Technical Appendix to Macrodynamics of Economics: A Bibliometric History

François Claveau* Yves Gingras†

June 15, 2016

Contents

*Francois.Claveau@USherbrooke.ca, Université de Sherbrooke, Canada Research Chair in Applied Epistemology, and Centre interuniversitaire de recherche sur la science et la technologie.

†Gingras.Yves@UQAM.ca, Université du Québec à Montréal, Canada Research Chair in History and Sociology of Science, and Centre interuniversitaire de recherche sur la science et la technologie.

This document is a technical companion to the article *Macrodynamics of Economics: A Bibliometric History*. It gives additional information on data and empirical procedure.

1 Data

1.1 Selection of economics journals

The Observatoire des sciences et des technologies (OST, UQAM) assigns each journal in the Web of Science to a field of research. Its classification is based on the one used by the US National Science Foundation (NSF) since the 1970s (National Science Board, 2006, Appendix table 5-39). The first step in our sampling of mainly-economics and partly-economics journals is to take the 387 journals assigned to the field 'economics'. These journals are all kept as mainly-economics journals in this study.

Since fields are mutually exclusive in the OST classification, some journals are not assigned to the field 'economics', but their contents pertain mainly or partly to economics. The Web of Science includes a 'Web of Science Category' variable, each journal being associated to one or more subjects. By extending the sample to include journals having a term with the stem 'econom' in their subjects, we reach 514 journals.

Among the journals added to the sample based on their subject field, some must be considered as mainly economics (e.g., *Journal of Finance*) and others as partly economics (e.g., *Annals of the New York Academy of Sciences*). The list of strings in Table 1 is used to discriminate among these journals. These strings are either stems highly related to economics, translations of economics terms, or full titles of journals judged by manual inspection to have substantial economics contents. The journals having at least one of these strings in their titles are included in the category of mainly-economics journals. The remaining 26 journals are considered partly-economics journals.

econ	tax	derivatives
admin	employ	forecasting
business	market	ekon
finan	risk	wirtschaft
trade	real estate	betrieb
industr	management	Value in Health
capital	productivity	Journal of Agrarian Change
labo	technolo	Kommunist

Table 1: *Strings related to economics*

Finally, 44 journals in the Web of Science are neither assigned to the field 'economics' nor to a subject with stem 'econom', but have this stem in their title. We manually inspected these publications to either leave them out of the corpus (9 journals), include them as partly-economics journals (12 journals) or include them as mainly-economics journals (23 journals).[1]

Appendix A includes the list of all journals in the sample. The number of journals has increased through time (see the Figure in the main article), with especially pronounced growth rates from the mid-1960s to the mid-1970s and since 2005. These periods of extreme growth are more likely due to the peculiar history of the database than to the evolution of economics. The first period corresponds to the early years of the database, the *Science Citation Index* being first published in 1964 (Wouters, 2006). The most recent years correspond to the first time the Web of Science faces a direct competitor: Scopus, launched in late 2004.

Three caveats have to be kept in mind for this first step of the sampling procedure. First, in sampling exclusively from journals, this study does not cover other modes of scientific communication, notably books.

[1] In the three categories, the most important publications in terms of number of documents are: *Journal of Economic Entomology* for left out journals with 24,029 documents, *Forest Policy and Economics* for partly-economics journals (966 documents), and *Finance a Úvěr - Czech Journal of Economics and Finance* for mainly-economics journals (542 documents). As the numbers attest, the decision to exclude the *Journal of Economic Entomology* is significant while the other journals are a lot smaller.

In economics, journal articles have been central since a long time – in contrast to other social sciences and humanities. This restriction is thus unlikely to significantly affect the relevance of this study. Second, the sample is necessarily restricted to what the Web of Science covers, which is not all journals in economics. Yet, the database has aimed, since its early years, at indexing "the world's most important and influential journals" (Testa, 2012). Since scientific influence is known to be highly concentrated, we can be confident that the 549 journals in this study give a good *macro* picture of what economics has been about through time. Third, some journals included in Web of Science and publishing some economics contents will fail to be selected by the sampling procedure. For instance, some economics articles have been published in generalist journals such as *Science* and *Nature*. But again, the results of this study are unlikely to be significantly affected by these imperfections. All in all, we can be confident that most of the important journals directly relevant to economics have been included in this first step. The next step attempts to avoid the opposite bias: having many documents in the corpus that have little to do with economics.

1.2 Documents

This study focuses on three types of documents—articles, notes and reviews—because these are the journal contents typically containing references.[2] With all the journals in the sample, the number of these documents amounts to 534,078. Some of these documents do not include references. After removing these documents, we are left with 485,690 entries.

All the documents in the mainly-economics journals are kept in the final corpus (409,462 items). It is however necessary to trim the partly-economics journals to avoid including in the corpus a lot of documents that have nothing to do with economics. The procedure used is to keep only the documents having a word with the stem 'econom' in either their title, abstract or keywords. The rationale is that authors publishing an economics paper in a journal that is not mainly about economics will often signal their field by using words with such a stem. Among the 76,228 documents in these partly-economics journals, 6765 are kept in the final corpus.

The corpus is left with 416,227 documents. Figure 1 illustrates the annual growth rate of publications. Although growth fluctuates quite dramatically from year to year (especially before the 1960s), average annual growth has been more or less steady – the average annual growth rate for the whole sample is 0.06. The corpus reaches 100 publications a year in 1920; 500 publications in 1942; 1000 publications in 1956; 5000 publications in 1977; and 10,000 in 2006. For 2014 – the last year included in this study – the number of documents in the corpus is 19,196.

This step of the corpus selection is subject to its own issues. First, the procedure is permissive by including all the documents from the mainly-economics journals in the corpus, but some of these documents might be best assigned to a different discipline. Since the proportion of these documents is unlikely to be large and since there is no easy and reliable method to almost perfectly discriminate between economics and non-economics documents, the procedure used in this study seems reasonable. Second, the procedure is perhaps too restrictive when it comes to partly-economics journals. It is probable that some documents that can legitimately be considered 'economics' do not have a word with the stem 'econom' in the text fields available in the database. An alternative strategy could have been to use a keyword list in the spirit of the one developed for selecting journals (see Table 1). The weak point of this strategy comes from the fact that keywords used in economics are also used by other scholars – especially by other social scientists. Extending the keyword list for document selection beyond 'econom' can thus easily make the corpus a lot too permissive – that is, it might come to include far more than the 8.9 % of documents now included from the partly-economics journals, many of these documents being in fact false positives. It seems therefore better to use our restrictive sampling procedure.

Finally, the selection procedure faces a limitation in the database: the indexing of abstracts and keywords only started in the late 1980s. Figure 2 shows the proportion of documents published in a given year and having their abstracts or keywords indexed. From 0, it jumps to more than 50 % at the beginning of the 1990s and is converging toward unity since then. The implication of this data limitation is that only titles

[2] Articles are vastly more numerous than notes and reviews. They account for 95% of the documents in the final corpus.

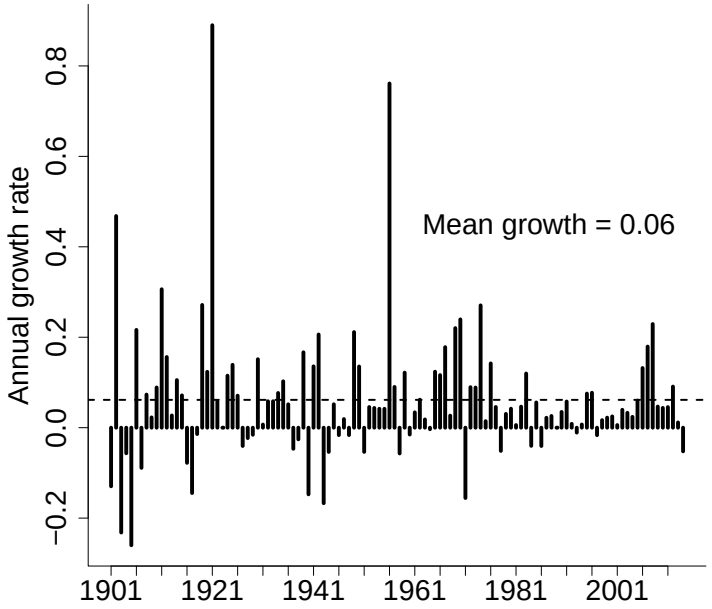


Figure 1: *Growth rates of documents in economics (full corpus)*

can be searched prior to the 1990s for the stem 'econom'. It is thus highly likely that a smaller proportion of *economics* documents in partly-economics journals are identified as such by the procedure before the 1990s. Indeed, Figure 3 is evidence of such bias: both the proportion of documents kept in partly-economics journals and the proportion of the retained documents in the total corpus have been increasing since the early 1990s. The perfect timing of these processes makes it unlikely that the bigger share going to documents from partly-economics journals is mainly due to a real increase in the share of economics documents in these journals. This evolution must be, at least partly, attributed to a higher rate of false negatives in the years prior to the indexing of abstracts and keywords. Note however the relatively small magnitudes involved. The retention rate oscillated around 0.04 in the decades before the 1990s and it rose to a maximum of 0.31 in recent years. This is a sharp increase, but it also leaves the vast majority of the documents in partly-economics journals out of our sample – which speaks in support of the necessity of trimming these journals to avoid the inclusion of a lot of irrelevant documents in our corpus. More importantly, Figure 3 shows that the documents retained from partly-economics journals make only a small proportion of the corpus even today; the maximum percentage reached until now is 2.7 %. There is thus little reasons to fear that the non-ideal sampling based on titles, abstracts and keywords will significantly bias the analysis.

1.3 References in documents

In this study, the analysis of the research dynamics in economics relies extensively on the references appearing in documents. At this point, the corpus includes a total of 11,216,142 references.

1.3.1 Available information to identify references

There are seven fields available in the database that can be used to identify each reference: two identifier fields, first author, year of publication, publication name (journal name for journal articles), volume and starting page. Life would be easy if the identifier fields did their job well enough. Unfortunately, the first identifier has a really poor coverage: only 0.2 % of the references in the corpus have a value for this field. The

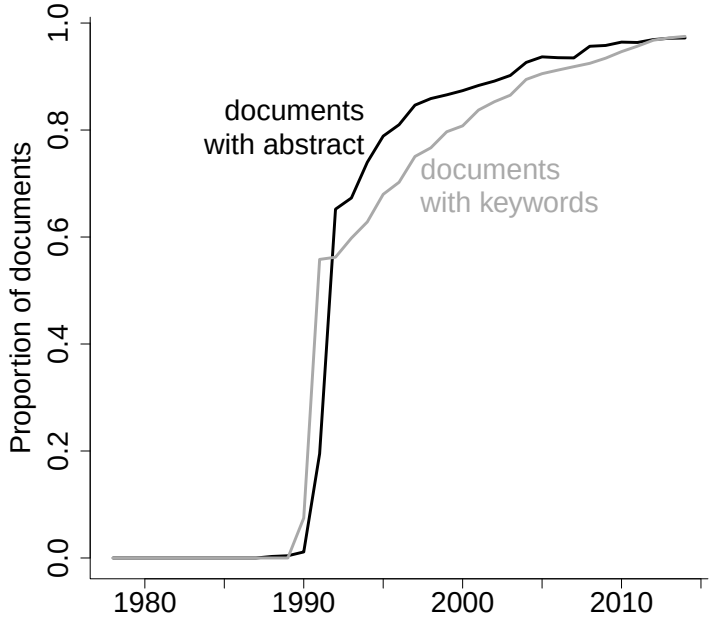


Figure 2: *Proportion of documents having an abstract and keywords*

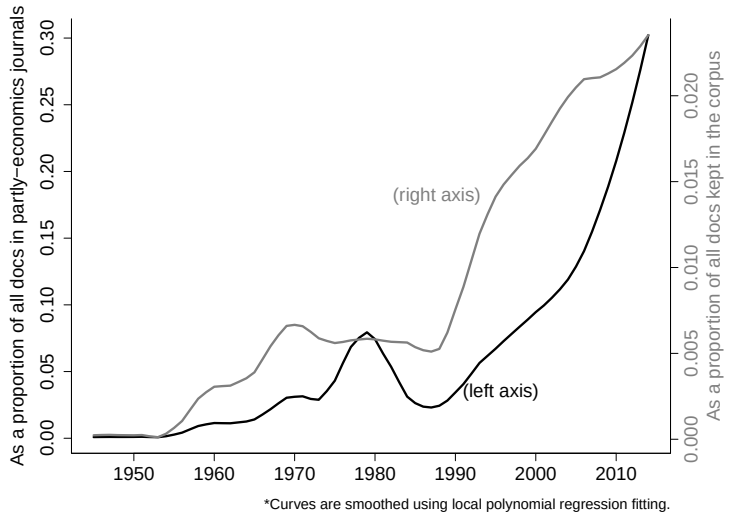


Figure 3: *Documents kept from partly-economics journals*

second identifier field has a significantly better coverage of 60.2 %. Using this field would still imply dropping a large proportion of the references in the corpus. Is it possible to do better? This second identifier is the benchmark that we should try to surpass; we will call it the 'WoS ID' because it is supplied by Thomson's Web of Science. The first identifier is uniformly worse than the WoS ID and will not be discussed any further.[3]

The goal is to create a new identifier field by combining the information from the fields available in the database. There are three desiderata in constructing such a field. First, one would like to identify as many references as possible, thus beating the coverage of the other identifier fields. Second, the same identifier should not be given to two references that actually stand for different documents (spurious match). Third, different identifiers should not be given to references to the same document (spurious mismatch).

If the fields 'first author', 'year', 'publication', 'volume' and 'first page' were all filled without transcription errors, the best identifier would be to concatenate these five fields. Each reference would have an identifier, thus maximizing coverage. The same document would have the same identifier every time it is referred to, thus minimizing spurious mismatch. Finally, since it is highly unlikely that two documents have the same true values for all these fields, the number of spurious matches would be extremely low.

Now, the first problem is that fields are not filled systematically. Almost all author (97.9 %), year (98.4 %) and publication (100 %) fields have values; but the volume and first page fields are far more sparsely populated (56.5 % and 61.9 %). As much as 34.9 % of the references have neither a volume nor a page. This distribution of missing values is mainly explainable by the fact that a great proportion of cited documents are not journal articles and thus do not have a volume and a (meaningful) starting page. For instance, books (e.g., Keynes' *General Theory*) and reports (e.g., the annual *OECD Economic Outlook*) should not be expected to have values for these fields. But they typically have values for the other fields. In particular, the publication field indexes a shortened version of their titles (for instance, `GENERAL THEORY EMPLO` or similar for Keynes' and something like `ECON OUTL` for the OECD's).

Second, fields also suffer from transcription errors or discrepancies. Take Keynes' *General Theory* again for illustration. One would expect the fields author, year and publication to be filled by `KEYNES-JM`, `1936` and `GENERAL THEORY EMPLO`, and the fields volume and starting page to be empty. 1453 references have exactly these values for the five fields. Yet, there are clearly more references to this classic book in the database. 2464 references have `KEYN` as part of the author field and `GENERAL TH` as part of the publication field. There is no reason to believe that even a handful of these matches are spurious. In consequence, at best only 59 % of the references to Keynes' *General Theory* have the expected values in all the fields.

Tables 2 gives the most common values for the six fields in the references to Keynes' book. For each field except WoS ID, the expected value is by far the most common. However, the proportion of spurious mismatches is compounded by asking that all fields match perfectly. The first column of the Table reports the WoS IDs for this document, which should ideally be unique and span all the references to it. All to the contrary, the vast majority of the references to Keynes' *General Theory* have no value for this field. Furthermore, the references to this book having a WoS ID are divided among 269 different values for this field. There is thus both extremely bad coverage and a high degree of spurious mismatch for this book. This result seems to be generalizable to books and reports: the field WoS ID does a really poor job at identifying these references.[4]

One field requiring particular attention is 'first author'. When the author has multiple first names or has initials, the field tends to exhibit a high level of spurious heterogeneity. This is not such a big issue for the author of the *General Theory* – the vast majority of citations to this book are indexed with author `KEYNES-JM` (see Table 2). But take, for instance, Wooldridge's (2002) highly-cited book *Econometric Analysis of Cross Section and Panel Data*. This document is cited 2001 times with author `WOOLDRIDGE-JM`, but also 836 times

[3] By uniformly worse, we mean that, for *any* subset of the corpus, the best the first identifier does is to do as good as the WoS ID. There is no reference having the first identifier but failing to have a WoS ID (strict subset for coverage). Furthermore, any two references sharing the same first identifier also share the same WoS ID (no possibility of reduction in spurious mismatch); and the reverse holds too (no possibility of reduction in spurious match).

[4] One key reason why it does a poor job is because having a value for the 'first page' field is an almost prerequisite for having a WoS ID and different pages result in different WoS IDs. Unfortunately, first page does not make sense for books; when the field has a value, it is not the true value.

WoS ID	freq	Author	freq	Year	freq
Null	2072	KEYNES-JM	2339	1936	1824
6971266	14	KEYNES	89	1964	169
27941631	7	KEYNES-J	22	1973	112
43900376	6	KEYNES-M	3	*Null*	101
43900345	5	KEYNESJM	3	1935	54
45603138	5	KEYNE	1	1979	21
Vol	freq	Publication	freq	Page	freq
Null	2361	GENERAL THEORY EMPLO	2283	*Null*	1974
14	38	GENERAL THEORY	61	383	24
7	28	GENERAL THEORY 2	19	201	10
13	19	GENERAL THEORY AFT 2	15	CH12	9
29	16	THE GENERAL THEORY O	13	32	9
12	1	GENERAL THEORY 1	11	372	7

Table 2: *Frequencies of field variations for references to Keynes' GT*

with author `WOOLDRIDGE-J` (i.e., missing the last initial).[5] The author field is thus likely to be an important source of spurious mismatch that should be dealt with.

Other documents – journal articles in particular – are far better and more easily identified. Take, for instance, Kahneman and Tversky's 'Prospect Theory' published in *Econometrica* in 1979. A customized search in the corpus locates 2626 references to this article, 89.6 % of which are identified by matching on the expected values for the five fields besides WoS ID.[6] There is also a unique WoS ID for these identified references; the identifier does well here. Table 3 reports some of the variation in the entries for the six fields. The deviant values have extremely low frequencies.

WoS ID	freq	Author	freq	Year	freq
45628	2352	KAHNEMAN-D	2603	1979	2623
10917822	61	TVERSKY-A	6	1982	3
14337332	55	KAHNEMAN-P	4		
Null	25	KAHNEMAN-A	2		
Vol	freq	Publication	freq	Page	freq
47	2576	ECONOMETRICA	2610	263	2393
46	20	PROSPECT THEORY ANAL	6	91	61
Null	16	ECONOMETRICA MAR	5	313	55
4	3	ECONOMETRICA J ECONO	5	*Null*	23

Table 3: *Frequencies of field variations for references to Kahneman & Tversky (1979)*

1.3.2 Procedure for the matching ID

A new identifier – called 'matching ID' for now on – has been created by the following procedure. First, only the first initial has been kept in the author field – for instance, `KEYNES-JM` becomes `KEYNES-J`. The goal of this first step is to reduce spurious mismatches stemming from differences in how journals or authors treat first names and initials. A risk of this change is that it could introduce spurious matches – i.e., two authors being seemingly the same because the initial differentiating them has been deleted. However, our *a*

[5] Similarly, Nelson and Winter's (1982) *Evolutionary Theory of Economic Change* has 1180 hits with `NELSON-RR` and 468 hits with `NELSON-R`.

[6] The expected values for the fields are: author=`KAHNEMAN-D`, year=`1979`, publication=`ECONOMETRICA`, volume=`47`, start page=`263`. To find discrepant references, the corpus has been searched for entries with author as being `KAHNEMAN|KANEMAN|TVERSKY`, publication as `ECONOMETRIC|PROSPECT TH` and year as being between 1975 and 1985 (i.e. '|' means 'or').

posteriori analysis of the matching ID has been unable to detect even a single case of this introduction of spurious matches in the corpus.[7]

Second, only a subset of the references has been kept. These are the references having either a WoS ID or having an empty volume field but values for author, year and publication. The last group of references are mostly books and reports (they usually do not have a meaningful value for volume); only the ones having enough information to avoid spurious matches are kept. This subset represents 93.5 % of the references in the corpus. This is a net improvement over the coverage of the WoS ID (60.2 %). It might be possible to increase coverage even more, but the possible gains are becoming smaller while the risk of spurious ID attribution increases.[8]

Finally, a matching ID has been attributed to each reference in the subset by enforcing two conditions:

1. References having the same WoS ID are given the same matching ID;

2. References without a volume and for which the author, year and publication fields perfectly match are given the same matching ID.

Meeting these two conditions jointly should help minimize spurious mismatches because the procedure benefits both from the fuzzy matching that has been done in the construction of WoS ID and from the extra information coming from taking into account the references that had no WoS ID. Indeed, 74,752 WoS ID out of 2,153,666 are found to be redundant.

Take again Keynes' *General Theory*. Among the 2464 references found above with a customized search, a single matching ID is attribute to 1754 (a 71.2 % accurate matching). This identification is already pretty good. The main reason preventing it to be even better is clear from Table 2: many references are to later editions. The field year thus exhibits a relatively high level of heterogeneity.

In sum, the matching ID generated with this procedure has a broad coverage and it arguably does better than WoS ID with respect to spurious mismatches. When it comes to spurious matches, the relative performance of the matching ID is harder to assess. In our various dealings with the corpus, no case of spurious match came to our attention. Furthermore, the truncation of the author field has been systematically investigated to find spurious matches and nothing was found. Finally, the references missing some key values (e.g. author or year) have been dropped: this choice sacrifices some coverage in order to decrease the risk of spurious matches.

1.3.3 Quality change of the references

The coverage of the matching ID is not uniform through time, as shown in Figure 4. The difference between the minimum coverage of 12.8 % in 1905 and the maximum coverage of 97.5 % in 2014 is extreme. This is due to the fact that the indexing of references has become more systematic. As the tables in Appendix B illustrate, there are far more empty fields in the references of documents from the early twentieth century than there are for the more recent documents. Since the data quality is poor for the period before 1956 – i.e., the coverage being below 2/3 – we define our period as starting in 1956 and we focus on analyzing economics after this point.

1.3.4 Distribution of references

In the main article, we report that the mean number of references per document as steadily grown through time. Figure 5 depicts other properties of the distribution of references through time. From Panel (a), we see that the normalized spread around the growing mean has been decreasing through time, with the

[7] Almost all different spellings of authors are of the sort `SEN-A` and `SEN-AK` or `BECKER-G` and `BECKER-GS` – obviously, two ways to refer to the *same* authors. Furthermore, since a document is not identified solely by the content of its author field, it is highly unlikely that all the fields match – including the truncated author field – although the authors are not the same.

[8] 47.1 % of the references excluded from the subset have no WoS ID but have values for author, year, publication and volume. These references are the prime candidates for broadening the subset of references having a matching ID. The other references have no values in at least one of the following fields: author, year and publication. It seems clearly better to left them out of the analysis.

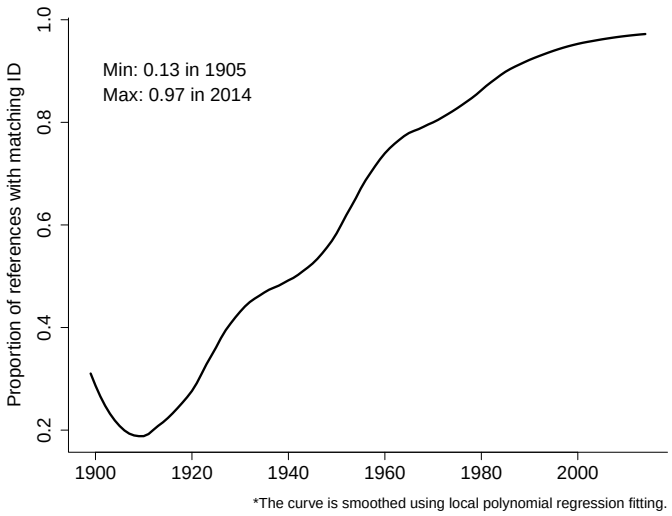


Figure 4: *Coverage of the matching ID through time*

notable exception of a spike in the mid 1980s. Panel (b) confirms that something peculiar happened at this moment: a few documents with extreme numbers of references skewed the distribution. A closer inspection of documents by z-scores found a few with extreme values; documents having typically more than a thousand references. As a precautionary measure, 8 documents with extremely long lists of references have been removed from our corpus.

2 Empirical procedure

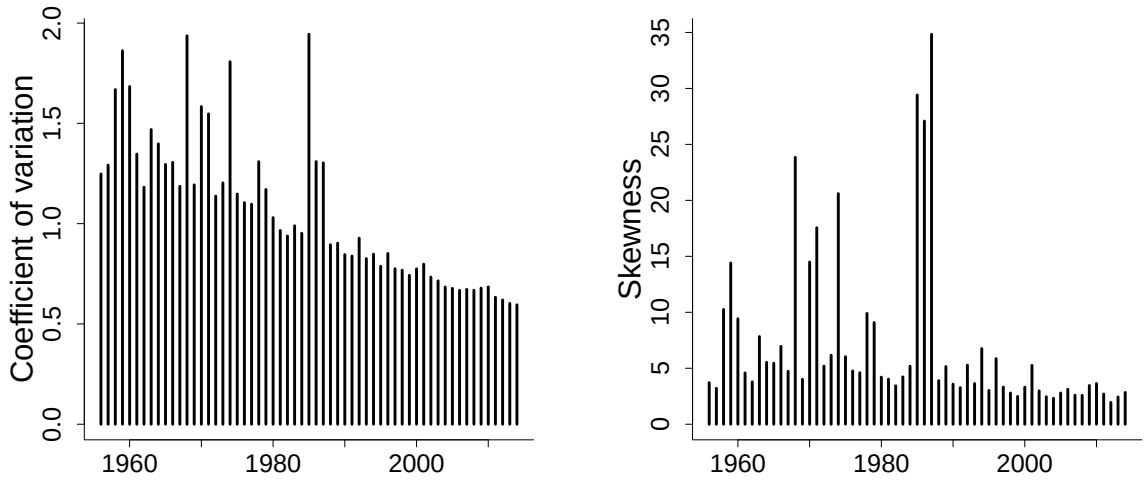
2.1 Further information on bibliographic coupling

Co-citation analysis (Small, 1973) is more frequently used in the literature than bibliographic coupling. In co-citation analysis, the nodes are the *cited* documents. If two nodes are co-cited in a citing document, there is an edge between the two. The weight between two nodes in co-citation analysis is proportional to the number of times the two nodes are co-cited in the corpus.

Bibliographic coupling is used here instead of co-citations because it capture more directly the property of concern: the cognitive proximity of two papers in economics. In contrast, a map based on co-citations -1- would include non-economics papers (but cited in economics) and -2- would be a more indirect measure of the nodes' cognitive proximity. Two papers can be quite often co-cited, while having been quite distant contributions at the time of publication. For instance, 80 documents co-cite Becker's (1965) famous article 'A Theory of the Allocation of Time' and Heckman's (1979) no less famous 'Sample Selection Bias as a Specification Error', though the two papers are on quite different topics (in fact, they share no reference and Heckman does not cite Becker). The relatively high number of co-citations of these two articles is direct evidence that the resources of both are sometimes simultaneously brought together in economics, but it is hardly evidence that the two papers are making similar contributions – i.e. that they are cognitively close. In contrast, if two citing documents cite both Becker (1965) and Heckman (1979), it is evidence that these two citing documents are cognitively close.

As described in the main article, a moving window of a couple of years is used for the dynamic network. A routine in R has been written to create, by bibliographic coupling, a weighted edge list from the references of all documents.[9] In our main specification, only documents that were published less than 5 years apart

[9] In R (R Core Team, 2013), the routine was made a lot quicker by the use of the packages `data.table` (Dowle et al., 2013),



(a) *Normalized spread around the mean* **(b)** *Asymmetry of the distribution*

Figure 5: *Properties through time of the distribution of references per document*

are coupled because more distant documents never appear together in the network. We also provide below [**REFS**] results for a moving window of 3 years.

2.2 Further information on automated community detection

The algorithm used in this study, like many popular algorithms for community detection, attempts to maximize "modularity", which is "a measure of the quality of a particular division of a network" (Newman and Girvan, 2004, p. 7). Suppose that we have a weighted network with n nodes. The edges of this network are represented in a symmetric, $n \times n$ matrix $\mathbf{A}$, where element A_{ij} is equal to the weight of the edge between nodes i and j (and is zero if there is no edge). The degree k_i of a node i – i.e., the total weight of its edges – is

$$k_i = \sum_j A_{ij}. \tag{1}$$

The total weight m of the edges for the whole network is

$$m = \frac{1}{2} \sum_{ij} A_{ij}. \tag{2}$$

Given a partition P on this network, we denote by c_i the community in which node i is.[10] The fraction F of weighted edges connecting nodes *within* communities is

$$F = \frac{\sum_{ij} A_{ij} \delta(c_i, c_j)}{2m}, \tag{3}$$

where $\delta(c_i, c_j)$ is an indicator function equal to 1 if $c_i = c_j$ (if the two nodes are in the same community) and 0 otherwise. F measures how much connections are concentrated within communities rather than

`foreach` (Revolution Analytics and Weston, 2013b) and `doParallel` (Revolution Analytics and Weston, 2013a).

[10] A partition of a set divides all elements in *non-overlapping* subsets. An implication is that the procedure used in this study does not allow a node (a document) to belong to more than one community. Though this can be seen as a limitation, it is reasonable to hypothesize that each publication will *typically* be located in what we intuitively regard as a single specialty. It would therefore be surprising that using some of the recent algorithms detecting *overlapping* communities (see Xie et al., 2013) would radically change the global structure of detected specialties.

between them. It is thus a measure of *community disconnection*. Now, the modularity Q of partition P is F subtracted by the expected fraction of weighted edges that would fall inside communities if edges where distributed randomly, but keeping the degrees of nodes fixed. With this random allocation, the expected weight between nodes i and j is $k_i k_j / 2m$ instead of A_{ij}. Starting from equation (3), we reach an expression for modularity Q (Newman, 2004, p. 6):

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j). \tag{4}$$

Among modularity-based algorithms, we use the Louvain method, which has been shown to perform especially well on various networks (Blondel et al., 2008).[11] This algorithm is simple. It starts by putting each node in a distinct community. It then goes through the nodes one by one, evaluating the change in modularity that would result from moving the node out of its community and into one of its neighbour's community. The transfer maximizing modularity is enacted, provided the change in modularity is positive. The algorithm processes each node sequentially with as many loops as necessary until no further improvement in modularity is possible with this procedure. This ends the first iteration of the algorithm. For the next iteration, the nodes in each community are collapsed into one node, and the sequential procedure is repeated at the level of the community-nodes – i.e., for each community-node, the algorithm considers which merge with another community (if any) maximizes modularity. The algorithm ends this second iteration (which may involve many loops over the community-nodes) when modularity cannot be improved by moving single community-nodes. For the next iteration, the new communities are themselves considered nodes and the procedure is repeated. When an iteration does not produce any new merge, the algorithm stops: the community assignment of the nodes in the original network is given by the aggregate node in which they are included.

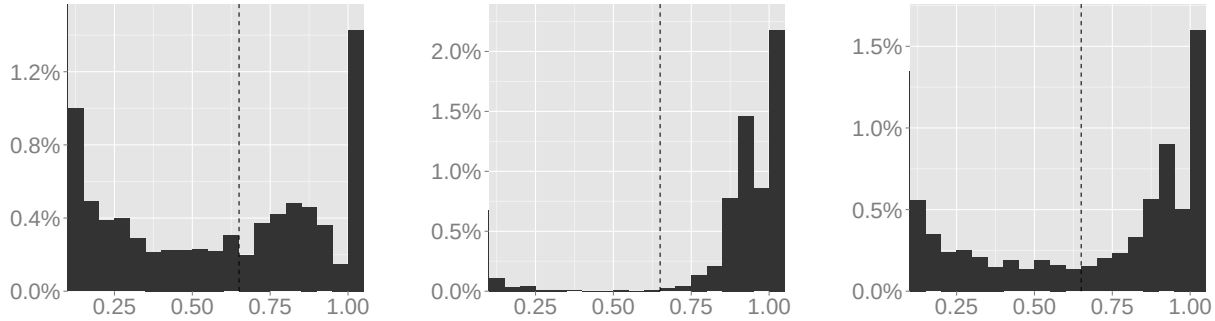
The Louvain method is meant for static networks. In contrast, this study analyses a dynamic network with a multi-year window moving through time. How can the Louvain method be adapted to a dynamic network? The simplest procedure would be to run the algorithm separately for each time window and to then use a metric to find out which cluster in window $w + 1$ can legitimately be considered the child of a cluster in w.

The metric γ_{ij} is described in the main document. Here, we will first discuss the chosen thresholds of γ_{ij}. We will then explain why the procedure of running the Louvain method for each time window independently is not satisfactory and we will describe and justify the chosen procedure.

A threshold must be chosen for γ_{ij} to decide when a cluster j inherits sufficiently from i to be considered a child of i. The chosen threshold in this study is 0.65. For values of γ_{ij} above this threshold, cluster j is considered a child of i. Why this value? Note that the distribution of γ_{ij} is strongly bimodal. For more than 90 % of the *i*-*j* pairs, γ_{ij} is lower than 0.1. As can be seen in Figure 6, the distribution of γ_{ij} peaks again when the measure approaches 1. In between the two extremes, frequencies are extremely low. Figure 6 presents the distribution of γ_{ij} for different versions of the community detection algorithm (versions to be discussed below). The chosen threshold seems especially appropriate for the version in panel (c), which is the one that we in the end use. In this case, the frequency of γ_{ij} starts increasing just after 0.65. The distribution of γ_{ij} is most dramatic for the version shown in panel (b). In this case, choosing a threshold between 0.5 and 0.8 would make little difference to the outcome because frequencies are almost null in this interval. For the version in panel (a), the number of children detected would be more sensitive to the threshold.

The other threshold distinguishes cluster splits from simple dispersion. It is set to 0.25. The distinction between a splitting and a dispersing cluster is of second-order importance for this study. The threshold of 0.25 can be justified by the distribution of γ_{ij} (i.e., Figure 6 shows that it is around where the measure settles to low frequencies), but a different threshold could have been chosen with little consequences (only a different share of clusters dying by dispersion rather than splitting).

[11] Although we use the igraph package in R for handling network graphs (Csardi and Nepusz, 2006), we do not use the implementation of the Louvain method in this package because it does not include the option of initializing the algorithm with a customized community assignment. Instead, we use version 0.2 of the C++ implementation provided by the developers of the Louvain method on their website (last accessed Nov. 3rd 2015). We simply wrote a function in R to interact with this program.

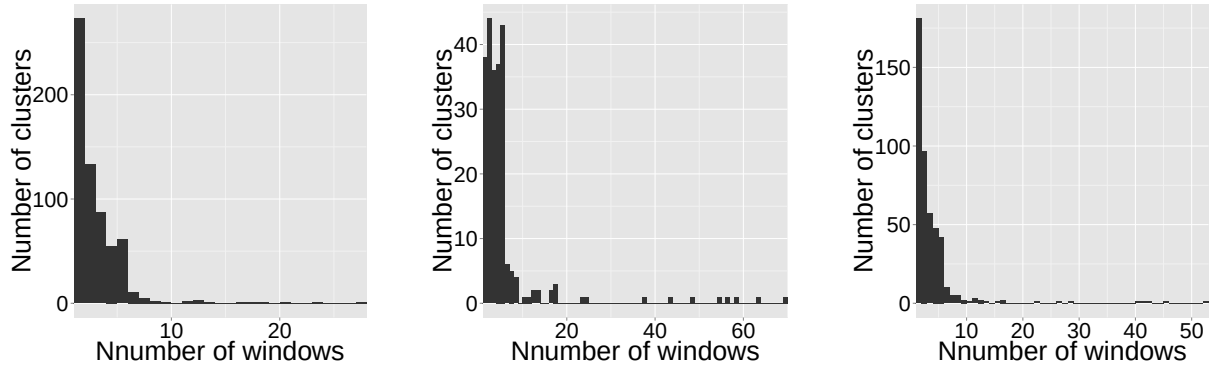


(a) *Unrestricted version,* 92 % *are smaller than 0.1*

(b) *Restricted version,* 94 % *of* γ_{ij} *are smaller than 0.1*

(c) *Low-level version,* 93 % *of* γ_{ij} *are smaller than 0.1*

Figure 6: *Distribution of* γ_{ij} *for different versions of the algorithm. Dotted line is the threshold for parenthood attribution. All instances of* γ_{ij} *below 0.1 are not shown (i.e., the vast majority).*



(a) *Unrestricted version*

(b) *Revised version*

(c) *Low-level version*

Figure 7: *Number of clusters by their life lengths (i.e., the number of time windows during which they exist) for different versions of the algorithm.*

We now turn to discussing the choice of the community detection algorithm. The most straightforward way to use the Louvain method to identify clusters in a *dynamic* network is to consider each time window as an independent network and use the original algorithm on each of them. Once all these independent community assignments are produced, one uses the method described in the previous subsection to identify communities in $w + 1$ as children of communities in w. This procedure is what we call the 'unrestricted' version of the algorithm because there is no restriction carried from the past on the algorithm as applied to window $w + 1$.

Aynaud and Guillaume (2010, pp. 510-11) have documented an undesirable property of the unrestricted algorithm: extreme instability to small changes in the network. In one simulation using a network with 9377 nodes, removing a single node at random leads to more than a thousand nodes being moved in the identified community structure. This outcome is clearly an artefact of the detection method: intuitively, a community should be a lot more resilient to the disappearance of one of its members.

In consequence of this instability, the unrestricted algorithm as applied to our evolving network of documents gives the picture of short-lived clusters. Panel (a) of Figure 7 shows the life lengths of the clusters detected by this version over the whole period considered (from time window 1899-1903 to 2009-2013). In

total, 640 clusters are differentiated, having an average life of 2.6 windows. According to this version, the longest lived cluster existed for 27 windows.[12] As one can see in Table 4, the clusters in the last period were pretty young.

Version of algorithm	Age of clusters
Unrestricted	27, 23, 16, 13, 5, 5, 5, 1, 1, 1
Restricted	69, 63, 58, 56, 54, 24, 23
Low-level	52, 45, 42, 28, 16, 12, 11, 8

Table 4: *Age of clusters in last time window*

There is a need for a greater temporal stability of the cluster assignments. Aynaud and Guillaume (2010, pp. 510-11; see also Aynaud, 2011, sec. 3.5) propose to add temporal stability to the Louvain method by including information from the past in the initial community assignment of the current window. Remember that, in the original Louvain method, each node starts in a distinct community. What we will call the restricted version of the algorithm is different only at this point: for all time windows except the first one, the initial community assignment is the final assignment from the previous window plus each *new* node being in a single-node community. If node d_i and d_j were assigned to the same community in window w and they still exist in $w+1$, they start in the same community in the latter window. If node d_{new} was not in w, it starts alone in a community. From this initial assignment, the algorithm runs as before: for each node, it is considered whether moving it to a different community improves modularity, and so forth. This procedure means that, although d_i and d_j start in the same community, they might end up in different communities. Yet, the fact that each time window starts with pre-existing clusters increases the temporal stability of the detected clusters.

Indeed, clusters in our economics network survive far longer with this restricted version than with the unrestricted version, as can be seen from Panel (b) of Figure 7. On average, the life length of a cluster is now 5.7 windows (instead of 2.6). The total number of clusters detected drops to 234 (instead of 640), 15 of which having life lengths above 15 windows and the longest lived having existed for 69 windows.

A potential issue with this restricted version is that, to achieve more stability, it gives away too much on the adequacy of the partition for each time window. As Figure 8 shows, the restricted algorithm does indeed worse than the unrestricted version with respect to the modularity of its assignments; its modularity scores are on average 0.02 below the scores of the unrestricted algorithm. How big is this difference? As a point of comparison, Blondel et al. (2008, Table 1) report modularity scores achieved by four algorithms on four networks of more than a thousand nodes. The original Louvain method outperforms the other algorithms by an average difference in modularity of 0.07. The restricted versions does quite better than this average, but a different version might be able to have an even smaller gap.

Another reason to refrain from adopting the restricted version is that, as Figure 9 shows, it systematically produces partitions with less communities than the unrestricted version, at least since the 1950s. This property might not be problematic; the temporal instability of the unrestricted algorithm might have as a collateral effect the identification of too many communities per window. Yet, it seems wise to devise a third version of the Louvain method to validate that the restriction does not give rise to significant artefacts. This third version is also meant to be a middle way between the extreme volatility of clusters given by the unrestricted algorithm and the, perhaps extreme, stability of clusters given by the restricted algorithm. This stability seems especially strong since the 1970s, as is obvious from Table 5. According to the unrestricted algorithm, the most recent period has been tumultuous, with 247 births and deaths of clusters. In sharp contrast, the restricted version counts only 11 such events – from this perspective, economics seems a long quiet river.

The third version of the algorithm, that we call the 'low-level restriction algorithm' or 'low-level' for short, also initializes the algorithm for each window (except the first) with information from the community

[12] What was this cluster? It lived from window 1984-1988 to 2010-2014. The reader can guess its identity by its 4 characteristic stemmed keywords: `wage, labor, school, estimation`. See the main document for a description of how these keywords are arrived at.

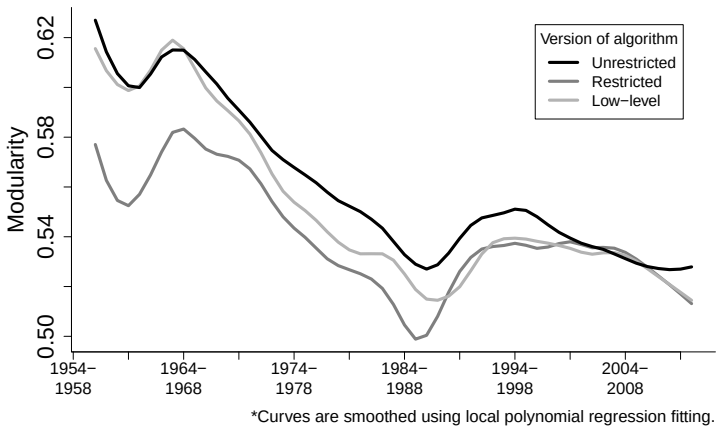


Figure 8: *Evolution of the modularity of economics*

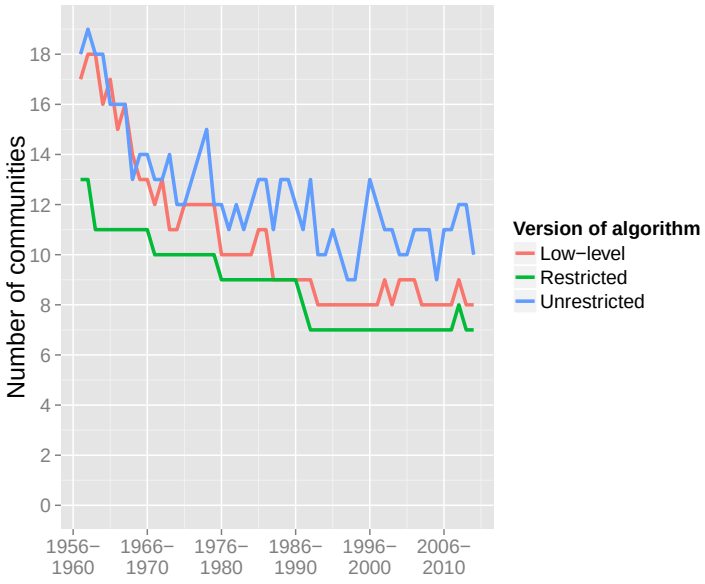


Figure 9: *Comparison of the number of clusters selected by each algorithm*

Version	Births	Deaths
Unrestricted	124	123
Restricted	5	6
Low-level	22	25

Table 5: *Events detected by each version of the algorithm since window 1970-1974.*

assignment of the previous window. The difference with the restricted version is that it does not use the *final* assignment, but the assignment given by the first iteration of the algorithm. Remember that, in the first iteration, the Louvain method moves nodes one by one between communities until modularity cannot be improved by this single-node displacement. At the end of this first iteration, the method gives a community assignment. In the following iterations, whole communities considered as community-nodes are moved. Each iteration in which there has been some movements produces a distinct community assignment. If there has been multiple iterations, the final assignment contains necessarily *fewer* communities than the first assignment. In the restricted version, the final assignment of $w-1$ was used to initialize the algorithm for w. In the low-level version, the first assignment is used. This version thus starts each window with smaller, more fragmented communities. This initialization makes it easier for the set of new nodes in w to restructure the previous communities.

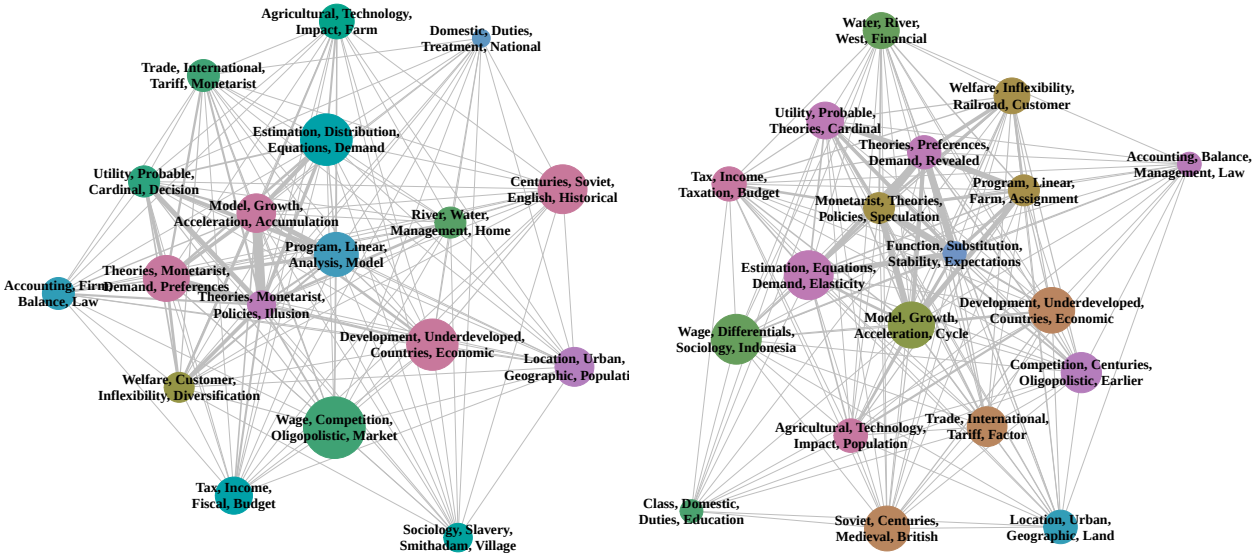
How does this low-level version of the Louvain method behave? For all the properties that we have looked at, it stands in between the unrestricted and the restricted versions. Panel (c) of Figure 6 shows that its distribution of γ_{ij} is less extreme than the two others: the frequency mass between 0.25 and 0.65 is smaller than for the unrestricted version and larger than the restricted version. For the whole period, it detects 465 clusters (between 234 and 640), with an average life of 3.3 windows (between 2 and 4.2). As can be seen from Panel (c) of Figure 7, the number of long-lived clusters (living more than 15 windows) is also in between the results from the other two algorithms (there are 10 such clusters, in between 6 and 15), and these long-lived clusters have an average life length of 32.8 windows (in between 20.2 and 37.2). The longest-lived cluster reaches an age of 52 windows (in between 27 and 69). At the end of the period considered, the life lengths of the clusters were also between the ones reported by the unrestricted and restricted algorithms (see Table 4).

What about the quality of the partitions produced by the low-level version? As shown in Figure 8, this version does better than the restricted version for almost all the period. Over the whole period, the average gap in modularity compared to the unrestricted version is closed by 40 %, from 0.020 to 0.012. The gap was already quite small compared with the reference class of other algorithms; it now becomes almost negligible.

Continuing the comparison between the low-level version and the other two, we see in Table 5 that it identifies 47 events (cluster births and deaths) since the 1970s (between the 11 and 247 events of the other versions). Finally, Figure 9 shows that, compared to the restricted version, it typically identifies a number of communities per window closer to the one of the unrestricted version. All in all, the low-level version outperforms the unrestricted version in terms of stability without almost not sacrificing, on average, modularity. It is the algorithm that we will be using.

But does the choice among these three algorithms really make a difference to the results? We already saw that it makes a huge difference to the temporal stability of the detected clusters (for instance, in the number of events reported, see Table 5). But what about the snapshots at any point in time? Are they vastly different depending on the algorithm?

Let us first casually compare the network representation at the beginning and at the end of the period for two of the three versions: the (original) unrestricted version and the low-level version (the version we will use). Figure 10 depicts the network in window 1956-1960 according to the two versions and Figure 11 does the same for window 2010-2014. The first striking shared property is the number of clusters: the network is highly fragmented in both versions at the beginning of the period and is made of a handful of clusters at the end. This positively correlated movement in the number of clusters could already be seen in Figure 9. At the end of our period, the original version of the algorithm outputs 10 clusters while the low-level version gives 8 clusters. If one takes the time to look more closely at the clusters, it will also appear that many



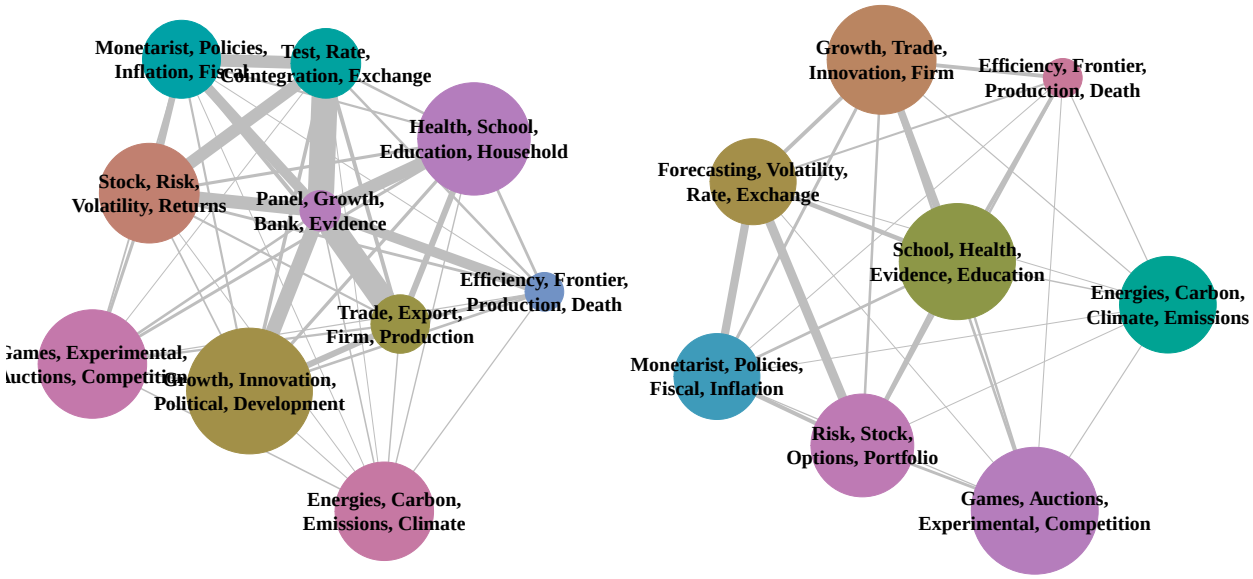
(a) *Original algorithm (unrestricted).* **(b)** *Selected algorithm (low-level).*

Figure 10: *Network of economics in the first time window (1956-1960) according to two versions of the algorithm.*

keywords that are grouped together in one version are grouped alike in the other version, and this holds for the beginning and the end of the period.

Focusing on the last window, we can look more carefully at the overlap of the clusters detected by the two algorithms. The table in panel 11c gives the fraction of the documents in each cluster detected by the unrestricted version (the rows) that are allocated to each cluster detected by the low-level version (the columns). What we see is that the members of almost all clusters from the original algorithm are highly concentrated in one cluster detected by the low-level version. Only the cluster with ID 638 does not have a sister cluster with a ratio greater than 0.75. This cluster has keywords `Growth, Innovation, Political, Development` in window 2010-2014. Around a third of its documents goes to a cluster (ID 401) with keywords `Games, Auctions, Experimental, Competition`, and around a half goes to a cluster (ID 442) with keywords `Growth, Trade, Innovation, Firm`. If we look at the columns, we see that this last column-cluster integrates a big fraction of another row-cluster (ID 640), which has keywords `Trade, Export, Firm, Production`. There is also an another column-cluster that integrates two row-clusters: it is the one with ID 454 (keywords: `School, Health, Evidence, Education`) which combines row-cluster ID 562 (keywords: `Health, School, Education, Household`) with row-cluster ID 631 (keywords: `Panel, Growth, Bank, Evidence`).

Finally, we can use a measure of the similarity of partitions called normalized mutual information (NMI) (Ana and Jain, 2003) to systematically track similarity of partitions through time. Denote by $\mathbf{D}$ the set of documents that are partitioned and denote by $\mathbf{C_1}$ and $\mathbf{C_2}$ the set of communities given by two methods of partitioning $\mathbf{D}$. A matrix N can be constructed where each row corresponds to a community in $\mathbf{C_1}$ and each column to a community in $\mathbf{C_2}$. The element n_{ij} of this matrix is the number of documents in $\mathbf{D}$ appearing *both* in community i of $\mathbf{C_1}$ and j of $\mathbf{C_2}$. Denote by $n_{i.}$ the sum of all the elements in row i of matrix N – that is, $n_{i.}$ gives the number of documents in $\mathbf{D}$ that are in community i according to the first method. Define $n_{.j}$ in a similar fashion (for the column). Finally, denote the sum of all the elements of N by n – i.e.,



(a) *Original algorithm (unrestricted).* **(b)** *Selected algorithm (low-level).*

	-401-	-437-	-442-	-454-	-459-	-462-	-463-	-464-
-562-	0.1			0.9				
-576-		1.0						
-601-					0.9			
-612-		0.1				0.9		
-627-	0.8							0.2
-628-		0.2						0.8
-631-				0.9				
-638-	0.3		0.5				0.1	
-639-						0.1	0.8	
-640-	0.1		0.8				0.1	

(c) *Fraction of documents being in the row-cluster according to the original algorithm and being in the column-cluster according to the selected algorithm.*

Figure 11: *Network of economics in the last time window (2010-2014) according to two versions of the algorithm*

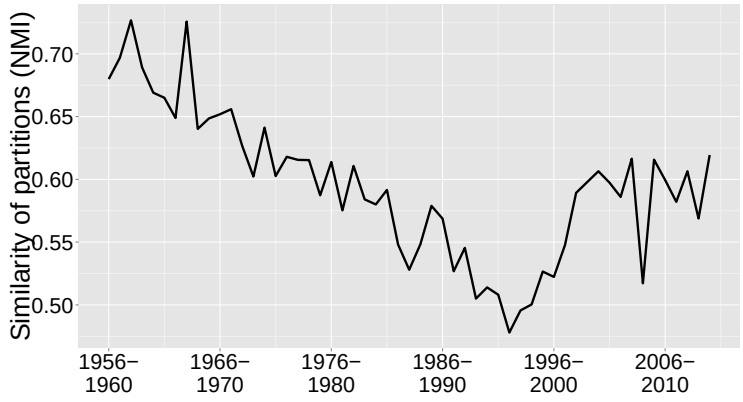


Figure 12: *Similarity of the community partitions from the original Louvain method and the selected algorithm.*

the cardinality of $\mathbf{D}$. The measure of normalized mutual information between $\mathbf{C_1}$ and $\mathbf{C_2}$ is:

$$NMI(\mathbf{C_1}, \mathbf{C_2}) = \frac{-2 \sum_{i \in \mathbf{C_1}} \sum_{j \in \mathbf{C_2}} n_{ij} \ln \left(\frac{n_{ij} n}{n_{i\cdot} n_{\cdot j}} \right)}{\sum_{i \in \mathbf{C_1}} n_{i\cdot} \ln \left(\frac{n_{i\cdot}}{n} \right) + \sum_{j \in \mathbf{C_2}} n_{\cdot j} \ln \left(\frac{n_{\cdot j}}{n} \right)}$$

This measure gives an indication of how predictable one community allocation is given the other allocation. When it approaches 1, the partitions are more similar.

Figure 12 depicts the evolution of this measure when we are comparing the partitions from the original algorithm and the algorithm that we have selected. We find that the two community partitions were most similar in the earliest time windows, that similarity was trending downward up to the early 1990s and has been trending mildly upward since then. This is an comforting non-monotonic pattern: though our chosen algorithm is initialized with restrictions for each new window, it does not systematically diverge more and more from a partition made without partition from the past.

Nevertheless, we must conclude that the choice of the algorithm has an impact on what economics looks like in each window, although it does not transfigure clusters to the point of making them unrecognizable. Since the version of the Louvain method that we have devised performs well with respect to both modularity and temporal stability of clusters, it is advisable to stick to it in our main analysis without trying to systematically build bridges to the results of the other, non-optimized versions.

2.2.1 Main output under alternative specifications

Operationalizing a general method inevitably involves the introduction of assumptions that might be questioned. In the main article and this appendix, we have attempted to justify our choices. Some of them clearly impact the results: for instance, choosing the version of the algorithm greatly influences the temporal stability of the uncovered specialty structure and mildly influences the specialty structure at any point in time. For such assumptions, we have argued that the option we use is more reasonable than alternatives. Other choices, though they need to be made, might not matter much for the result. In this section, we look at two assumptions that have been questions by readers of an early version of the paper and argue that changing these assumptions in the direction indicated by these readers do not greatly impact the results.

The first assumption is about the length of our time windows. We chose 5 years, but we accept that there is no strong argument to stop at exactly this length. To check whether our results are robust to a change in this assumptions, we chose a length that is at one extreme of what we deem reasonable: 3 years.[13] We

[13] If time windows were even shorter, that would mean that more than half of the documents in window $w+1$ would not be in window w. A window of 3 years was also the length suggested by a referee.

reran the algorithm without changing any of the other assumptions.

Figure 13 is our main representation of the results that can be compared to the analogous representation in the main article or the web platform. As one would expect, the specialty structure is more volatile if we use 3-year windows: we have 150 clusters over the period compared to 101 with our favored specification. This volatility seems to be mainly noise: some successions of specialties are interrupted and start as new successions two to three time windows later. For instance, what one might call public finance jumps from row 7 to row 8 and a specialty focusing on labor markets moves from row 9 to row 7. These jumps make the graph harder to read, but the overall specialty structure is highly similar. In particular, the two versions attain analogous partitions at the end of the period. Table 6 reports how the documents of the last time window of the 3-year specification are distributed in the last time window of the 5-year specification: the two specifications give 8 clusters each and each cluster has a clear sister cluster in the other specification. Since the results of the two specifications are highly similar, but the 3-year specification is harder to interpret because of excess volatility, we conclude that it is reasonable to stay with our original specification where time windows have a length of 5 years.

	-401-	-437-	-442-	-454-	-459-	-462-	-463-	-464-
-554-			0.8	0.2				
-576-		0.1				0.9		
-589-		1.0						
-590-				0.8				
-596-					0.8		0.2	
-602-		0.2						0.8
-603-	0.6		0.1			0.1		0.1
-604-	0.1						0.9	

Table 6: *Fraction of documents being in the row-cluster in window 2012-2014 of the 3-year specification that are in the column-cluster of window 2010-2014 of the 5-year specification*

The second assumption that has attracted criticism is our decision to consider as clusters, in each time window, only communities that have at least 1 % of the documents. A referee rightly noted that this restriction means something quite different in absolute terms at the beginning and at the end of the period. For window 1956-1960, 1 % of the discipline represents 62 documents; for window 2010-2014, it represents 954 documents. To test whether this restriction significantly distorts the overall representation of economics, we used an alternative (weaker) restriction: a community must count at least 10 documents to be registered as a cluster.

As can be seen in Figure 14, the main consequence of changing this restriction is a more cluttered representation: there are numerous extremely small clusters that cohabit with the main clusters that we have in our main specification. Furthermore, these small clusters are usually short-lived (average life is 2.1 time windows and the longest-lived clusters survive 9 windows) and do not form lasting successions of specialties. In short, we do not detect much that is worth noting in a history of the *macro*dynamics of economics below our original threshold. Since having numerous small clusters can make it harder for the reader to see the big picture, we conclude that it is reasonable to keep our original threshold.

2.3 Further information on keyword retrieval

Define $\mathbf{T}$ to be a set of preprocessed words, which can include many times the same stem s. The subset $\mathbf{w}_s \subset \mathbf{T}$ is made of all the elements of $\mathbf{T}$ that are stem s. The frequency of s in $\mathbf{T}$, denoted $f_{\mathbf{T}}(s)$, is simply the number of elements of $\mathbf{w}_s$.

Taking any subset $\mathbf{T}'$ of some $\mathbf{T}$, we can identify what stems distinguish most this subset from the full set by using the log-likelihood measure proposed by Rayson and Garside (2000):

$$LL(s, \mathbf{T}', \mathbf{T}) = f_{\mathbf{T}'}(s) * \ln\left(\frac{f_{\mathbf{T}'}(s)}{\mathrm{E}_{\mathbf{T}'}(s)}\right) + f_{\mathbf{T}}(s) * \ln\left(\frac{f_{\mathbf{T}}(s)}{\mathrm{E}_{\mathbf{T}}(s)}\right) \quad (5)$$

Figure 13: *Specialties through time in economics with time* ***windows of 3 years*** *(instead of 5). Only clusters surviving at least two windows are shown. Height represents relative share of documents per window, connected clusters are either related through merge or split. Stemmed keywords for each cluster are the most frequent keywords associated with this cluster during its lifetime. For clusters surviving at least five windows, the four most frequent keywords are shown; for shorter-lived clusters, only the most frequent keyword is used. Row identifiers are shown on both sides of the image to facilitate referencing.*

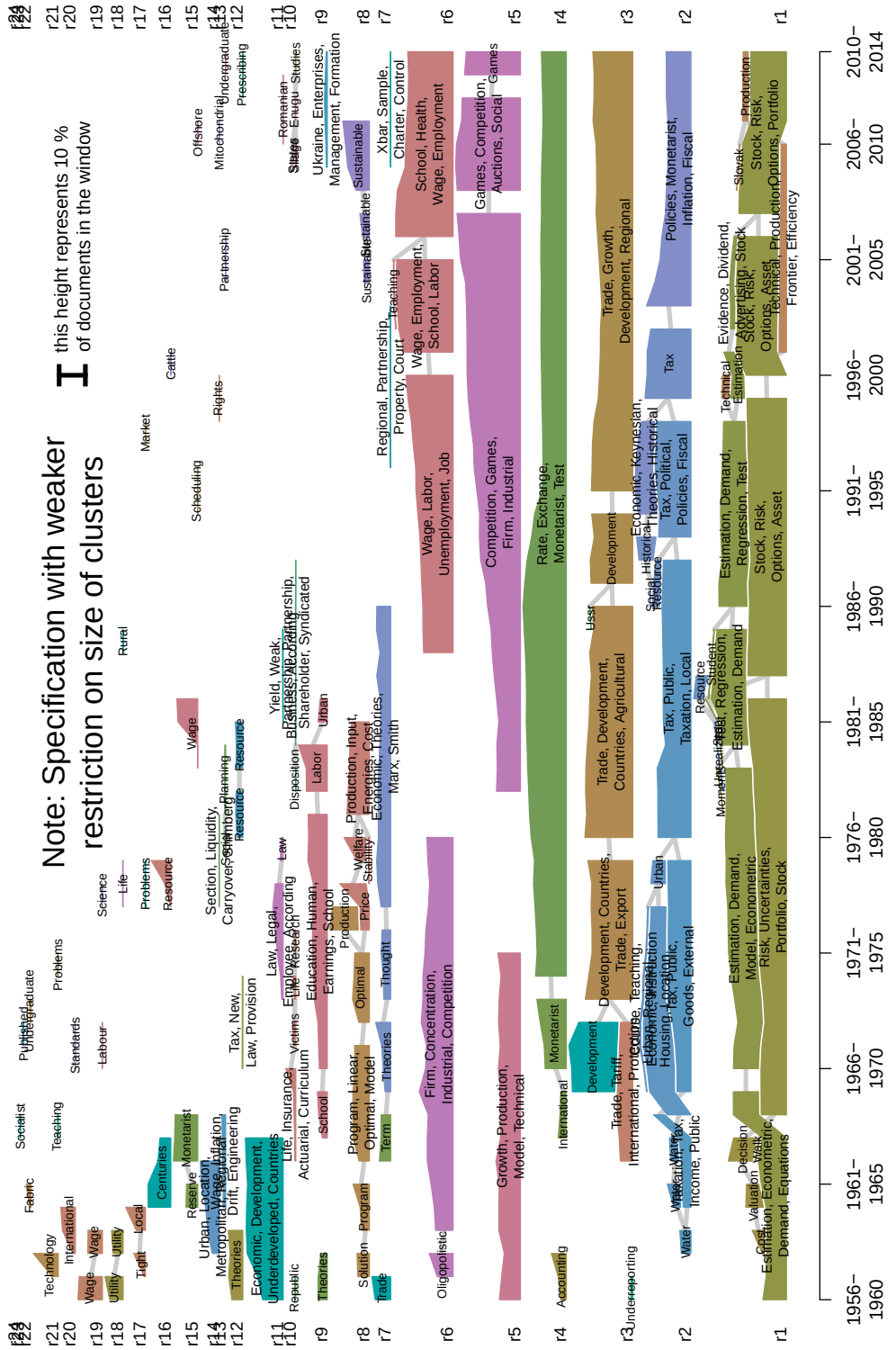


Figure 14: *Specialties through time in economics with all communities of* ***at least 10 documents*** *being counted as clusters. Only clusters surviving at least two windows are shown. Height represents relative share of documents per window, connected clusters are either related through merge or split. Stemmed keywords for each cluster are the most frequent keywords associated with this cluster during its lifetime. For clusters surviving at least five windows, the four most frequent keywords are shown; for shorter-lived clusters, only the most frequent keyword is used. Row identifiers are shown on both sides of the image to facilitate referencing.*

where $\mathrm{E}_{\mathbf{T}}(s)$ and $\mathrm{E}_{\mathbf{T}'}(s)$ are the expected number of occurences of stem s in $\mathbf{T}$ and $\mathbf{T}'$, given by:

$$\mathrm{E}_{\mathbf{T}}(s) = n_{\mathbf{T}} * \frac{f_{\mathbf{T}'}(s) + f_{\mathbf{T}}(s)}{n_{\mathbf{T}'} + n_{\mathbf{T}}}$$

and

$$\mathrm{E}_{\mathbf{T}'}(s) = n_{\mathbf{T}'} * \frac{f_{\mathbf{T}'}(s) + f_{\mathbf{T}}(s)}{n_{\mathbf{T}'} + n_{\mathbf{T}}}$$

The stems with the highest values of $LL(s, \mathbf{T}', \mathbf{T})$ are considered to be the ones most conspicuously differentiating $\mathbf{T}'$ from $\mathbf{T}$. There are two types of stems having a high value of LL. First, the relative frequency of a stem in $\mathbf{T}'$ might be far higher than in $\mathbf{T}$, leading to a high value of the first term in equation (5). Second, the opposite might hold: the stem might be far more frequent elsewhere, leading to a high value of the second term. This second type of stems identifies the subset $\mathbf{T}'$ by pointing to what it is not about. The first type is more straightforward. We only focus on this type of stems i by restricting the keywords to the ones with $f_{\mathbf{T}'}(s) > \mathrm{E}_{\mathbf{T}'}(s)$. Unless otherwise indicated, $\mathbf{T}'$ is the set of title words in one cluster at one time window and $\mathbf{T}$ is the set of all title words at this same time window.

Appendices

A Journals included in this study

Here is the complete list of the 549 journals included in this study and selected with the procedure described in subsection 1.1. The 387 titles in **bold** are assigned to the discipline 'economics' by the NSF classification; the 124 titles in *italics* are kept as mainly-economics journals though they are in a different discipline according to the NSF classification. Finally, the 38 journals in gray are partly-economics journals.

Acta Oeconomica, **Actual Problems of Economics**, **Actualite Economique**, **Advances in Econometrics**, **Advances in Econometrics : A Research Annual**, **African Development Review-Revue Africaine de Developpement**, *African Economic History*, **Agrekon**, *Agribusiness*, *Agricultural Administration*, *Agricultural Administration and Extension*, **Agricultural Economics**, **Agricultural Economics Research**, **Agricultural Economics-zemedelska Ekonomika**, Agricultural Science in Finland, Agricultural and Food Science in Finland, **Ajia Keizai-journal of The Institute of Developing Economies-**, *Akron Business and Economic Review*, **American Economic Journal-applied Economics**, **American Economic Journal-economic Policy**, **American Economic Journal-macroeconomics**, **American Economic Journal-microeconomics**, **American Economic Review**, **American Economist**, **American Journal of Agricultural Economics**, **American Journal of Economics and Sociology**, **American Law and Economics Review**, **Amfiteatru Economic**, **Annales D Economie Et de Sociologie Rurales**, **Annales de Sciences Economiques Appliquees**, **Annales-economies Societes**, **Annals of Economic and Social Measurement**, **Annals of Economics and Finance**, **Annals of Public and Co-operative Economy**, Annals of Regional Science, Annals of The New York Academy of Sciences, **Annual Review of Economics**, **Annual Review of Financial Economics**, **Annual Review of Resource Economics**, **Applied Economic Perspectives and Policy**, **Applied Economics**, **Applied Economics Letters**, *Aquaculture Economics & Management*, **Archiv Fur Soziale Gesetzgebung Und Statistik**, **Argumenta Oeconomica**, **Asia-pacific Journal of Accounting & Economics**, **Asian Economic Journal**, **Asian Economic Papers**, **Asian Economic Policy Review**, *Asian Journal of Technology Innovation*, **Asian-pacific Economic Literature**, Astin Bulletin, *Atlanta Economic Review*, *Australian Economic History Review*, **Australian Economic Papers**, **Australian Economic Review**, **Australian Journal of Agricultural Economics**, **Australian Journal of Agricultural and Resource Economics**, **B E Journal of Economic Analysis & Policy**, **B E Journal of Macroeconomics**, **B E Journal of Theoretical Economics**, **Baltic Journal of Economics**, **Bell Journal of Economics**, **Bell Journal of Economics and Management Science**, *Berichte Uber Landwirtschaft*, *Betrieb*, *Betriebswirtschaftliche Forschung Und Praxis*, **British Tax Review**, **Brookings Papers on Economic Activity**, **Bulletin for International Fiscal Documentation**, **Bulletin of Economic Research**, *Bulletin of Indonesian Economic Studies*, Bulletin of Rural Economics and Sociology, **Bulletin of The Oxford University Institute of Economic and Statistics**, **Cahiers Economiques de Bruxelles**, **Cambridge Journal of Economics**, *Cambridge Journal of Regions Economy and Society*, *Canadian Farm Economics*, **Canadian Journal of Agricultural Economics-revue Canadienne D Economie Rurale**, **Canadian Journal of Economics**, **Canadian Journal of Economics & Political Science**, **Canadian Journal of Economics-revue Canadienne D Economique**, **Cato Journal**, **Center for Settlement Studies. Research Reports. University of Manitoba**, **Cepal Review**, **Cesifo Economic Studies**, **China Agricultural Economic Review**, **China Economic Review**, **China & World Economy**, *Chinese Economic Studies*, **Cliometrica**, **Communist Economies & Economic Transformation**, *Computational Economics*, **Contemporary Economic Policy**, **Contemporary Policy Issues**, **Contributions To Political Economy**, *Cuadernos de Economia Y Direccion de La Empresa*, **Custos E Agronegocio**, *Defence and Peace Economics*, **Desarrollo Economico**, **Desarrollo Economico-revista de Ciencias Sociales**, **Developing Economies**, **Development and Change**, **Dynamic Games and Applications**, *E & M Ekonomie A Management*, **Eastern African Economic Review**, *Eastern European Economics*, **Ecological Economics**, **Econ Journal Watch**, **Econometric Reviews**, **Econometric Theory**, **Econometrica**, **Econometrics Journal**, **Economia Chilena**, **Economia Mexicana-nueva Epoca**, **Economia Politica**, **Economic Bulletin for Europe**, **Economic Computation and Economic Cybernetics Studies and Research**, **Economic Development Quarterly**, **Economic Development and Cultural Change**, *Economic Geography*, *Economic History Review*, **Economic History Review-first Series**, **Economic Inquiry**, **Economic Journal**, **Economic Modelling**, **Economic Planning**, **Economic Policy**, **Economic Record**, **Economic Research-ekonomska Istrazivanja**, **Economic Systems**, **Economic Systems Research**, **Economic Theory**, **Economic and Business Bulletin**, **Economic and Labour Relations Review**, **Economic and Social Review**, **Economica**, **Economica-new Series**, **Economics Letters**, *Economics & Human Biology*,

Economics & Politics, **Economics and Philosophy**, **Economics of Education Review**, **Economics of Governance**, **Economics of Planning**, **Economics of Transition**, **Economics-the Open Access Open-assessment E-journal**, **Economie Appliquee**, Economie Et Medecine Animales, **Economist**, **Economist-netherlands**, *Economy and Society*, **Ekonomicko-matematicky Obzor**, **Ekonomicky Casopis**, *Ekonomiska Samfundets Tidskrift*, **Ekonomista**, **Ekonomska Istrazivanja-economic Research**, *Emerging Markets Finance and Trade*, **Emerging Markets Review**, **Empirica**, **Empirical Economics**, **Energy Economics**, **Energy Efficiency**, **Energy Journal**, **Energy Policy**, **Energy Sources Part B-economics Planning and Policy**, *Engineering Costs and Production Economics*, *Engineering Economist*, *Engineering and Process Economics*, *Environment and Development Economics*, **Environmental & Resource Economics**, **Estudios de Economia**, *Eurasian Geography and Economics*, Europe-asia Studies, **European Economic Review**, *European Journal of Health Economics*, *European Journal of Law and Economics*, *European Journal of Political Economy*, **European Journal of The History of Economic Thought**, **European Review of Agricultural Economics**, *European Review of Economic History*, **Experimental Economics**, *Expert Review of Pharmacoeconomics & Outcomes Research*, *Explorations in Economic History*, **Explorations in Entrepreneurial History**, **Federal Reserve Bank of St Louis Review**, **Feminist Economics**, *Finance A Uver-czech Journal of Economics and Finance*, *Finanzarchiv*, **Fiscal Studies**, **Food Policy**, Forest Policy and Economics, Futures, **Games and Economic Behavior**, **Geneva Papers on Risk and Insurance Theory**, **Geneva Papers on Risk and Insurance-issues and Practice**, **Geneva Risk and Insurance Review**, *George Washington Journal of International Law and Economics*, **German Economic Review**, **German Journal of Agricultural Economics**, **Giornale Degli Economisti E Annali Di Economia**, **Global Economic Review**, **Hacienda Publica Espanola**, *Health Economics*, *Health Economics Policy and Law*, **Historical Materialism-research in Critical Marxist Theory**, *History of Economic Ideas*, **History of Political Economy**, **Hitotsubashi Journal of Economics**, Homme Et La Societe, **Housing Educators Journal**, **Housing Finance Review**, **Iktisat Isletme Ve Finans**, **Imf Economic Review**, **Imf Staff Papers**, Independent Review, Indian Economic and Social History Review, **Indian Journal of Economics**, **Industrial and Corporate Change**, **Industrialization and Productivity**, **Industry and Innovation**, **Information Economics and Policy**, *Insurance Mathematics & Economics*, **Inter-american Economic Affairs**, **International Development Review**, **International Economic Review**, *International Environmental Agreements-politics Law and Economics*, **International Finance**, **International Food and Agribusiness Management Review**, **International Journal of Economic Theory**, *International Journal of Finance & Economics*, *International Journal of Forecasting*, **International Journal of Game Theory**, *International Journal of Health Care Finance & Economics*, **International Journal of Industrial Organization**, **International Journal of Production Economics**, **International Journal of Social Economics**, **International Journal of Transport Economics**, *International Labour Review*, **International Monetary Fund Staff Papers**, **International Review of Economics & Finance**, **International Review of Law and Economics**, *International Tax and Public Finance*, **Investigacion Economica**, **Investigaciones Economicas**, **Inzinerine Ekonomika-engineering Economics**, *Iowa Agricultural and Home Economics Experiment Station Research Bulletin*, *Iowa Agricultural and Home Economics Experiment Station Special Report*, Irish Journal of Agricultural and Food Research, **Itea-informacion Tecnica Economica Agraria**, **Jahrbuch Fur Sozialwissenschaft**, **Jahrbucher Fur Nationalokonomie Und Statistik**, **Japan and The World Economy**, **Japanese Economic Review**, **Japanese Economic Studies**, **Japanese Economy**, *Jcms-journal of Common Market Studies*, *Journal of Accounting & Economics*, **Journal of African Economies**, *Journal of Agrarian Change*, **Journal of Agricultural Economics**, **Journal of Agricultural Economics Research**, **Journal of Agricultural and Resource Economics**, **Journal of Applied Econometrics**, **Journal of Applied Economics**, *Journal of Australian Political Economy*, *Journal of Banking & Finance*, *Journal of Behavioral Finance*, **Journal of Behavioral and Experimental Economics**, **Journal of Business Economics and Management**, **Journal of Business & Economic Statistics**, **Journal of Choice Modelling**, **Journal of Comparative Economics**, **Journal of Competition Law & Economics**, Journal of Consumer Affairs, **Journal of Cultural Economics**, **Journal of Developing Areas**, **Journal of Development Economics**, **Journal of Development Studies**, **Journal of Econometrics**, **Journal of Economic Behavior & Organization**, **Journal of Economic Dynamics & Control**, **Journal of Economic Education**, **Journal of Economic Geography**, **Journal of Economic Growth**, *Journal of Economic History*, **Journal of Economic Inequality**, **Journal of Economic Interaction and Coordination**, **Journal of Economic Issues**, **Journal of Economic Literature**, **Journal of Economic Perspectives**, *Journal of Economic Policy Reform*, *Journal of Economic Psychology*, **Journal of Economic Studies**, **Journal of Economic Surveys**, **Journal of Economic Theory**, Journal of Economic and Social Measurement, **Journal of Economics & Management Strategy**, *Journal of Economics and Business*, **Journal of Economics-zeitschrift Fur Nationalokonomie**, **Journal of Empirical Finance**, **Journal of Energy in Southern**

Africa, **Journal of Environmental Economics and Management**, **Journal of Environmental Management**, **Journal of Evolutionary Economics**, **Journal of Farm Economics**, *Journal of Finance*, **Journal of Financial Econometrics**, *Journal of Financial Economics*, **Journal of Financial Stability**, *Journal of Financial and Quantitative Analysis*, *Journal of Forecasting*, **Journal of Forest Economics**, *Journal of Health Economics*, **Journal of Home Economics**, *Journal of Home Economics Research*, *Journal of Housing Economics*, **Journal of Human Capital**, **Journal of Human Development and Capabilities**, **Journal of Human Resources**, **Journal of Industrial Economics**, **Journal of Institutional Economics**, **Journal of Institutional and Theoretical Economics-zeitschrift Fur Die Gesamte Staatswissenschaft**, **Journal of International Development**, *Journal of International Economic Law*, **Journal of International Economics**, **Journal of International Financial Markets Institutions & Money**, *Journal of International Law and Economics*, **Journal of International Money and Finance**, **Journal of International Trade & Economic Development**, **Journal of Korea Trade**, **Journal of Labor Economics**, **Journal of Land and Public Utility Economics**, **Journal of Law Economics & Organization**, **Journal of Law & Economics**, **Journal of Macroeconomics**, **Journal of Mathematical Economics**, *Journal of Media Economics*, Journal of Mental Health Policy and Economics, **Journal of Monetary Economics**, **Journal of Money Credit and Banking**, **Journal of Neuroscience Psychology and Economics**, **Journal of Pension Economics & Finance**, *Journal of Policy Analysis and Management*, **Journal of Policy Modeling**, **Journal of Political Economy**, **Journal of Population Economics**, **Journal of Post Keynesian Economics**, *Journal of Productivity Analysis*, **Journal of Public Economic Theory**, **Journal of Public Economics**, **Journal of Real Estate Finance and Economics**, *Journal of Real Estate Research*, Journal of Regional Science, **Journal of Regulatory Economics**, *Journal of Risk and Insurance*, **Journal of Risk and Uncertainty**, **Journal of Rural Economics and Development**, Journal of Social Political and Economic Studies, **Journal of Socio-economics**, **Journal of Sports Economics**, *Journal of Taxation*, **Journal of The American Institute of Planners**, **Journal of The American Real Estate and Urban Economics Association**, **Journal of The Asia Pacific Economy**, Journal of The Economic and Social History of The Orient, **Journal of The European Economic Association**, *Journal of The History of Economic Thought*, **Journal of The Japanese and International Economies**, *Journal of Transport Economics and Policy*, Journal of Transport Geography, *Journal of Urban Economics*, *Journal of World Trade*, **Keio Economic Studies**, *Kommunist*, **Korean Economic Review**, **Kyklos**, **Labour Economics**, **Land Economics**, Landbauforschung Volkenrode, **Latin American Economic Review**, **Lecture Notes in Economics and Mathematical Systems**, **Macroeconomic Dynamics**, **Malayan Economic Review**, *Managerial and Decision Economics*, **Manchester School**, **Manchester School of Economic and Social Studies**, *Marine Resource Economics*, Maritime Economics & Logistics, *Matekon*, *Mathematical Finance*, **Mathematical Social Sciences**, Metalworking Economics, **Metroeconomica**, *Mississippi Valley Journal of Business and Economics*, **Mondes En Developpement**, **Moorgate and Wall Street**, **National Tax Journal**, **National Westminster Bank Quarterly Review**, **Nationalokonomisk Tidsskrift**, **Nber General Series-national Bureau of Economic Research**, **Nber Macroeconomics Annual**, **Nber Occasional Papers-national Bureau of Economic Research**, **Nebraska Journal of Economics and Business**, *Networks & Spatial Economics*, **New England Economic Review**, **New Political Economy**, **New Telecom Quarterly**, **New Zealand Economic Papers**, **North American Journal of Economics and Finance**, **Occasional Papers Rural Development Committee and South Asia Program-cornell University**, Occasional Papers in Economic and Social History-university of Hull, *Oklahoma Current Farm Economics*, **Open Economies Review**, **Oxford Bulletin of Economics and Statistics**, **Oxford Economic Papers-new Series**, **Oxford Review of Economic Policy**, **Pacific Economic Bulletin**, **Pacific Economic Review**, **Panoeconomicus**, Papers in Regional Science, *Pharmacoeconomics*, **Politicka Ekonomie**, Politics Philosophy & Economics, **Portuguese Economic Journal**, **Post-communist Economies**, Post-soviet Affairs, *Post-soviet Geography and Economics*, **Prague Economic Papers**, **Problemas Del Desarrollo**, Problems of Communism, *Problems of Economic Transition*, *Problems of Economics*, **Psychologie V Ekonomicke Praxi**, **Public Choice**, **Public Finance Quarterly**, **Public Finance Review**, **Public Finance-finances Publiques**, *Qme-quantitative Marketing and Economics*, **Quantitative Economics**, *Quantitative Finance*, *Quarterly Journal of Business and Economics*, **Quarterly Journal of Economics**, *Quarterly Review of Economics and Business*, *Quarterly Review of Economics and Finance*, **Rand Journal of Economics**, **Real Estate Economics**, **Recherches Economiques de Louvain-louvain Economic Review**, *Regional Science and Urban Economics*, Regional Studies, **Resource and Energy Economics**, **Resources and Energy**, *Review of Agricultural Economics*, *Review of Black Political Economy*, *Review of Business and Economic Research*, *Review of Derivatives Research*, *Review of Development Economics*, *Review of Economic Conditions in Italy*, **Review of Economic Design**, **Review of Economic Dynamics**, **Review of Economic Statistics**, **Review of Economic Studies**, **Review of Economics and Statistics**, **Review of Economics of The Household**, **Review of Environmental Economics**

and Policy, **Review of Finance**, *Review of Financial Studies*, **Review of Income and Wealth**, **Review of Industrial Organization**, **Review of International Economics**, **Review of International Organizations**, **Review of International Political Economy**, **Review of Keynesian Economics**, **Review of Network Economics**, **Review of Radical Political Economics**, **Review of Social Economy**, *Review of World Economics*, **Revista Brasileira de Economia**, Revista Usem, Revista de Ciencias Sociales, **Revista de Economia Aplicada**, **Revista de Economia Mundial**, *Revista de Historia Economica*, *Revista de Historia Industrial*, **Revue Canadienne D Etudes Du Developpement-canadian Journal of Development Studies**, **Revue D Economie Politique**, Revue D Etudes Comparatives Est-ouest, **Revue Economique**, **Revue de L Est**, *Rivista Di Economia Agraria*, *Rivista Di Politica Economica*, **Rivista Internazionale Di Scienze Economiche E Commerciali**, **Romanian Journal of Economic Forecasting**, **Scandinavian Journal of Economics**, **Scottish Journal of Political Economy**, **Series-journal of The Spanish Economic Association**, **Singapore Economic Review**, **Skandinavisk Aktuarietidskrift**, **Small Business Economics**, **Social Choice and Welfare**, *Social Science & Medicine Part C-medical Economics*, **Social and Economic Administration**, **Social and Economic Studies**, *Socio-economic Planning Sciences*, **Socio-economic Review**, **South African Journal of Economic and Management Sciences**, **South African Journal of Economics**, **Southern Economic Journal**, *Soviet Economy*, Soviet Studies, **Spanish Economic Review**, **Spatial Economic Analysis**, **Staff Papers Brookings Institution**, **Structural Change and Economic Dynamics**, **Studies in Nonlinear Dynamics and Econometrics**, **Survey Research Methods**, **Swedish Journal of Economics**, **Technological and Economic Development of Economy**, **Theoretical Economics**, **Theory and Decision**, *Tijdschrift Voor Economische En Sociale Geografie*, *Tourism Economics*, **Transformations in Business & Economics**, Transport Policy, Transportation Research Part A-policy and Practice, Transportation Research Part B-methodological, Transportation Research Part E-logistics and Transportation Review, **Trimestre Economico**, **Utilities Policy**, *Value in Health*, **Veterinary Economics**, *Virginia Agricultural Economics*, *Virginia Polytechnic Institute Extension Service-virginia Agricultural Economics*, **Weltwirtschaftliches Archiv-review of World Economics**, **Western Economic Journal**, **Western Journal of Agricultural Economics**, *Work Employment and Society*, **World Bank Economic Review**, **World Bank Research Observer**, **World Development**, **World Economy**, **World Trade Review**, **Yale Economic Essays**, **Zbornik Radova Ekonomskog Fakulteta U Rijeci-proceedings of Rijeka Faculty of Economics**, **Zeitschrift Fur Betriebswirtschaft**, **Zeitschrift Fur Nationalokonomie**, **Zeitschrift Fur Nationalokonomie-journal of Economics**, *Zeitschrift Fur Wirtschafts-und Sozialwissenschaften*, *Zeitschrift Fur Wirtschaftsgeographie*

B Glimpses at the data on references

matching ID	Author	Year	Publication	Vol	Page
Null	*Null*	*Null*	ARCH SOMMAIRES ROUEN	*Null*	226
Null	*Null*	*Null*	EC J	14	47
Null	MITCHELL	*Null*	ORG LABOR	*Null*	CH20
Null	*Null*	1904	B NATL ASS WOOL DEC	*Null*	*Null*
Null	WRIGHTC	*Null*	FORUM	6	223
435141	*JOINTSELCOMMRA	1872	REP JOINT SEL COMM R	*Null*	R34
Null	DEROUSIERSM	*Null*	LABOR QUESTION	*Null*	127
Null	BUTTERWORTH	*Null*	TREATIES LAW RELATIN	*Null*	168
Null	WRIGHTC	*Null*	B I INT	8	110
Null	*Null*	*Null*	12 CENSUS	9	568
Null	*Null*	1898	P 7 ANN CONV	*Null*	19
Null	GRINLING	*Null*	BRIT RAILWAYS BUSINE	*Null*	161
Null	*Null*	*Null*	COMPTES RENDUS ECHEV	2	30
Null	*Null*	*Null*	CAHIERS ETATS GEN NO	2	396
Null	*Null*	*Null*	REG BUREAU	2	140

Table 7: *Random sample of the references in 1905*

matching ID	Author	Year	Publication	Vol	Page
1984597	HOLMES-M	2008	S E EUROPE J EC	1	9
744650	DOWNING-M	1996	J ENVIRON ECON MANAG	30	316
2246495	BURT-R	1992	STRUCTURAL HOLES SOC	*Null*	*Null*
1448606	VASSALOU-M	2004	J FINANC	59	831
454347	GROSSMAN-G	2001	SPECIAL INTEREST POL	*Null*	*Null*
1809530	LOMAS-K	2010	BUILD RES INF	38	1
2078157	KNIES-G	2008	J APPL SOC SCI STUD	128	75
1145317	ALSAMARRAI-S	1998	ECON EDUC REV	17	395
1460689	FOGEL-R	2004	ECON DEV CULT CHANGE	52	643
1824785	SEIFTER-A	2010	GEOSPATIAL HEALTH	4	135
1642982	OKAMOTO-D	2007	SOC SCI RES	36	1391
1498012	SCHWINDTBAYER-L	2005	J POLIT	67	407
449489	HENSHER-D	2005	APPL CHOICE ANAL PRI	*Null*	*Null*
1344877	JONES-C	2002	AM ECON REV	92	220
3329853	LEVY-A	2008	UTILITY VALUES HLTH	*Null*	*Null*

Table 8: *Random sample of the references in 2014*

References

Ana, L. and Jain, A. (2003), "Robust data clustering," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings*, vol. 2, pp. II–128–II–133 vol.2.

Aynaud, T. (2011), "Détection de communautés dans les réseaux dynamiques," Thèse de doctorat, Université Pierre et Marie Curie.

Aynaud, T. and Guillaume, J.-L. (2010), "Static community detection algorithms for evolving networks," in *2010 Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, pp. 513–519.

Becker, G. S. (1965), "A Theory of the Allocation of Time," *The Economic Journal*, 75, 493–517.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008), "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2008, P10008.

Csardi, G. and Nepusz, T. (2006), "The igraph software package for complex network research," *InterJournal*, Complex Systems, 1695.

Dowle, M., Short, T., Srinivasan, S. L. w. c. f. A., and Saporta, R. (2013), *data.table: Extension of data.frame for fast indexing, fast ordered joins, fast assignment, fast grouping and list columns.*, r package version 1.8.10.

Heckman, J. J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

National Science Board (2006), *Science and Engineering Indicators 2006*, Arlington, VA: National Science Foundations.

Nelson, R. R. and Winter, S. G. (1982), *An Evolutionary Theory of Economic Change*, Cambridge, Mass: Belknap Press of Harvard University Press.

Newman, M. E. J. (2004), "Analysis of weighted networks," *Physical Review E*, 70, 056131.

Newman, M. E. J. and Girvan, M. (2004), "Finding and evaluating community structure in networks," *Physical Review E*, 69, 026113.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

Rayson, P. and Garside, R. (2000), "Comparing Corpora Using Frequency Profiling," in *Proceedings of the Workshop on Comparing Corpora*, Stroudsburg, PA, USA: Association for Computational Linguistics, CompareCorpora '00, pp. 1–6.

Revolution Analytics and Weston, S. (2013a), *doParallel: Foreach parallel adaptor for the parallel package*, r package version 1.0.6.

— (2013b), *foreach: Foreach looping construct for R*, r package version 1.4.1.

Small, H. (1973), "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the American Society for Information Science*, 24, 265–269.

Testa, J. (2012), "The Thomson Reuters Journal Selection Process," Web of Knowledge Website, http://wokinfo.com/essays/journal-selection-process/.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press.

Wouters, P. (2006), "Aux origines de la scientométrie," *Actes de la recherche en sciences sociales*, 164, 11.

Xie, J., Kelley, S., and Szymanski, B. K. (2013), "Overlapping Community Detection in Networks: The State-of-the-art and Comparative Study," *ACM Comput. Surv.*, 45, 43:1–43:35.