s41562-023-01562-4

McKinney, W. (2010). Data structures for statistical computing in python. *SciPy, 445*, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Moed, H. F., Aisati, M. H., & Plume, A. (2013). Studying scientific migration in Scopus. *Scientometrics, 94*(3), 929–942. https://doi.org/10.1007/s11192-012-0783-9

Mohnen, M. (2022). Stars and brokers: Knowledge spillovers among medical scientists. *Management Science, 68*(4), 2513–2532. https://doi.org/10.1287/mnsc.2021.4032

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of web of science and Scopus: A comparative analysis. *Scientometrics, 106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Muschallik, J., & Pull, K. (2016). Mentoring in higher education: Does it enhance mentees' research productivity? *Education Economics, 24*(2), 210–223. https://doi.org/10.1080/09645292.2014.997676

Oettl, A. (2012). Reconceptualizing stars: Scientist helpfulness and peer performance. *Management Science, 58*(6), 1122–1140. https://doi.org/10.1287/mnsc.1110.1470

Rose, M. E. (2022). Small world: Narrow, wide, and long replication of Goyal, van der Leij and Moraga-González (JPE 2006) and a comparison of EconLit and Scopus. *Journal of Applied Econometrics, 37*(4), 820–828. https://doi.org/10.1002/jae.2886

Rose, M. E., & Kitchin, J. R. (2019). pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. *SoftwareX, 10*, 100263. https://doi.org/10.1016/j.softx.2019.100263

Schmidt, C. M., et al. (2016). Gaps in affiliation indexing in Scopus and PubMed. *Journal of the Medical Library Association, 104*, 2. https://doi.org/10.5195/jmla.2016.60

Teodoridis, F., Bikard, M., & Vakili, K. (2019). Creativity at the knowledge frontier: The impact of specialization in fast- and slow-paced domains. *Administrative Science Quarterly, 64*(4), 894–927. https://doi.org/10.1177/0001839218793384

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics, 10*(2), 347–364. https://doi.org/10.1016/j.joi.2016.02.003

Widmann, R., Rose, M. E., & Chugunova, M. (2022). Allegations of sexual misconduct, accused scientists, and their research. *Max Planck Institute for Innovation & Competition Research Paper*. https://doi.org/10.2139/ssrn.4260210

Yadav, A., McHale, J., & O'Neill, S. (2023). How does co-authoring with a star affect scientists' productivity? Evidence from small open economies. *Research Policy, 52*(1), 104660. https://doi.org/10.1016/j.respol.2022.104660