Fine-grained classification of journal articles by relying on multiple layers of information through similarity network fusion: the case of the Cambridge Journal of Economics

Fine-grained classification of journal articles by relying on multiple layers of information through similarity network fusion: the case of the *Cambridge Journal of Economics*

Alberto Baccini[1] Lucio Barabesi[1] Martina Cioni[1] Eugenio Petrovich[2] Daria Pignalosa[1]

[1]Department of Economics and Statistics, University of Siena, Italy
[2]Department of Philosophy and Education Sciences, University of Torino, Italy and Department of Philosophy, Tilburg University, The Netherlands

Abstract

This paper explores the possibility of classifying journal articles by exploiting multiple information sources, instead of relying on only one information source at a time. In particular, the Similarity Network Fusion (SNF) technique is used to merge the different layers of information about articles when they are organized as a multiplex network. The method proposed is tested on a case study consisting of the articles published in the *Cambridge Journal of Economics*. The information about articles is organized in a two-layer multiplex where the first layer contains similarities among articles based on the full-text of articles, and the second layer contains similarities based on the cited references. The unsupervised similarity network fusion process combines the two layers by building a new single-layer network. Distance correlation and partial distance correlation indexes are then used for estimating the contribution of each layer of information to the determination of the structure of the fused network. A clustering algorithm is lastly applied to the fused network for obtaining a classification of articles. The classification obtained through SNF has been evaluated from an expert point of view, by inspecting whether it can be interpreted and labelled with reference to research programs and methodologies adopted in economics. Moreover, the classification obtained in the fused network is compared with the two classifications obtained when cited references and contents are considered separately. Overall, the classification obtained on the fused network appears to be fine-grained enough to represent the extreme heterogeneity characterizing the contributions published in the *Cambridge Journal of Economics*.

Keywords— Similarity network fusion; Generalized distance correlation; Partial distance correlation; Multilayer social networks; Communities in networks; Topic modeling.

Acknowledgments: The research is funded by the Italian Ministry of University, PRIN project: 2017MPXW98. We thank Alessandra Durio who contributed to the work by doing all the processing for the construction of the similarity matrices based on bags of words and topic modeling.

1 Introduction

The classification of scientific papers is a complex task accomplished either by experts or by using suitable algorithms. While expert classification is based on a thorough understanding of the metadata, content, and context of the papers to be classified, algorithmic classification usually relies on a single feature of the papers, such as content, keywords, or references/citations. For example, algorithmic classifications based on content use distant reading techniques to reveal groups of papers that share similar topics. Alternatively, algorithmic classifications relying on citations use citation relationships between papers to individuate clusters of papers with either overlapping bibliographies, as in bibliographic coupling, or that are frequently cited in the literature, as in co-citation. These clusters can be mapped, in turn, to research fields or subdisciplines.

This article explores a new method for algorithmic classification that relies on the integration of multiple features of papers, instead of one feature at a time. To this aim, we propose to adopt a technique called Similarity Network Fusion (SNF) [42], which is able to merge different layers of information about papers when they are organized as a multiplex network.

In this article, we focus in particular on two standard features of scientific papers: the full-text and the cited references. The method here proposed comprises three main steps. First, the similarities among articles are organized in a two-layer multiplex. The first layer conveys similarities based on full-text features, specifically, word frequencies, and topics extracted using the Latent Dirichlet Allocation technique [8]; the second layer similarities based on cited references, specifically, bibliographic coupling similarities. In the second step, SNF is used to synthesize the information available in the two separate layers into a new single-layer network, where the previous similarities are properly combined and fused. The fusion process is unsupervised and leverages the structural properties of each layer. By using distance correlation and partial distance correlation indexes, we can measure the contribution of each layer of information to the determination of the structure of the resulting network. In the third and last step, a clustering algorithm is applied to the fused network to obtain a classification of the papers. By comparing the results of clustering on the fused network with the results of clustering on each of the starting layers, it is possible to evaluate the coherence between the classification based on all the information available, and the classifications obtained by exploiting only one kind of information at a time.

This method is tested on a suitable case study, i.e., the papers published in the *Cambridge Journal of Economics* in the period 1985-2013. The *Cambridge Journal of Economics* is one of the leading non-mainstream economics journals. It was founded in the 1970s with the main purpose of providing a forum for post-Keynesian, Marxist, and Sraffian scholars, but hosts contributions from all other schools of heterodox thought, such as those in the institutionalist and evolutionary traditions. For a lively account of the events that led to the birth of the journal and to grasp the spirit that has characterized it since its inception, see [31]. Precisely because its distinguishing character lies in the plurality of approaches, the journal is not dedicated to a single field of economic analysis, but covers a multiplicity of topics ranging from microeconomics to macroeconomics and economic policy, with contributions of both theoretical and applied nature. The papers published in the *Cambridge Journal of Economics* are thus characterized by an extreme heterogeneity both from the viewpoint of their analytical approach and from the viewpoint of the topics covered. Hence the task of classifying its papers is particularly difficult since groups of papers may be defined not only in terms of different topics but also of approaches to the same topic. The choice of the *Cambridge Journal of Economics* is thus particularly challenging for testing the capacity of the new methodology proposed here to generate a meaningful and interpretable fine-grained classification of articles.

The paper is organized as follows. In Section 2 a short review of relevant literature is provided.

Section 3 describes the network data. Section 4 describes the workflow of the exploratory analysis. Section 5 defines the similarities among articles based on cited references and topics. In Section 6 the dependence between similarity matrices is discussed. In Section 7 the classifications of articles obtained in the two separate layers of cited references and topics are presented and compared. Section 8 describes the use and the results obtained by using SNF. Section 9 discusses and interprets the results. Section 10 concludes by suggesting further steps for the present research.

2 A short literature review

According to Glanzel and Schubert [20] "the classification of science into a disciplinary structure is at least as old as science itself". From Aristotle to Medieval logicians up to Nineteenth-century positivists, philosophers and scientists have proposed numerous classificatory schemes for organizing human knowledge [18]. In the Twentieth century, many concurring general classification schemes have become established, such as the Dewey classification, the OECD's fields of science, and Web of Science (WoS) or Scopus categories (for a short review see the entry "research fields" in [39]). These general schemes, however, "are too broad to adequately capture the more complex, fine-grained cognitive reality" [17]. Hence, uncountable attempts have been developed for classifying disciplines at the desired fine-grained level.

From a theoretical point of view, the delineation of scientific fields consists in partitioning the objects of the analysis into groups by using some classification technique. These techniques are divided by Zitt et al. [43] into three distinct groups: (i) ready-made or institutional classifications of science which originate from scientists or librarians and do not entail the use of bibliometrics; each "artifact" is classified by experts who assign it the correct label after a thorough evaluation based on its contents, context, and metadata; (ii) ex-post classifications where experts attach disciplinary or sub-disciplinary labels to clusters of "artifacts" obtained by network analysis techniques applied to relevant bibliometric or scientometric data; (iii) classifications that rely on highly supervised schemes functional to efficient information retrieval.

A stream of literature, recently reviewed in [6], is focused on the classification of journals. However, most of the literature considers the paper rather than the journal as the basic unit of scientific fields, and hence as the target of the classification. Similarities between pairs of papers have been defined in terms of contents by considering title and abstracts [11], keywords or text [2]; in terms of citation relations such as direct citations [33, 34] or co-citations [35] or bibliographic coupling [24]; or, finally, in terms of social attributes such as authorship [28]. Many works aim to compare scientific field classifications emerging from the use of different definitions of similarity relations among articles [10, 25, 26]. Sjögårde and Ahlgren [33] compare the choice of different parameters of modularity optimization for obtaining a fine-grained classification of disciplines.

By and large, the big part of the literature devoted to scientific fields delineation uses an approach based on a single-layer network, by exploiting only one type of information for classification or for the definition of a science map [30]. Zitt et al. [43] labeled as "hybridization" or "multinetwork approaches" those contributions that try to combine different layers of information for scientific fields delineation and classification purposes. A short review of the origin of these methods is available in [23], while a recent review is proposed by [43]. They highlighted that the use of multiple layers of information requires the heavy intervention of researchers for their integration. Indeed, for instance, Glänzel and Thijs [21] combined information obtained by bibliographic coupling and textual similarities to delineate clusters of articles and topics in astronomy and astrophysics. To this end, they built a weighted linear combination of the cosine similarities and studied the different results obtained by the adoption of different weights.

Previously, Colliander [15] had combined a measure of article similarity based on technical terms with one based on cited references, by using a complex weighted algorithm. The hybrid methodology proposed in [23] incorporates many assumptions about statistical distributions of data and the use of inverse-χ^2 method for integrating information about citation and text-based similarities.

The methodology proposed in this paper for integrating different layers of information about papers was applied in informetrics for the first time to the classification of scholarly journals. Specifically, Baccini et al. [6] used the similarity network fusion technique for classifying journals by considering a three-layer network: the first layer is generated by bibliograhic coupling among journals, the second considers the crossed presence of the same authors contributing to different journals, the third the crossed presence of the same scholars in the editorial boards of different journals. The technique of integration is purely exploratory and completely unsupervised; it does not require assumptions about the statistical distribution of data and about the weight to assign to different information during the integration process. Moreover, as will be illustrated later on, it is possible to measure ex-post the contribution of each layer to the structure of the fused final network.

The case-study developed here regards economics. Only in the last decade have economists developed interest in quantitative methods as tools to improve their understanding of the structure and evolution of their discipline. A special issue of the *Journal of Economic Methodology* is devoted to the "quantitative turn" in the history of economic thought [16]. Several contributions have employed network analysis tools, mostly by exploiting citation information, both in the form of bibliographic coupling and co-citation analysis. For instance, Claveau and Gingras [14] use bibliographic coupling to measure the cognitive similarity between articles in a corpus of over 400,000 documents retrieved from the Web of Science database and construct a dynamic network analysis that leads to identifying families of research fields and their evolution over time. They are thus able to identify the emergence and decline of subfields within economics since the late 1950s and reconstruct the history of specialties in the discipline. Truc, Claveau, and Santerre [40] rely on co-citation between articles published in the two main economic methodology journals over the past three decades in order to appraise the standard interpretation of the developments in the field. They generate three co-citation networks, one for each decade under consideration, and, by observing continuities and changes across the networks, they assess the main historical trends put forth in the existing interpretive literature. The first application of the LDA topic modeling technique to the economics discipline was carried out by [3]. They show how topic modeling can be used to construct a map of economics over time and detect key developments in the structure of the discipline. An attempt at combining citation and content information was carried out by García, Otero, and Salazar [19]. They analyze the developments in the field of consumption modeling over forty years by constructing co-citation networks and combining them with semantic evidence in the form of the most frequent strings of words used in the abstracts of co-citing articles. Developments in consumption modeling are detected by observing how clusters of co-citations and semantic changes co-evolved over the time period under consideration.

3 Data

Cited references data and textual data were retrieved from Web of Science (WoS) and JSTOR databases, respectively. JSTOR archival journal collection includes more than 2,800 academic journals across the humanities, social sciences, and natural sciences from 1,200 publishers from 57 countries. The time span of the analysis was determined by the data availability in the two databases: WoS started recording the *Cambridge Journal of Economics* in 1985, whereas JSTOR does not provide access to its most recent issues because of JSTOR policies. At the time of data retrieval (2019), the last complete available year

Table 1: Top-10 most cited documents in the *Cambridge Journal of Economics* (1985-2013)

Rank	Author	Year	Title	Citations
1	Keynes, J.M.	1936	The General Theory of Employment, Interest and Money	172
2	Marx, K.	1867 (vol I) 1885 (vol II) 1894 (vol III)	Das Kapital	133
3	Sraffa, P.	1960	Production of Commodities by Means of Commodities	115
4	Marshall, A.	1890	Principles of Economics	88
5	Lawson, T.	1997	Economics and Reality	86
6	Smith, A.	1776	An Inquiry into the Nature and Causes of the Wealth of Nations	70
7	Nelson, R.R. Winter, S.G.	1982	An Evolutionary Theory of Economic Change	63
8	Williamson, O.E.	1985	The Economic Institutions of Capitalism. Firms, Markets, Relational Contracting	62
9	Schumpeter, J.A.	1942	Capitalism, Socialism and Democracy	54
10	Kalecki, M.	1971	Selected Essays on the Dynamics of the Capitalist Economy 1933-1970	52

was 2013. Thus, the time span was set from 1985 to 2013.

WoS data, including the cited references of the records, were retrieved from WoS web platform, whereas the n-grams used in the topic modeling were retrieved from JSTOR Data for Research platform. In both cases, the query was based on the title of the target journal. The records in the two datasets were then matched using the volume, number, and name of the first authors. Any record without cited references or appearing in only one of the datasets was excluded, so that the final dataset included 1,344 records. Note that all types of documents published by the *Cambridge Journal of Economics*, not only research articles, were included.

To improve the reliability of citation analysis, cited references were cleaned using the CRExplorer software [38]. CRExplorer individuates and merges variants of the same reference through an algorithm based on string similarity. The process was humanly supervised to avoid wrong merging and individuate further variants to unify. Special attention was reserved for books and historical references, which are very common in this journal, in order to merge all of their possible variants. Of the 45,611 distinct cited references appearing in the raw dataset, 42,072 distinct cited references remained after this consolidation process (-7.8%). A significant number of cited references (84.7%) collects only 1 citation in the entire dataset. Table 1 shows the top-10 most common cited references in the dataset.

The 1,344 papers retrieved from JSTOR contained a total of 7,540,085 non-distinct words. This corpus of textual data was prepared in a suitable way for the topic modeling analysis using several Python scripts [7]. Pre-processing included the normalization of words with anomalous characters such as numbers or accents, the removal of too short and too long words, which stem from errors in the original documents, and the filtering out of stop-words. To scale down the size of the vocabulary, moreover, words were reduced to their root form (stems) through stemming, which was implemented with the **nltk** library. Lastly, rare stems that occurred in less than three papers in the corpus and common stems that occurred in more than 95% of the documents were removed. The processing of the textual data resulted thus in a vocabulary of 79,262 distinct stems distributed among 1,344 documents.

4 The workflow of the exploratory analysis

Figure 1 shows the workflow adopted for the analysis. As anticipated, the basic building blocks of the analysis are two different ways of representing similarity relations among articles, based respectively on cited references and contents (words and topics extracted with LDA). Thus, the first step of the analysis consists in defining the similarity networks among articles. A similarity network is obtained by considering cited references, while seven different networks are built by considering words and topics. The use of seven different similarity networks based on words and topics is due to the difficulty of satisfactorily define the number of topics, as will be explained later on.

The association among the structures of these similarity networks is computed for verifying the coherence of the information contained in them. The networks are then partitioned in communities or clusters of articles by using Louvain algorithm [9]. The clusters of papers obtained are compared by using suitable statistical techniques. If a weak association is detected among the structures of similarity networks based on cited references and on words and topics, and if the article clusters are weakly coherent, it can be concluded that information conveyed by cited-references and words and topics is different.

This conclusion may suggest the utility of integrating the different information available. This is done by applying the similarity network fusion technique. It consists in fusing the similarity network based on cited references, with each of the seven similarity networks based on words and topics. A suitable statistical technique is applied for measuring the contribution of the two layers to the structure of the fused network.

Hence, the association between the seven fused networks is computed for verifying the coherence among the network structures obtained. The Louvain algorithm is then applied to the fused networks for clustering articles. The clusters in the fused networks are compared among them and with the ones obtained with cited references and words and topics separately. The usefulness of the fusion will finally be validated by an expert interpretation of the classifications obtained.

5 Building similarity networks

As to the layer based on cited references, each paper is characterized by the set of its cited references, and the similarity between two articles is computed by considering their common cited references. The basic information is organized in a bipartite network, where the first set of nodes contains citing articles and the second set of nodes contains the cited references. Edges link each citing paper to its cited references. Similarities between each pair of citing articles are then computed by using the classical Jaccard similarity coefficient [22] and organized in a square similarity matrix. More precisely, if A and B represent the sets of cited references of two different papers, the Jaccard coefficient is defined as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \ , \tag{1}$$

where $|\cdot|$ denotes the cardinality of a set. It is apparent that $0 \leq J(A,B) \leq 1$. Hence, the similarity between two articles is proportional to the number of references cited by both papers: when two papers cite exactly the same set of references, i.e. when $A = B$, the maximum similarity $J(A,B) = 1$ occurs. In contrast, the minimum similarity $J(A,B) = 0$ is achieved when two papers have no common references, i.e. when $A \cap B = \emptyset$.

As to the layer of information based on contents, the basic idea is to measure similarity between articles in terms of similarity of contents. The adoption of a distant reading perspective is straightforward and it is here implemented in two different ways. The first one is very simple and does not require theoretical assumptions: each paper is characterized by its “Bags of Words” (hereafter BoW), i.e. the frequent distribution of lexical items used in it. Thus, the similarity between each pair of papers is computed by considering their common words. For implementing this approach, the basic information is organized in a bipartite network where the first set of nodes contains articles and the second set contains the words used in it. Edges link each article to its words with a weight proportional to the word

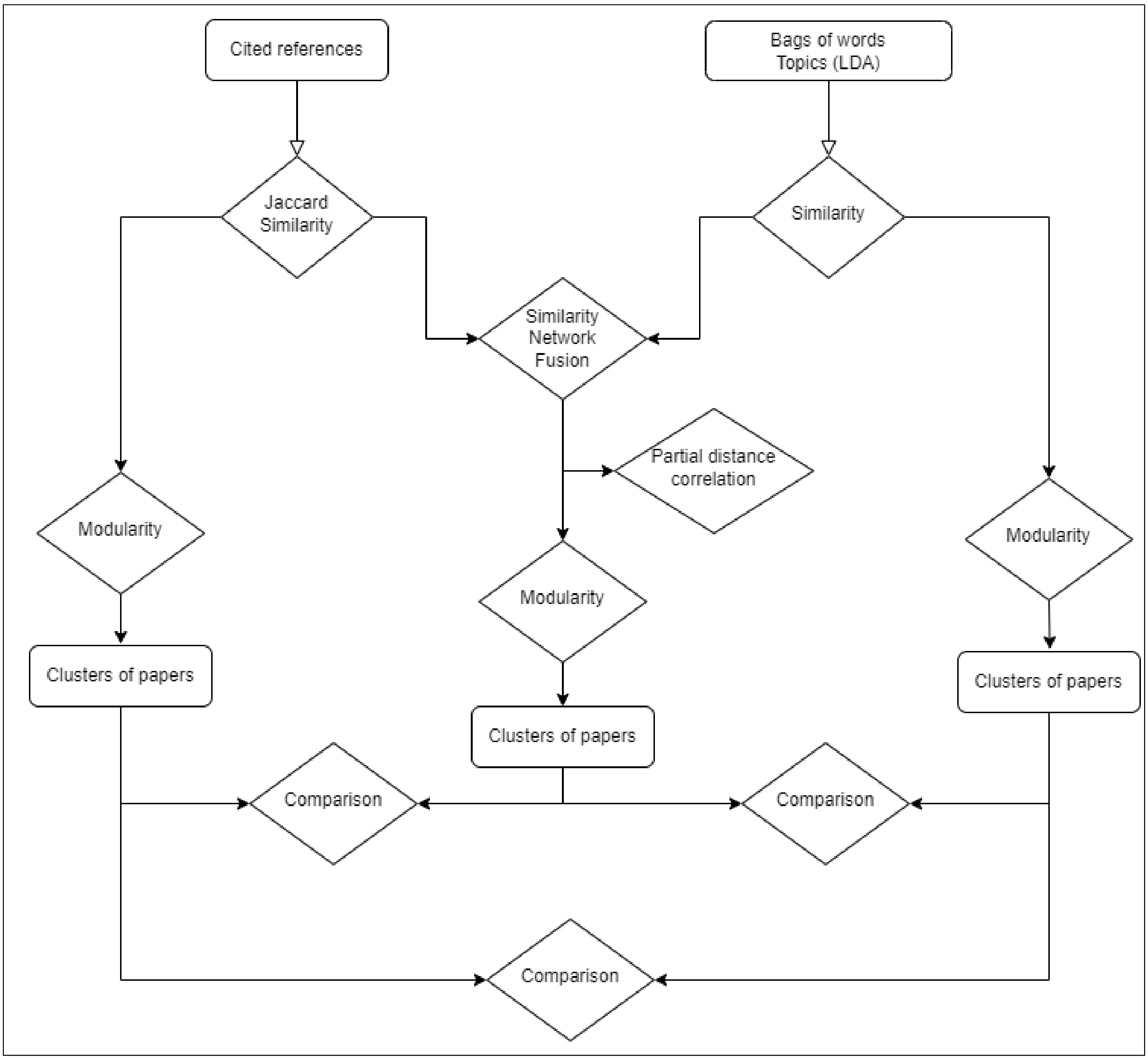

Figure 1: The workflow of exploratory analysis.

frequency in the article. Under this setting, it is presumed that the style of an article is mainly revealed through the choice of words, and in particular by the very frequent words in the whole corpus. For more details, see the survey provided by Jacques Savoy [32, Chapter 3] in a useful monograph dealing with stilometry. The word types do not have a precise meaning, and they induce a stylistic description, which is independent of the topics of the underlying article. To be more explicit, let us assume that there exist L most frequent word types in the selected articles, without taking punctuation marks or numbers into consideration. In addition, let p_{il} be the relative frequency of the l-th word in the corpus for the i-th paper. Thus, the matching between a pair of papers may be computed by using a similarity concept based on the total variation measure, and given by

$$S_{ij} = 1 - \frac{1}{2} \sum_{l=1}^{L} |p_{il} - p_{jl}| \ . \tag{2}$$

Actually, the normalized total variation measure between the i-th and j-th article is nothing else than $1-S_{ij}$. The total variation measure is usually (and very naturally) adopted for comparing two categorical distributions (see e.g. [1]). In turn, we have that $0 \leq S_{ij} \leq 1$. In particular, if two articles have the same relative frequency of words in the corpus, the maximum similarity $S_{ij} = 1$ occurs. On the other hand, the minimum similarity $S_{ij} = 0$ is reached when the two articles adopt completely different words. In the stylometric framework, it is worth noting that $1 - S_{ij}$ is actually the Labbé's inter-textual distance (for more details see [32]).

The second approach is based instead on topic modeling. Under this framework, a pre-defined number of "topics", say K is considered and the Latent Dirichlet Allocation (LDA) method is carried out on the whole set of articles [8]. A detailed account of LDA is given by [32, section 7.3]. It should be remarked that the definition of "topic" is not a subject heading under LDA, as commonly intended. Instead, LDA produces K lists of word types and their distributions in the whole corpus. Each word list and the corresponding word distribution is defined as a topic in LDA. Hence, we also have a topic distribution for each paper. In such a case, we obtain a bipartite network where the first set of nodes contains articles and the second set the topics. Edges link each paper to the topics with a weight proportional to the topic probability for each paper. Again, similarities between pairs of papers can be computed by using the above-mentioned similarity index based on the total variation measure. However, in this case, p_{il} turns out to be the relative probability of the l-th topic for the i-th paper and L is replaced by K.

As anticipated, LDA requires the topic number K as a parameter to be defined. Actually, K may be considered in some sense as a "smoothing" parameter, and no obvious selection rule can be generally given. A possible empirical strategy consists in defining different numbers of topics for testing the stability of the results obtained. In our database, results are computed by assuming $K = 5, 10, 15, 20, 25, 30$.

In sum, the primary outputs of the data analysis are: (i) a paper similarity matrix based on cited references; (ii) a paper similarity matrix based on BoW; (iii) a set of 6 paper similarity matrices based on topics, each one referred to the selected topic number K.

6 Do similarity matrices contain the same information?

In order to compare the dependence between the similarity matrices, the generalized distance correlation R_d suggested by [37] is adopted. Its interpretation is similar to the squared Pearson correlation coefficient: R_d is defined in the interval $[0, 1]$. Values close to zero indicate no or very weak association; larger values indicate a stronger association, which is perfect for $R_d = 1$ - similar considerations hold for $\sqrt{R_d}$ (for more details, see [29]).

The generalized distance correlations between matrices based on topics reported in Table 2 allow checking whether the information obtained changes by using different topic numbers K. Results indicate that when a K higher than 5 is chosen, the information obtained is substantially similar, with values of generalized distance correlations generally higher than 0.9. The similarity matrix obtained by setting $K = 5$ has, by contrast, the lowest association.

The generalized distance correlation between the matrices based on topics and the similarity matrix based on BoW allows, on the other hand, controlling if the use of BoW and topic modeling produces similar information in terms of similarities between articles. The generalized distance correlations also reported in Table 2 are stable at about 0.71 whatever the number of topics. This result can be interpreted as indicating that the use of topics does not entail a relevant loss of information with respect to the use of BoW.

Table 2: Generalized distance correlation between article similarity matrices.

dCor	Topics_5	Topics_10	Topics_15	Topics_20	Topics_25	Topics_30	Bags of words	Cited References
Topics_5	1	0.82975	0.82816	0.82849	0.81031	0.79929	0.71164	0.40801
Topics_10		1	0.92928	0.90471	0.8897	0.88892	0.71164	0.46437
Topics_15			1	0.95045	0.9338	0.9255	0.70745	0.49207
Topics_20				1	0.97351	0.96069	0.71183	0.50102
Topics_25					1	0.98026	0.71032	0.5144
Topics_30						1	0.70574	0.51935
Bags of Words							1	0.45511
Cited References								1

Finally, the generalized distance correlations between the article similarity matrix based on cited references and, respectively, the matrix based on BoW and the matrices based on topics indicate if the information contained is associated or not. In the case of a very high distance correlation between matrices, one could argue that the choice of one or the other matrix is not relevant, as the information conveyed by both matrices is highly associated. On the opposite, a very low distance correlation could indicate that the matrix based on cited references and matrices based on BoW or topics convey different information and then both kinds of information should be considered. In the case at hand, the last column of Table 2 shows that the generalized distance correlation between the matrix based on cited references and the matrices based on topics has an intermediate value that tends to a slight growth according to the number of topics. Also the generalized distance correlation between the matrix based on cited references and the matrix based on BoW has an intermediate value. The ambiguity of all these results suggests, however, that the information conveyed by cited references and contents is not highly associated and, as a consequence, both information sources should be taken into consideration.

7 Classifications of articles based on modularity

As anticipated, the classifications of papers are developed by using the Louvain algorithm based on modularity [9]. The algorithm is applied to the paper similarity matrices built by considering topics, BoW and cited references. The number of clusters and the corresponding values of modularity are reported in Table 3.

The Louvain algorithm applied to the paper similarity matrix based on cited references produces 51 clusters (modularity 0.41). Only 6 clusters have more than 100 articles for a total of 1,278 articles (95.1% of the total number of articles), 4 contain less than 8 articles (for a total of 22 articles, 1.6%), and the remaining 41 clusters are composed mainly by isolated articles (for a total of 44 articles 3.3%).

When the Louvain algorithm is applied to similarity matrices based on topics, for any number of topics, it produces 5 clusters with the only exception of the case of Topic_10 which results in 4 clusters. The modularity values are very similar for any number of topics.

Table 3: Clusters of articles obtained through Louvain algorithm applied to article similarity matrices based on cited references, topics and bags of words.

	n. of clusters	Modularity
Topics_5	5	0.351041
Topics_10	4	0.315298
Topics_15	5	0.310147
Topics_20	5	0.320850
Topics_25	5	0.324278
Topics_30	5	0.324139
BoW	3	0.039607
Cited References	51	0.409726

Finally, when the similarity of articles in terms of BoW is considered, the Louvain algorithm produces 3 clusters with very low modularity, indicating that it is unable to neatly separate clusters of articles.

In Table 4, the values of Cramer's V are reported for measuring the associations between classifications obtained for different numbers of topics, bags of words, and cited references. All the classifications obtained by considering topics are highly associated. The classifications for Topic_15, Topic_20, and Topic_25 have particularly high values of Cramer's V.

Analogously, all the clusters obtained for different numbers of topics are highly associated with the clusters obtained for bags of words: this provides further support to the observation made above that the use of topics does not entail a loss of information compared to the use of BoW.

Finally, the association between clusters based on cited references and the ones based on topics and bags of words shows intermediate values. This is coherent with the result obtained by considering the generalized distance correlations: the information conveyed by the similarity network based on cited references and the ones based on BoW and topics is not highly associated.

Table 4: Values of Cramer's V for measuring the association between clusters obtained from article similarity matrices based on cited references, topics and bags of words.

	Topics_5	Topics_10	Topics_15	Topics_20	Topics_25	Topics_30	BoW	CR
Topics_5	1	0.72165	0.67551	0.6988	0.72272	0.72783	0.8061	0.63879
Topics_10		1	0.7976	0.7767	0.7888	0.78278	0.79833	0.57177
Topics_15			1	0.8265	0.77156	0.76054	0.79762	0.52129
Topics_20				1	0.8417	0.80093	0.79243	0.64479
Topics_25					1	0.89111	0.7918	0.53521
Topics_30						1	0.79447	0.53811
BoW							1	0.51175
CR								1

In sum, this step of analysis suggests that article similarity matrices based on topics and on BoW contain entirely similar information and result in overlapping papers classifications. The similarity matrix based on cited references instead appears to convey information different from the matrices based on topics and bags of words, and, accordingly, the classification obtained from it does not overlap with the others.

8 Similarity network fusion

The adoption of a Similarity Network Fusion technique (SNF) permits to synthesize in a single layer the information contained in multiple layers [41, 42, 6, 5]. In the present case, the SNF is realized by considering the two layers of the multiplex formed by cited references and by one of the matrices based on topics, or, alternatively, by bags of words. SNF integrates into a unique similarity matrix the pair of original matrices by means of the Cross Diffusion Process (CDP) [41, 42]. CDP is an unsupervised iterative procedure that reinforces very strong links present in the single layers, and at the same time weakens links that are common to all the layers. Unlike the other techniques recalled in the literature review, SNF requires neither the choice of weights to combine different layers of information nor assumptions about the statistical distribution of the data.

As a result, SNF generates a new network where nodes are articles and the edges are weighted according to the new similarity values obtained through CDP. In fact, the iterative procedure enriches the information of a single layer with that coming from the other layer. SNF maintains densely connected groups of articles and reinforces their links; at the same time, SNF makes more visible low weighted links that are present in both the layers, as they may represent stable relationships among groups of articles.

SNF was realized in the `R` environment by using the `SNF` function in the **SNFtool** package (R Core Team, 2020).

Seven fused networks are realized, corresponding to the fusion of the similarity matrix based on cited references, with each of six similarity matrices based on topics (Fused_5, ..., Fused_30), and with the matrix based on bags of words (Fused_BoW).

The generalized distance correlation can be used for evaluating how much of the information contained in the two single layers is reported also in the fused networks. The second column of Table 5 reports the generalized distance correlation between the similarity matrix based on cited references and each of the seven fused networks. The correlation is high and it is also stable for a number of topics greater than 5 and BoW.

The third column of Table 5 reports the generalized distance correlation between each of the similarity matrices based on topics and on BoW used for each fusion, and each of the seven fused networks. Correlation is very low for Topic_5 and tends to grow with the number of topics. The correlation between the Fused_BoW matrix and the matrix based on BoW as an ambiguous value.

Table 5: Generalized distance correlation between fused matrices and similarity matrices based on topics and cited references.

	Similarity matrices based on:	
Fused matrices	cited references	topics
Fused_5	0.9567	0.33768
Fused_10	0.76	0.64455
Fused_15	0.76597	0.6735
Fused_20	0.76685	0.69133
Fused_25	0.76663	0.70341
Fused_30	0.76883	0.70031
Fused_BoW	0.76281	0.54714

For verifying if the SNF generates different structures of fused networks according to the number of topics, it is possible to use again the generalized distance correlations among fused matrices as reported in Table 6. It shows that the distance correlations among the fused matrices are very high for all the numbers of topics and only slightly lower for bags of words. As a result, the structures of the fused matrices are similar and contain very similar information.

Table 6: Generalized distance correlation between fused matrices.

	Fused_5	Fused_10	Fused_15	Fused_20	Fused_25	Fused_30	Fused_BoW
Fused_5	1	0.73994	0.74282	0.742	0.74025	0.74221	0.74703
Fused_10		1	0.98197	0.96931	0.96712	0.96474	0.96399
Fused_15			1	0.98039	0.97576	0.97121	0.9575
Fused_20				1	0.98449	0.97666	0.95694
Fused_25					1	0.98724	0.9552
Fused_30						1	0.95214
Fused_BoW							1

8.1 Contribution of the two layers to the fused network

The partial distance correlation R_d^* proposed by Székely and Rizzo [36] can be used for analyzing the contribution of each layer to the similarity structure of the fused networks, as proposed by [4]. It measures the degree of association between the similarity matrix of the fused network and a layer, by removing the effect of the second layer. The partial distance correlations were evaluated in the R-computing environment (R Core Team, 2020) by using the `pdcor` function in the package **energy**. The computed values of these coefficients are reported in Table 7.

The partial distance correlations indicate that the layer of topics give the major contribution to the structure of the fused networks, irrespective of the number of topics chosen. When the layer based on bags of words is considered, instead, cited references give the major contribution to the structure of the fused network. This is probably due to the difficulty in clustering the layer based on BoW, as previously noted.

Table 7: Partial distance correlation between the seven fused networks and the two corresponding layers. Each row reports the values of partial distance correlation between the fused networks and one of the layers, conditioned on the other layer.

		Fused_5	Fused_10	Fused_15	Fused_20	Fused_25	Fused_30	Fused_BoW
Fused, Topics|Cited references	$\sqrt{R_d^*}$	0.1768	0.37107	0.40099	0.42678	0.43928	0.42847	0.22674
Fused, Cited references|Topics	$\sqrt{R_d^*}$	0.31864	0.31076	0.31204	0.30748	0.30214	0.30345	0.34617

8.2 Classification of papers in the fused networks

Also in this case, the classification of articles in the seven fused networks is conducted by using Louvain algorithm. Table 8 reports the number of resulting clusters and the modularity values. The number of clusters lies between 7 and 9 for fusions based on topics, while it goes up to 11 when the fusion is based on bags of words. The values of modularities are always similar.

Table 8: Clusters of articles obtained through Louvain algorithm applied to fused networks.

	n. of clusters	Modularity
Fused_5	7	0.427179
Fused_10	7	0.426463
Fused_15	9	0.430192
Fused_20	8	0.431632
Fused_25	9	0.431812
Fused_30	8	0.430127
Fused_BoW	11	0.445219

The stability of the classifications between all the possible pairs of fused networks is also explored by estimating the Cramer's V, as reported in Table 9. These values indicate that the classifications have a relatively high degree of association that tends to grow with the number of topics used in the fusion. The clustering obtained for the Fused_BoW matrix has high and stable degrees of association with all the clustering obtained in the other fused matrices.

Table 9: Values of Cramer's V for measuring the association between classifications of articles in the fused networks.

	Fused_5	Fused_10	Fused_15	Fused_20	Fused_25	Fused_30	Fused_BoW
Fused_5	1	0.624	0.61886	0.67017	0.71679	0.61291	0.84775
Fused_10		1	0.66071	0.58883	0.59487	0.58463	0.79917
Fused_15			1	0.69633	0.73132	0.74926	0.71154
Fused_20				1	0.73035	0.73666	0.74585
Fused_25					1	0.64443	0.788
Fused_30						1	0.70443
Fused_BoW							1

9 Main characteristics of groups of papers

As anticipated, the papers published in the *Cambridge Journal of Economics* are characterized by an extreme heterogeneity both from the point of view of their analytical approach and of the topics covered. This heterogeneity is reflected in the low value of distance correlation (0.50102, see Table 2) between the papers' similarity matrices obtained from textual information and from cited references. On the other hand, the classifications obtained for fused networks are highly stable. Hence, in what follows the Fused_20 network case only is analyzed and discussed in detail. Note that the case of Fused_20 is the less favorable for illustrating the utility of fusing information instead of relying on only one of the two layers of information. Indeed, the value of Cramer's V for the association between clusters obtained in the network based on cited references and on Topic_20 is the highest, i.e. the information conveyed by the two networks has the highest level of association with respect to all the other cases. Hence, valuable results in the most challenging case are a good premise to argue about the utility of fusion in less challenging cases, when the association between layers is lower.

More in detail, Figure 2 compares the clusters obtained in Topic_20 and cited-references networks. The Topic_20 network is partitioned by Louvain algorithm in 5 clusters reported in rows and labelled T1, ..., T5. The cited-references network is instead partitioned in 51 clusters, 45 of which contain either isolated papers or groups of less than 8 articles. For the sake of simplicity Figure 2 reports in columns the 6 largest clusters R1, ..., R6; plus a seventh one, labelled "R99", that contains the other 45 clusters. The largest 6 clusters collect 1,278 articles, i.e. the 95.1% of the total number of classified articles. The value of 0.64479 for Cramer's V indicates a moderate association between the classification based on Topic_20 and the one based on cited references. Figure 2 synthesizes the distinction between the two classifications: articles belonging to the same cluster on the basis of one information source are generally spread over at least three different clusters on the basis of the other.

We will get into more detail shortly by illustrating how the various groups of articles obtained from the fused network exploit such a richness of information. Meanwhile, it is worth noting some meaningful overlapping. In particular, in absolute terms, the two most important overlaps are the one between R4 and T1 (215 papers, i.e. 50% of R4 and 72% of T1) and the one between R4 and T4 (172 papers, 40% of R4 and 48% of T4). Now, T1 contains papers mainly pertaining to the fields of industrial and labor economics, while T4 gathers papers focused on the history of economic thought and on economic methodology. In turn, R4 is characterized by the prevalence of papers in the institutionalist and evolutionary traditions, two approaches that in terms of topics are primarily devoted to the theory of the firm and of industrial organization and which, having their methodological premises as a distinctive mark, devote a great deal of space to methodological research. This explains the overlapping: R4 shares with T1 many works on industrial economics, and with T4 many papers with strong methodological content. Notice that cluster R4 has as most cited references works by Robert Nelson and Sidney Winter, Oliver Williamson, and Joseph A. Schumpeter (ranking 7th, 8th and 9th in the entire database of cited

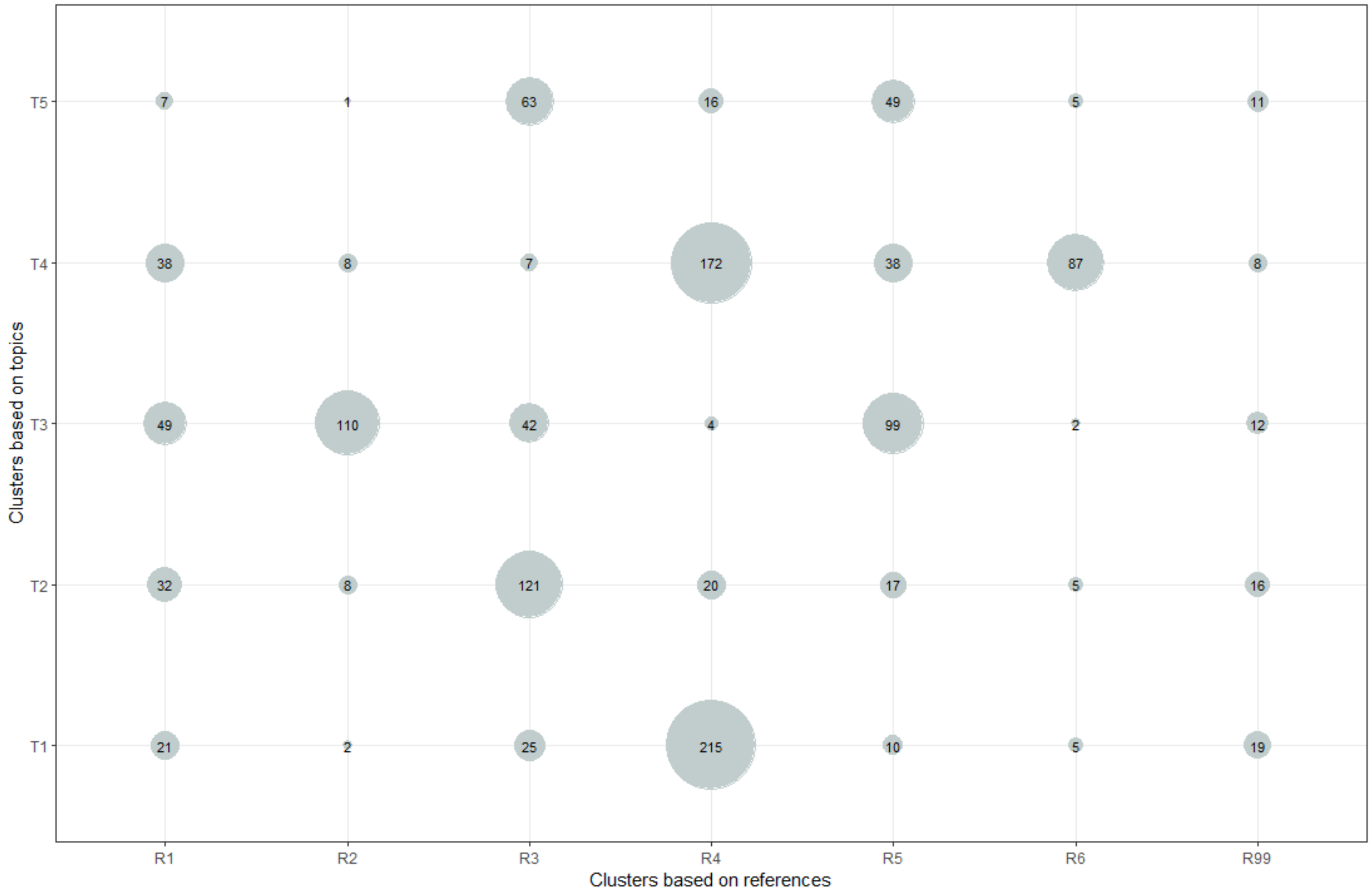

Figure 2: Comparison of clusters in Topics_20 and in cited references networks. The five clusters obtained through Louvain algorithm in the Topics_20 network are in the x-axis and labelled as T1, ..., T5. The six big clusters obtained through Louvain algorithm in the cited reference are in the y-axis and labelled as R1, ..., R6; R99 collects 45 clusters with isolated articles or with a size less than 8 articles.

references, see Table 1) that are clearly related to the institutionalist and evolutionary traditions. There are two other overlaps worth noting. First, the large majority of the articles in R6 are in T4. This is because R6 gathers methodological papers that are grouped in T4 with papers on the history of economic thought. Secondly, most of the papers belonging to R2 also belong to T3. R2 collects papers primarily connected to the Sraffian approach, which in T3 are combined with papers in the post-Keynesian tradition and with Marxist analyses so as to form a broader set that, following a now established practice, we can label "Cambridge Economics" [27].

Turning now to the analysis of the results obtained through the similarity network fusion, the Louvain algorithm detects 8 clusters in the Fused_20 network. Table 10 reports the number and the proportion of articles classified in each group; groups are labelled according to the expert observation described below; the color codes are the ones of Figure 3.

Table 10: Clusters in the Fused_20 network.

Cluster	Label	Color	N. of papers	%
F1	Classical and Marxian political economy	SkyBlue	210	15.6
F2	Monetary economics and history of macroeconomics	Black	197	14.7
F3	Growth and development economics	Orange	189	14.1
F4	Economics of institutions	Orchid2	246	18.3
F5	Economics of firms, industry, and technical change	VioletRed2	132	9.8
F6	Instability of capitalist economic systems	Acquamarine3	131	9.7
F7	Rural economies	Bisque2	13	1.0
F8	Economic methodology	PaleGreen3	226	16.8

Figure 3 shows the fused network. The alluvial plot [12] of Figure 4 compares the clusters obtained in the Fused_20 network with the ones present in the Topic_20 and Cited References network. The central stratum is composed of eight blocks corresponding to the clusters in the Fused_20 network. The blocks are coloured according to Table 10; their heights are proportional to the size of the clusters. The left stratum is composed of five blocks representing the clusters detected in the Topic_20 network; the right stratum by seven blocks representing the clusters in the Cited Reference network. The flows between the central stratum and the other two lateral strata represent how the clusters in the central stratum are composed in terms of articles clustered in the two lateral strata. The flows are coloured according to the clusters in the Fused_20 network, and the height of the flows is proportional to the size of the components contained in both blocks connected by the stream field. In the supplementary material Figure A.1 provides a different representation of the composition of clusters.

The first cluster F1 gathers 210 papers (15.6% of the articles) that primarily belong to the school of thought that can be defined as Classical and Marxian Political Economy, i.e., contributions in the tradition that was brought back to light by Piero Sraffa (1960) after it had been abandoned with the rise of the neoclassical paradigm. Note that Sraffa's work is the third most cited document in the dataset (see Table 1). The F1 cluster thus contains many papers on the history of economic thought dealing with the Classical economists and Marx. From the point of view of the topics addressed, it reflects the multiplicity of problems that formed the subject of these economists' analyses, mainly the theory of value and distribution, fiscal policy, the theory of international trade, and the analysis of long-term trends of the economic system. The cluster also contains papers dealing with the critique of the neoclassical approach, thus expanding on the work undertaken in this respect by Sraffa, as well as contributions to the debate on Marxist theory and its relationship to Classical political economy. Marx's *Capital* figures as the second most cited document in the dataset, while the *Wealth of Nations* by Adam Smith, who is arguably the father of Classical political economy, ranks sixth. Figure 4 shows that the Classical and Marxian Political Economy cluster brings together papers that the classification based on cited references places in two distinct groups: namely, contributions in the Marxist tradition

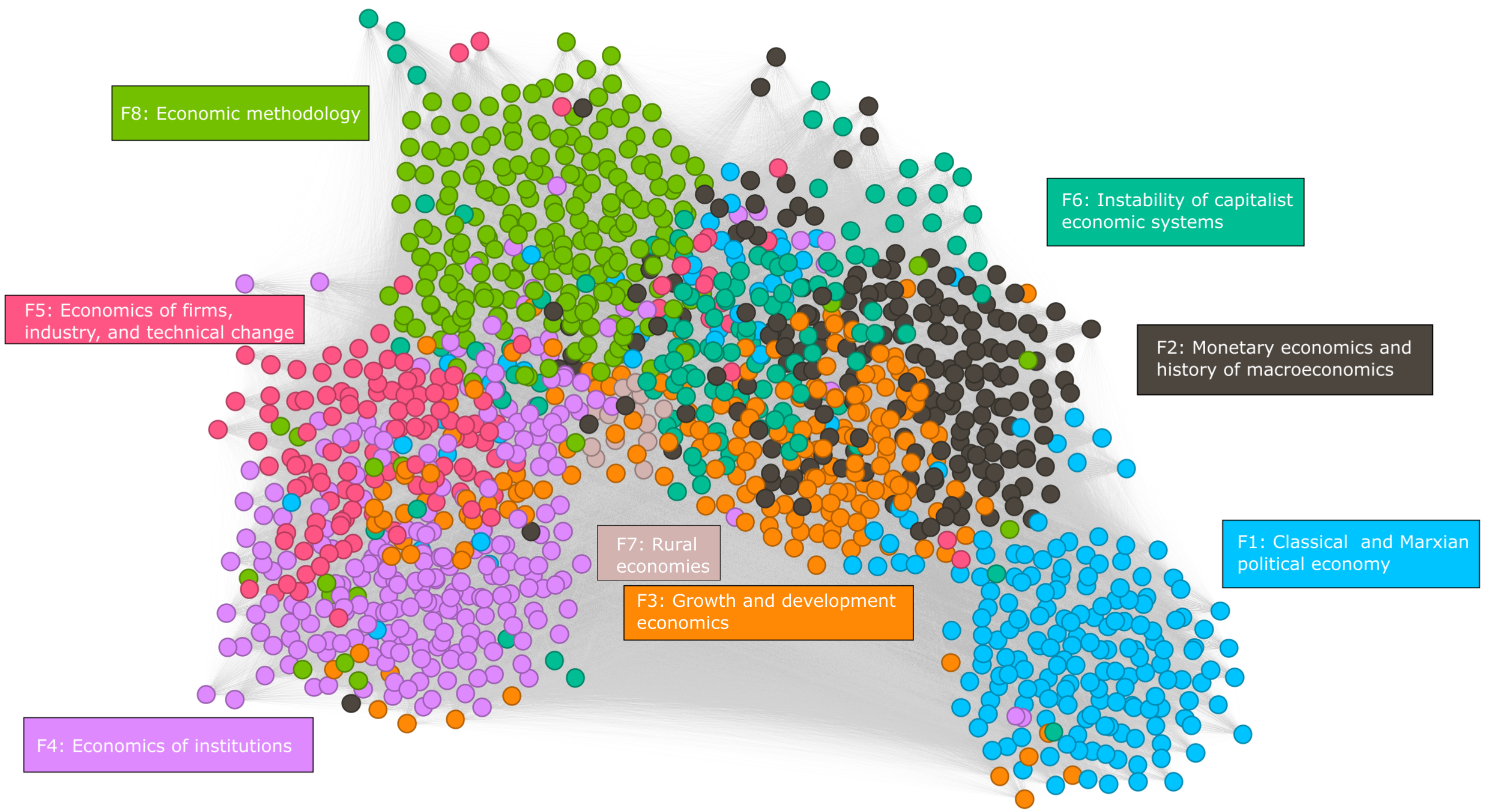

Figure 3: The fused similarity network in the *Cambridge Journal of Economics*. Nodes are articles. They are colored according to the clustering reported in Table 10, obtained through the Louvain algorithm based on modularity in the Fused_20 network. The graph is realized by GEPHI with `OpenOrd` and `Noverlap` visualization algorithms.

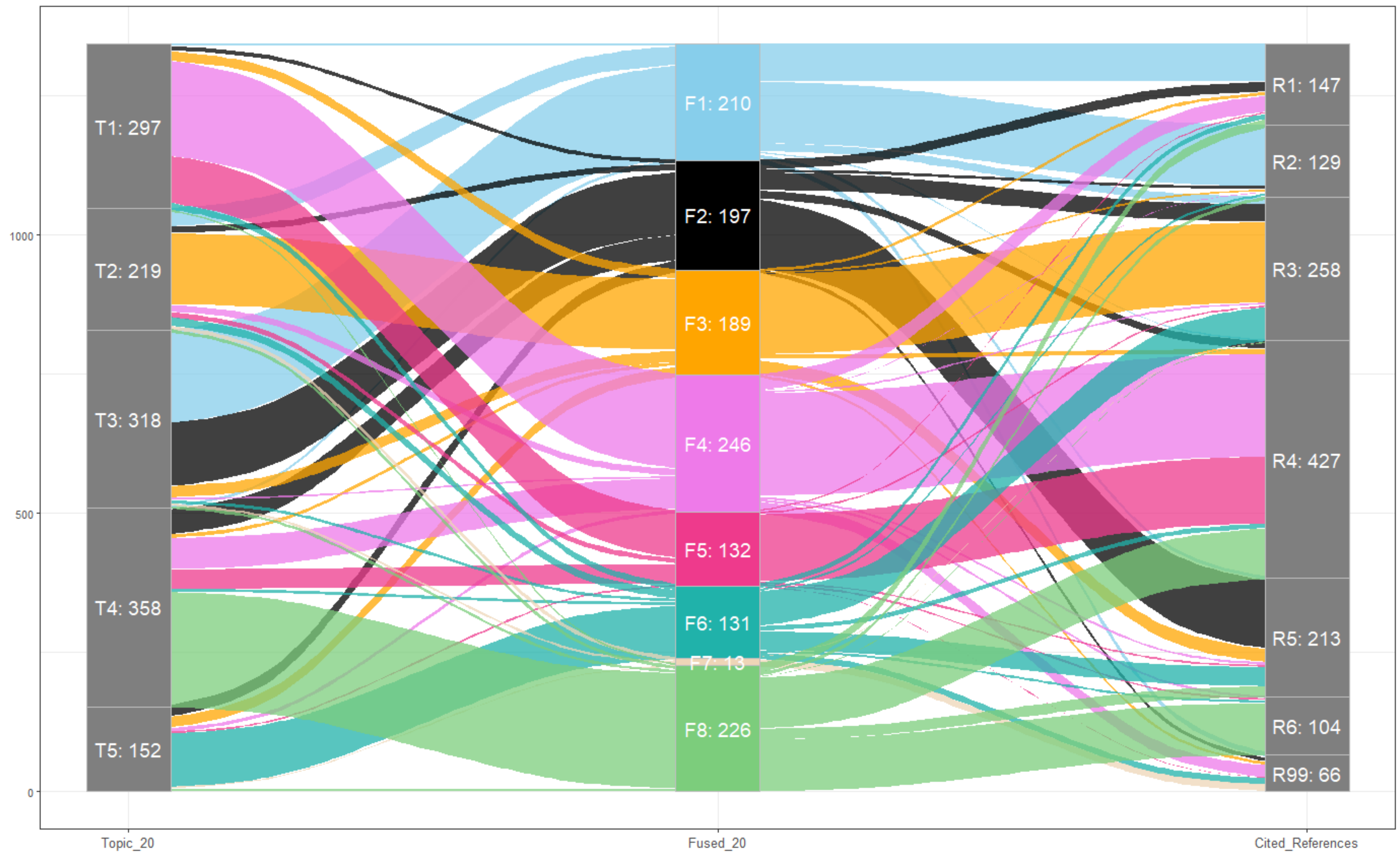

Figure 4: Clusters of *Cambridge Journal of Economics* articles in the fused network and in the networks based on Topic_20 and Cited References. The strata of the alluvial plot refer to the clusters obtained in each network. The central stratum and the flows of the plot are colored as in Figure 3. Strata report the codes and the number of articles of each cluster. The plot was realized in R environment by using the **ggalluvium** packages [13].

gathered in cluster R1, and works in the vein of classical economists and Sraffa gathered in cluster R2. The cluster obtained from the fused network draws from two groups that the citation information keeps separate, pointing at a new, significant connection. The connection is all the more meaningful in that it is not clearly identifiable even by adopting textual information as a basis for classification. Indeed, while it is true that the papers in the cluster come mostly from T3, there is a significant proportion of papers coming from T2, mainly papers in the Marxist tradition dealing with macroeconomic issues. Most importantly, the papers from T3 account for only half of total T3, a cluster in which the Sraffian-oriented papers are grouped together with contributions on Keynesian and post-Keynesian theory belonging to the rather comprehensive group that we already called "Cambridge economics".

Hence, the F1 cluster is mainly characterized in terms of the theoretical approach. For the F2 and F3 clusters, instead, it is rather the homogeneity of the topics dealt with that justifies the grouping, although a certain prevalence of contributions in the post-Keynesian tradition can be observed. Both clusters collect contributions in the field of macroeconomics. The F2 cluster gathers 197 papers (14.7% of the articles) focused on Monetary economics and the history of macroeconomics. Analogously to the first cluster, also in this case the fusion emphasizes a significant connection which is less evident on the basis of the two classifications based on citation and textual information. Indeed, the Monetary economics and the history of macroeconomics cluster brings together papers that on the basis of cited references are distributed mainly between clusters R3 and R5, while on the basis of content, these papers are partly in T3 and partly in T4. The significance of the grouping we observe combining the two sources of information can be appreciated by taking into account the fact that the economist who is universally regarded as the father of macroeconomics, John Maynard Keynes, made one of his most innovative contributions in the theory of money. Now, most of the papers on the history of macroeconomics published in the *Cambridge Journal of Economics* are devoted precisely to Keynes or to economists who draw on his contribution – post-Keynesian economists. These papers are thus largely concerned with monetary issues: in this regard, notice that Keynes's *General Theory* is the most cited document in the entire database. It seems therefore reasonable to group the contributions on monetary economics together with those on the history of macroeconomics.

The F3 cluster gathers 189 papers (14.7%) mainly devoted to Growth and development economics. Most of the papers classified in this group are classified as T2 and R3. In this case, the fused network tends to reproduce the classifications obtained by considering topics and cited references separately. Indeed, most of the papers in R3 belong to T2, and most of T2 papers belong to R3.

The F4 cluster in the fused network, gathering 246 articles (18.3%), is the largest and, at first sight, the most heterogeneous since it covers many research areas such as labor economics, comparative economic systems, social structure, law and economics. The label Economics of institutions indicates that all the areas of research represented in this cluster share a common thread in the analysis of the role of institutions in the economy. This theme is explored in different directions, ranging from the consideration of the role of social institutions to the analysis of economic institutions and, with reference to the latter, investigating both single markets, such as the labor market, and the general structure of the economic system. The Economics of institutions cluster brings together almost half of the papers belonging to R4 with one-fifth of the papers belonging to R1. At the same time, it groups more than half of the papers in T1 with 15% of the papers in T4.

As for the F5 cluster, this contains 132 papers (9.8%) mainly dealing with the Economics of firms, industry, and technical change. From the point of view of the theoretical paradigm, there is a prevalence in this group of the evolutionary approach, a school of thought whose main contributions relate precisely to the theory of the firm, the economics of innovation and technical change, and the theory of industrial organization. In this case, too, the classification based on the whole set of information available isolates a set of papers that according to the two classifications based on topics and cited references do not form a group of their own. Indeed, while almost two-thirds of the papers in F5 cluster come from T1, these papers within T1 are found together with contributions on labor economics that are gathered in the fourth cluster in the fused network. Similarly, while almost all of the papers in the fifth cluster come from R4, these papers represent just over a quarter of R4, where they are collected together with contributions that are again gathered in the fourth cluster in the fused network.

The F6 cluster contains 131 papers (9.7%) primarily concerned with the Instability of capitalist economic systems. The main issues addressed are related to financial markets, and financial fragility in particular, to the sustainability of fiscal policy, together with the large set of issues raised by the outbreak of the Great Recession, with a particular focus on the debate on the response to the crisis within the European Union. It is mainly the information coming from topics analysis that contributes to determining this cluster, with two-thirds of the papers coming from group T5; with respect to the classification based on cited references, this cluster mainly draws from two groups that collect macroeconomic papers, R3 and R5, in which however the subjects of instability and crises do not stand out distinctly.

The F7 cluster (labelled Rural economies) represents a negligible component of the network, as it only contains 13 articles (1.0%). It is worth mentioning, however, that this group has a clear characterization since all papers deal with topics related to rural economies, such as rural poverty, the role of informal credit markets, and agricultural production. These papers are classified into three different clusters according to topics and are almost all dispersed in R99.

Finally, the last F8 cluster gathers 226 articles (16.8%) and can suitably be labelled Economic methodology. Quite reasonably, along with a large majority of articles with strong methodological content, in this group we find contributions that deal with the question of pluralism in economics or, more generally, with the sociology of economics. Once again, it is the textual information that provides the greatest contribution, as the cluster is mostly made up of articles from group T4; conversely, it draws from two distinct groups from the point of view of the classification based on cited references: indeed, it is true that there is a group with a clear methodological character, R6, but as we have seen an important share of methodological papers are found in R4, which gathers contributions in the institutionalist and evolutionary traditions.

Overall, it appears that the extreme heterogeneity that characterizes the papers published in the *Cambridge Journal of Economics*, while making any attempt at classification more difficult, makes the similarity network fusion technique particularly valuable. The papers in our dataset lend themselves to being classified both on the basis of their analytical approach and on the basis of the topics covered, and the two classifications based on cited references and topics do not always favor the same criterion. It is therefore certainly interesting to combine textual and citation information in order to identify the criterion that yields the most strongly connected groups of papers.

10 Conclusion

The issue of using multiple sources of information for classifying papers and delineating research fields is an old problem in scientometrics. Usually, the delineation of scientific fields is conducted by considering only one layer of information at a time. There are classifications based on information about citations and references; and classifications that rely on content by adopting distant reading techniques. In most cases, the use of different sources of information results in different classifications of the same set of articles. So far, the attempts to use together multiple sources of information for classifying articles and delineating research fields have adopted techniques that require strong assumptions about statistical distributions of data, and about the weights to assign to different information when they are integrated. The Similarity

Network Fusion technique proposed here is an unsupervised technique able to integrate different layers of information in a single similarity network. SNF technique does not require any assumption about the statistical distribution of data, nor the choice of weights to be attributed to the layers of information when they are combined.

The case-study addressed in this work regards the classification of articles published in the *Cambridge Journal of Economics*. Founded in the 1970s, it is the leading generalist non-mainstream economics journal, open to contributions from diffent schools of economics. The task of classifying its articles is particularly difficult as they differ not only in terms of subjects but also in approaches to the same subject.

To this end, two layers of information are used: one based on bibliographic coupling, and the other based on contents. The bibliographic coupling served to define a similarity network among articles. By exploiting the full-text of articles, the similarity of contents was defined by using two different approaches. The first one was based on Bags of Words, i.e. on the relative frequency of words in articles; the second one, requiring more statistical assumptions, was based on LDA topic modeling and produced six different similarity networks corresponding to different pre-defined numbers of topics.

The first result of the paper consists in showing that Bags of Words and LDA produce similarity networks highly associated. Hence, the use of one or the other technique, and the adoption of different numbers of topics do not entail a relevant loss of information. Indeed, when the same technique of clustering is applied to the seven networks based on contents, the resulting classifications of articles are highly associated.

The second result confirms previous analyses: the similarity network based on cited references and the ones based on contents have a moderate level of association; hence they convey different information about articles. Indeed, when the same technique of clustering is applied to the network based on cited references and to the networks based on contents, the resulting classifications of articles have a low level of association.

Thus, the adoption of a technique for integrating information about cited references and contents appears fully justified.

The third result of the paper regards the application of the similarity network fusion technique. It results in seven different fused networks which have highly correlated structures, and which produced highly associated classifications of articles. A technique for showing the contribution of each layer to the structure of the fused network is also presented, and, in the present case, it shows that contents contributed more to the final result than cited references.

The classification obtained through SNF has been evaluated from an expert point of view, by inspecting whether it can be interpreted and labelled with reference to research programs and methodologies adopted in economics. Moreover, the classification obtained in the fused network is compared with the two classifications obtained when cited references and contents are considered separately.

Overall, the classification obtained on the fused network appears to be fine-grained enough to represent the extreme heterogeneity characterizing contributions published in the *Cambridge Journal of Economics*. The articles lend themselves to being classified both on the basis of their analytical approach and on the basis of the topics covered. The two classifications based on cited references and topics do not always favor the same criterion, thus resulting in less fine-tuned classifications.

The discussion of this last point has been conducted by considering the less favorable case, i.e. for the highest value of association between clusters obtained in the network based on cited references and the one based on topics. The discussion of results highlighted that the fine-grained classification obtained in the fused network appears qualitatively superior to the classifications obtained in the two layers separately. This suggests that fusion may be more effective in less challenging cases. In sum, the

SNF technique appears as a useful tool for the complex task of fine-grained classification of articles.

The results presented here are promising but more research is needed. A first step might consist in comparing the classification obtained from SNF with the expert classification of articles defined by institutions or scholars. In the research agenda of this group, there is the comparison of the classification of articles obtained for the articles of the *Cambridge Journal of Economics* with the classification (not so easily available) adopted by the *Econlit* database, maintained by the American Economic Association.

A second line of research might consist in extending the analysis by adding other layers of information to the two adopted here, for instance by adding the similarity network based on co-citation among articles, on mentions of articles in social media, or on keywords chosen by the authors. These extensions require a minimum adaptation of the setting presented in this work.

A third line of research might consist in applying SNF technique to the classification of papers and journals in other research fields than economics for verifying the suitability of the approach. It would surely be interesting to apply the technique to sets of papers much bigger than the one explored here and in a multidisciplinary context, in view of verifying the robustness of the instrument.

References

[1] Alan Agresti. *Categorical data analysis*. Vol. 792. John Wiley & Sons, 2012. ISBN: 0470463635.

[2] Per Ahlgren and Cristian Colliander. "Document–document similarity approaches and science mapping: Experimental comparison of five approaches". In: *Journal of Informetrics* 3.1 (2009), pp. 49–63. ISSN: 1751-1577. DOI: https://doi.org/10.1016/j.joi.2008.11.003. URL: https://www.sciencedirect.com/science/article/pii/S1751157708000680.

[3] Angela Ambrosino et al. "What topic modeling could reveal about the evolution of economics". In: *Journal of Economic Methodology* 25.4 (2018), pp. 329–348. DOI: 10.1080/1350178X.2018.1529215. eprint: https://doi.org/10.1080/1350178X.2018.1529215. URL: https://doi.org/10.1080/1350178X.2018.1529215.

[4] Alberto Baccini et al. "Intellectual and social similarity among scholarly journals: An exploratory comparison of the networks of editors, authors and co-citations". In: *Quantitative Science Studies* 1.1 (2019), pp. 277–289. DOI: 10.1162/qss_a_00006.

[5] Federica Baccini, Monica Bianchini, and Filippo Geraci. "Graph-Based Integration of Histone Modification Profiles". In: *Mathematics* 10.11 (2022), p. 1842.

[6] Federica Baccini et al. "Similarity network fusion for scholarly journals". In: *Journal of Informetrics* 16.1 (2022), p. 101226. ISSN: 1751-1577. DOI: https://doi.org/10.1016/j.joi.2021.101226. URL: https://www.sciencedirect.com/science/article/pii/S1751157721000973.

[7] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3.null (Mar. 2003), pp. 993–1022. ISSN: 1532-4435.

[9] Vincent D. Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008. DOI: 10.1088/1742-5468/2008/10/p10008.

[10] K. W. Boyack and R. Klavans. "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?" In: *Journal of the American Society for Information Science and Technology* 61.12 (2010), pp. 2389–2404. DOI: 10.1002/asi.21419.

[11] Kevin W. Boyack. "Investigating the effect of global data on topic detection". In: *Scientometrics* 111.2 (2017), pp. 999–1015. ISSN: 1588-2861. DOI: 10.1007/s11192-017-2297-y. URL: https://doi.org/10.1007/s11192-017-2297-y.

[12] Jason Cory Brunson. "ggalluvial: Layered Grammar for Alluvial Plots". In: *Journal of Open Source Software* 5.49 (2020), p. 2017. DOI: 10.21105/joss.02017.

[13] Jason Cory Brunson and Quentin D. Read. *ggalluvial: Alluvial Plots in 'ggplot2'*. R package version 0.12.4. URL: http://corybrunson.github.io/ggalluvial/.

[14] François Claveau and Yves Gingras. "Macrodynamics of Economics: A Bibliometric History". In: *History of Political Economy* 48.4 (2016), pp. 551–592. ISSN: 0018-2702. DOI: 10.1215/00182702-3687259. URL: https://doi.org/10.1215/00182702-3687259.

[15] Cristian Colliander. "A novel approach to citation normalization: A similarity-based method for creating reference sets". In: *Journal of the Association for Information Science and Technology* 66.3 (2015), pp. 489–500. ISSN: 2330-1635. DOI: https://doi.org/10.1002/asi.23193. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23193.

[16] José Edwards, Yann Giraud, and Christophe Schinckus. "A quantitative turn in the historiography of economics?" In: *Journal of Economic Methodology* 25.4 (2018), pp. 283–290. DOI: 10.1080/1350178X.2018.1529133. eprint: https://doi.org/10.1080/1350178X.2018.1529133. URL: https://doi.org/10.1080/1350178X.2018.1529133.

[17] Joshua Eykens, Raf Guns, and Tim C. E. Engels. "Fine-grained classification of social science journal articles using textual data: A comparison of supervised machine learning approaches". In: *Quantitative Science Studies* 2.1 (2021), pp. 89–110. ISSN: 2641-3337. DOI: 10.1162/qss_a_00106. URL: https://doi.org/10.1162/qss_a_00106.

[18] Nicholas Fisher. "The Classification of the Sciences". In: *Companion to the History of Modern Science*. Ed. by R.C. Olby. London: Routledge, 1990, pp. 853–868.

[19] Carlos García, Daniel Otero, and Boris Salazar. "The Drifting Influence of Hall's Random-Walk Hypothesis on Consumption Modeling". In: *History of Political Economy* 55.1 (Feb. 2023), pp. 103–143. ISSN: 0018-2702. DOI: 10.1215/00182702-10213653. eprint: https://read.dukeupress.edu/hope/article-pdf/55/1/103/1793411/103garcia.pdf. URL: https://doi.org/10.1215/00182702-10213653.

[20] Wolfgang Glänzel and András Schubert. "A new classification scheme of science fields and subfields designed for scientometric evaluation purposes". In: *Scientometrics* 56.3 (2003), pp. 357–367. ISSN: 1588-2861. DOI: 10.1023/A:1022378804087. URL: https://doi.org/10.1023/A:1022378804087.

[21] Wolfgang Glänzel and Bart Thijs. "Using hybrid methods and 'core documents' for the representation of clusters and topics: the astronomy dataset". In: *Scientometrics* 111.2 (2017), pp. 1071–1087. ISSN: 1588-2861. DOI: 10.1007/s11192-017-2301-6. URL: https://doi.org/10.1007/s11192-017-2301-6.

[22] P. Jaccard. "The distribution of the flora in the alpine zone". In: *New Phytologist* 11 (1912), pp. 37–50.

[23] Frizo Janssens, Wolfgang Glänzel, and Bart De Moor. "A hybrid mapping of information science". In: *Scientometrics* 75.3 (2008), pp. 607–631. ISSN: 1588-2861. DOI: 10.1007/s11192-007-2002-7. URL: https://doi.org/10.1007/s11192-007-2002-7.

[24] M. M. Kessler. “Comparison of the results of bibliographic coupling and analytic subject indexing”. In: *American Documentation* 16.3 (1965), pp. 223–233. ISSN: 0096-946X. DOI: https://doi.org/10.1002/asi.5090160309. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090160309.

[25] Richard Klavans and Kevin W. Boyack. “Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge?” In: *Journal of the Association for Information Science and Technology* 68.4 (2017), pp. 984–998. ISSN: 2330-1635. DOI: https://doi.org/10.1002/asi.23734.

[26] R. Kleminski, P. Kazienko, and T. Kajdanowicz. “Analysis of direct citation, co-citation and bibliographic coupling in scientific topic identification”. In: *Journal of Information Science* (2020). DOI: 10.1177/0165551520962775.

[27] Maria Cristina Marcuzzo et al. “Cambridge as a Place in Economics”. In: *History of Political Economy* 40.4 (Nov. 2008), pp. 569–593. ISSN: 0018-2702. DOI: 10.1215/00182702-2008-027. eprint: https://read.dukeupress.edu/hope/article-pdf/40/4/569/428966/HOPE404-01marcuzzo.pdf. URL: https://doi.org/10.1215/00182702-2008-027.

[28] Chaoqun Ni, Cassidy R. Sugimoto, and Jiepu Jiang. “Venue-author-coupling: A measure for identifying disciplines through author communities”. In: *Journal of the American Society for Information Science and Technology* 64.2 (2013), pp. 265–279. ISSN: 1532-2882. DOI: 10.1002/asi.22630. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22630.

[29] Marek Omelka and Sarka Hudecová. “A comparison of the Mantel test with a generalised distance covariance test”. In: *Environmetrics* 24.7 (2013), pp. 449–460. DOI: 10.1002/env.2238.

[30] E. Petrovich. “Science Mapping”. In: *ISKO Encyclopedia of Knowledge Organization* (2020). URL: https://www.isko.org/cyclo/science_mapping.

[31] Ashwani Saith. “The Cambridge Journal of Economics – A Forum of One’s Own”. In: *Review of Political Economy* 35.1 (2023). doi: 10.1080/09538259.2022.2104027, pp. 28–49. ISSN: 0953-8259. DOI: 10.1080/09538259.2022.2104027. URL: https://doi.org/10.1080/09538259.2022.2104027.

[32] Jacques Savoy. *Machine learning methods for stylometry*. Cham: Springer, 2020.

[33] Peter Sjögårde and Per Ahlgren. “Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics”. In: *Journal of Informetrics* 12.1 (2018), pp. 133–152. ISSN: 1751-1577. DOI: https://doi.org/10.1016/j.joi.2017.12.006. URL: https://www.sciencedirect.com/science/article/pii/S1751157717303371.

[34] Peter Sjögårde and Per Ahlgren. "Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties". In: *Quantitative Science Studies* 1.1 (2020), pp. 207–238. ISSN: 2641-3337. DOI: 10.1162/qss_a_00004. URL: https://doi.org/10.1162/qss_a_00004.

[35] Henry Small. "Co-citation in the scientific literature: A new measure of the relationship between two documents". In: *Journal of the American Society for Information Science* 24.4 (1973), pp. 265–269. ISSN: 0002-8231. DOI: https://doi.org/10.1002/asi.4630240406. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.4630240406.

[36] Gábor J. Székely and Maria L. Rizzo. "Partial distance correlation with methods for dissimilarities". In: *Annals of Statistics* 42.6 (Dec. 2014), pp. 2382–2412. DOI: 10.1214/14-AOS1255.

[37] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. "Measuring and Testing Dependence by Correlation of Distances". In: *The Annals of Statistics* 35.6 (2007), pp. 2769–2794.

[38] Andreas Thor et al. "Introducing CitedReferencesExplorer (CRExplorer): A program for reference publication year spectroscopy with cited references standardization". en. In: *Journal of Informetrics* 10.2 (May 2016), pp. 503–515. ISSN: 17511577. DOI: 10.1016/j.joi.2016.02.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S1751157716300049 (visited on 02/21/2019).

[39] Roberto Todeschini and Alberto Baccini. *Handbook of Bibliometric Indicators. Quantitative Tools for Studying and Evaluating Research.* Weinheim (Germany): Wiley-VCH, 2016. ISBN: 978-3-527-33704-0.

[40] Alexandre Truc, François Claveau, and Olivier Santerre. "Economic methodology: a bibliometric perspective". In: *Journal of Economic Methodology* 28.1 (2021), pp. 67–78. DOI: 10.1080/1350178X.2020.1868774. eprint: https://doi.org/10.1080/1350178X.2020.1868774. URL: https://doi.org/10.1080/1350178X.2020.1868774.

[41] Bo Wang et al. "Unsupervised metric fusion by cross diffusion". In: *IEEE Conference on Computer Vision and Pattern Recognition* (June 2012), pp. 2997–3004. DOI: 10.1109/CVPR.2012.6248029.

[42] Bo Wang et al. "Similarity network fusion for aggregating data types on a genomic scale". In: *Nature methods* 11 (Jan. 2014). DOI: 10.1038/nmeth.2810.

[43] M. Zitt et al. "Bibliometric delineation of scientific fields". In: *Springer Handbook of Science and Technology Indicators*. Ed. by W. Glänzel et al. Springer, 2019, pp. 25–68. DOI: 10.1007/978-3-030-02511-3_2.

A Supplementary figures

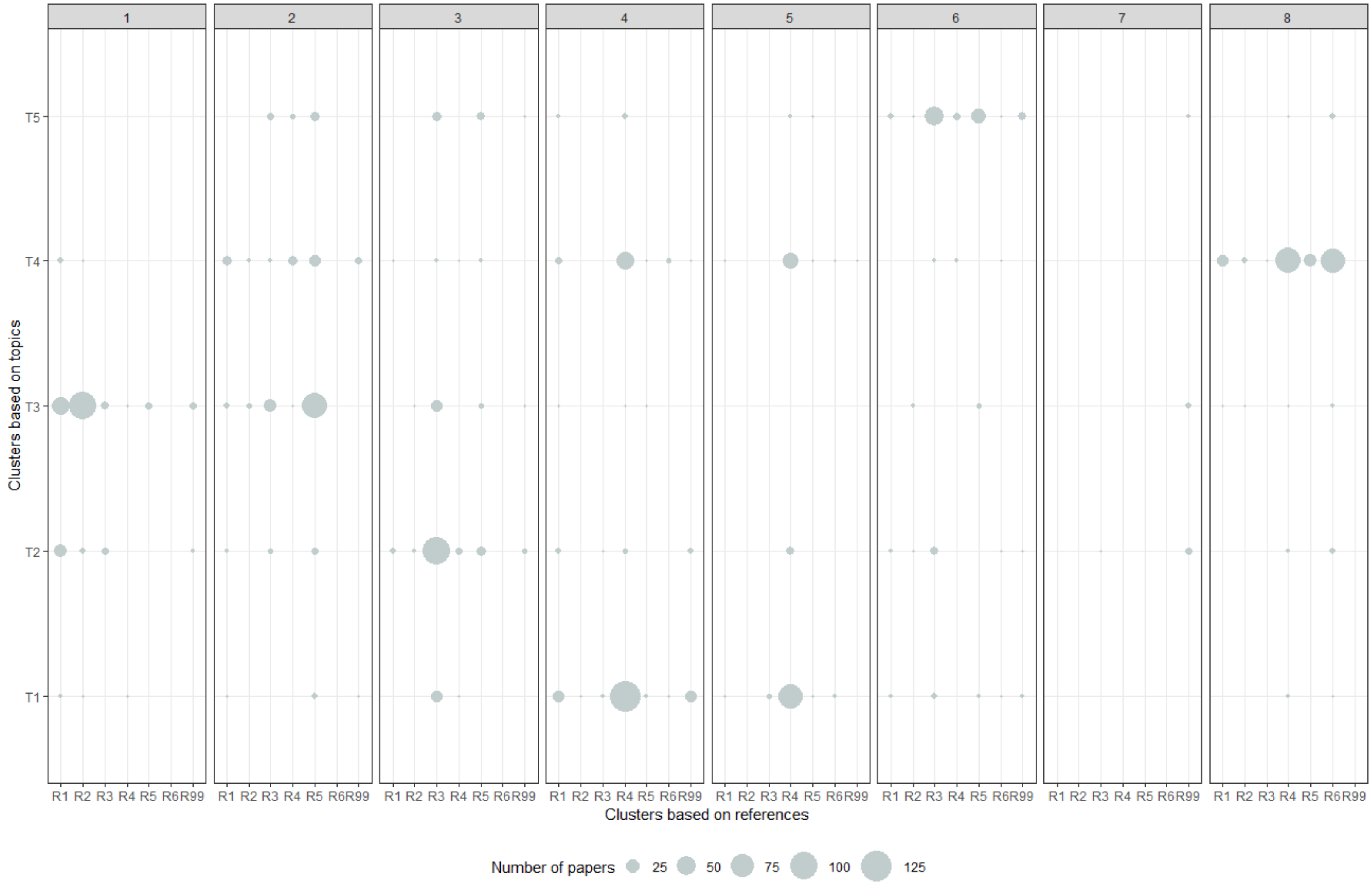

Figure A.1: Cross-distribution of articles from the *Cambridge Journal of Economics* in different clusters. Each panel represents one of the 8 clusters obtained by Louvain algorithm applied to the Fused_20 network. On the y-axis the 5 clusters obtained in the Topics_20 network are reported; on the x-axis the clusters obtained in the Cited references network. Size of points is proportional to the number of papers.