Journal of Informetrics 13 (2019) 354–386

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

Scientific community detection via bipartite scholar/journal graph co-clustering

Chiara Carusi*, Giuseppe Bianchi

Electronic Engineering Department, University of Rome "Tor Vergata", Italy

ARTICLE INFO

Article history:
Received 7 October 2018
Received in revised form 6 January 2019
Accepted 10 January 2019
Available online 4 February 2019

Keywords:
Community detection
Scientometrics
Clustering
Bipartite Graph analysis

ABSTRACT

This paper stems from the observation that researchers in different fields tend to publish in different journals. Such a relationship between researchers and journals is quantitatively exploited to identify scientific community clusters, by casting the community detection problem into a co-clustering problem on bipartite graphs. Such an approach has the potential of leading not only to the fine- grained detection of scholar communities based on the similarity of their research activity, but also to the clustering of scientific journals based on which are the most representative of each community. The proposed methodology is purely data-driven and completely unsupervised, and does not rely on any semantics (e.g. keywords or a-priori subjective categories). Moreover, unlike "flat" data structures (e.g. collaboration graphs or citation graphs) our bipartite graph approach blends in a joint structure both the researcher's attitude and interests (i.e., freedom to select the venue where to publish) as well as the community's recognition (i.e., acceptance of the publication on a target journal); as such may perhaps inspire further scientometric evaluation strategies. Our proposed approach is applied to the Italian research system, for two broad areas (ICT and Microbiology&Genetics), and reveals some questionable aspects and community overlaps in the current Italian scientific sectors classification.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

The problem of classification of science, perhaps "as old as science itself" (Glänzel & Schubert, 2003), is not only a speculative topic, but comes along with several pragmatic implications in terms of scholars' assessment and relevant carriers, as well as public funding and sharing of grants. Indeed, publication activity, citation habits, and perhaps even scientific productivity considerably differ among different research areas (Abramo & D'Angelo, 2011). Therefore, an appropriate classification of homogeneous scientific fields - and in turns communities of scholars - is a crucial precondition for a valid and fair scientometric evaluation of research (Glänzel & Schubert, 2003).

Taking one step further, many bibliometric maps have been proposed in scientometric literature (Boyack, Klavans, & Börner, 2005; Klavans & Boyack, 2009; Leydesdorff & Rafols, 2009; Moya-Anegón et al., 2004; Rosvall & Bergstrom, 2008; van Eck & Waltman, 2009), attempting not only to detect communities but also to provide a broader view of the "distance" of scientific fields, their structure and dynamics. A very natural input for such a classification effort is the information that can be directly or indirectly gathered by the scholarly publication system and the relevant bibliometric relationships (e.g. publications, citations, coauthorship). These networks are analysed in order to capture meaningful properties of the

* Corresponding author.
E-mail addresses: Chiara.Carusi@uniroma2.it (C. Carusi), Giuseppe.Bianchi@uniroma2.it (G. Bianchi).

https://doi.org/10.1016/j.joi.2019.01.004
1751-1577/© 2019 Elsevier Ltd. All rights reserved.

underlying research system, and are often used to determine the influence of bibliometric units like scholars or journals. Besides the immediate application in scientometric evaluation, "maps of science" are of general interest and have many potential applications in research policy issues (Noyons, 1999; Noyons & Calero-Medina, 2009). Disciplinary maps can indeed help university administrators and R&D managers understand their organization and support the decision-making process (Boyack et al., 2005), suggesting areas of opportunity, the amount of overlap (Noyons & Calero-Medina, 2009) in research and potential collaborations (Calero, Buter, Cabello Valdés, & Noyons, 2006) across different universities or even at a faculty level (Noyons & Calero-Medina, 2009). Understanding the inherent structure of the scientific landscape would make it easier not only to monitor research activity, but also to investigate about the future of science (Barabâsi et al., 2002).

Our specific interest for classification is also fueled by the practical importance that such a classification plays in the Italian academic system (a quite unique system in the European landscape, as discussed in Abramo, D'Angelo, and Murgia (2013)) in terms of promotions, call for positions, funding, and so on. The official scholar classification system used in Italian universities is currently top-down provided by the Italian Ministry of Education, University and Research. It comprises 367 "scientific disciplinary sectors" (SDS), aggregated first into 88 macro-sectors and then into 14 high-level areas.[1] According to this system, each scholar or researcher working in an Italian university must belong to one (and only one) specific scientific disciplinary sector, which should best reflect their expertise and main field of research. When it comes to the assessment of research productivity, choosing an appropriate classification system plays a crucial role, considering that, as a matter of fact, quantitative indicators such as the number of publications and citations vary significantly among different scientific disciplines. But to what extent is this "a-priori" classification consistently reliable? And how would it compare with respect to scientific communities inferred from automated analysis of suitable bibliometric networks? Other than providing more general insights, we believe that the methodologies and analyses presented in the remainder of the paper contribute to unveil some serious limitations and questionable aspects of the currently deployed official Italian categories.

Approach and contribution

Traditionally, research on bibliometric networks has usually focused either on citation networks or on collaboration networks, where the observation units are journals (Archambault, Beauchesne, & Caruso, 2011; Glänzel & Schubert, 2003; Leydesdorff & Rafols, 2009) linked together by bibliographic references or, respectively, scholars sharing authorship of articles (Calero et al., 2006; Newman, 2001, 2004b; Perianes-Rodríguez, Olmeda-Gómez, & Moya-Anegón, 2010; Rodriguez & Pepe, 2008). More recently, many classifications of science have also been proposed at the level of individual publications (Šubelj, van Eck, & Waltman, 2016; Boyack et al., 2005; van Eck & Waltman, 2017; Waltman & Eck, 2012), leveraging citation relations between papers.

All these bibliometric networks are however usually based on direct relations between publications, respectively in the form of citation between journals/papers or in the form of coauthorship between scholars, thus working very well when the aim of the analysis is to build a map of science or to identify groups of researchers working together. They are less suitable for our problem, since we want to group together scholars working on the same subject even in the case where they have not shared authorship of a paper or where they have not cited each other's work. Classification systems based on references can indeed be affected by the variability in citation habits among subfields (Archambault et al., 2011; Glänzel & Schubert, 2003), while coauthorship patterns are often driven by departmental and institutional affiliation (Rodriguez & Pepe, 2008).

Starting from these observations, in this paper we propose a different type of bibliometric network. A first underlying assumption behind this work is that communities of scholars working on the same subject can be detected looking at the distribution of their publications across different scientific journals, since scholars that typically publish their papers in the same journals are likely to be working on similar research topics. We find especially compelling the fact that a paper's acceptance to a journal denotes not only interest from the author (which is free to choose the journal, and hence the topic areas), but also indirectly certifies the author' skills, and the explicit recognition that the work is of interest and valuable for the relevant community – the author cannot decide to have her own paper accepted but the editor and the peer reviewers, arguably experts in the topic areas relevant to the journal, have to rate as adequate and appropriate the level of contribution of the paper in order to grant acceptance! In essence, we believe that the author-journal relation might be an harder to bias link for community classification than other links (such as citation or co-authorship) which are mainly influenced by the authors' freedom. We find that the same kind of author-journal bipartite network has also been used in Minguillo (2010), in that case to identify the core and periphery journals for the Spanish library and information science community according to the interaction of authors around journals. With respect to Minguillo (2010), in the present study we further leverage the duality of such two-mode network to propose a co-clustering methodology for both the author and the journal node sets. Indeed, the second reason which motivates our proposed bipartite author-journal graph approach is that the same structure and relation can be used not only to detect communities of authors, but also infer which are the representative journals for each research community in a pairwise manner. In other words, whereas most of science classification systems proposed in literature address the one-way clustering problem of scholar or journal classification, the methodology proposed in this paper naturally takes into account both these levels at the same time, leading to a unique classification system where scholar clusters and journal clusters are linked together in a pairwise manner.

In short, the contribution of this paper is the following:

[1] The classification system is available at https://www.cun.it/uploads/4079/Allegato_CAcademicFieldsandDisciplines.pdf?v=

1 we propose a new bibliometric approach based on the idea that academic activity is best described when looking at scholars and academic journals *simultaneously*, and which leverages a spectral technique to *jointly* address both the challenge of community detection for scholars and the problem of journal classification;
2 we analyze the research map emerging in two selected areas, the ICT and Microbiology&Genetics research fields, along with a multiscale analysis which highlights how the high-level research map breaks up into lower-level research trends when working at a different resolution;
3 we provide quantitative evidence that, at least for the two areas specifically investigated in the paper, the official scholar classification currently used in the Italian national academic system is characterized by non marginally overlapping sectors, thus making us question whether such an a-priori classification (which we recall is an exception in the European landscape) should not be abandoned in favor of an unsupervised research community classification based solely on the structural properties of a bibliometric network.

2. Methodology

The two datasets specifically used in this work collect information about academic publications retrieved from the Scopus database (see Appendix), and authored by at least an Italian Faculty member listed in the national "Cineca" database[2] over a reference period of time (we non-restrictively chose the sixteen years period 2000-2016). For each paper, we collected information on its title, on its author(s), on the journal it was published in and on its publication date. We excluded from our analysis conference papers, and we processed the data, cleared ambiguities, and cleaned the data as detailed in the Appendix.

For building the datasets, we chose two broad scientific areas. The first one, information and communication technology (ICT, specifically analyzed in Section 3), was selected because it is the authors' area of expertise. As such, knowing the relevant scientific communities, it was relatively easy for us to interpret/justify the findings. Our second selected example area was microbiology and genetics; this area was chosen because of its complex nature, with scholars involved from very different disciplines (medicine, biology, chemistry, and even agriculture!) and therefore being its classification in communities a quite challenging goal (see Section 4).

The final and cleaned version of the dataset covers a total 47 718 papers published in 1454 journals by 2582 scholars for ICT, and a total of 61 950 papers published in 2099 journals by 2111 scholars for the Microbiology&Genetics area. Suitably aggregating the information retrieved from the Scopus database, a scholar-by- journal *publication matrix A* can be defined for each example area, setting each entry of the matrix $a_{k,l}$ equal to the total number of papers that scholar k has published in journal l within the reference period of time.

2.1. The scholar-journal bipartite network

As discussed in the introductive section, in this paper we focus our analysis on a particular type of bibliometric network, to which we will refer as the *scholar-journal network*, where nodes represent scholars and journals linked together whenever a scholar has her paper(s) published in a scientific journal. Due to the way it is defined, the scholar-journal network displays the following properties:

- it is *bipartite* - since edges represent publications by an author in a journal, there are no direct links between two scholars or between two journals;
- it is *undirected* - even though it could seem intuitive to think of publications as directed edges from scholars to journals, in this specific network we want to capture the duality of the relation between the two sets of nodes;
- it is *weighted* - by putting positive weights on the edges, we capture information on how many papers a particular scholar has published in the same journal (for our analysis we need to take into account not only where scholars publish their papers, but also their exact publication distribution among those journals).

Formally, the scholar-journal network can be represented as a graph $\lfloor G = (\lfloor S, \lfloor J, \lfloor E)$ where $\lfloor S = s_1, \ldots, s_n$ and $\lfloor J = j_1, \ldots, j_m$ are respectively the set of scholars and the set of journals they have published their papers in, whereas $\lfloor E$ is the set of edges $\{(s_k, j_l) : s_k \in \lfloor S, j_l \in \lfloor J\}$, weighted according to the number of papers scholar s_k has published in journal j_l. Starting from the $n \times m$ scholar-by-journal publication matrix A, which accounts for the total number of papers each scholar has published in each journal, the $(n+m) \times (n+m)$ adjacency matrix of $\lfloor G$ can be written in the following form:

$$M = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$$

where we have ordered the vertices so that the first n vertices index the scholars (node set $\lfloor S$) while the last m index the journals (node set $\lfloor J$).

[2] The Cineca database is freely accessible at *http://cercauniversita.cineca.it/php5/docenti/cerca.php*

In the next subsection we will derive a simultaneous clustering for both scholars and journals resorting to a bipartite spectral graph partitioning approach, originally proposed in Dhillon (2001) to conduct co-clustering of words and documents in a document collection. Even though the application is very different, this latter problem presents many aspects in common with ours. In fact, the basic premise behind the methodology of Dhillon (2001) lies in the observation that word clustering induces document clustering and vice versa; in the present work we will leverage the exact same duality in the publication relation between scholars and journals. Results obtained on the ICT scholar-journal network are provided in the next subsection to better illustrate the methodology. The ICT publication graph $\lfloor G$ is composed of $n = |\lfloor S| = 2582$ scholar vertices, $m = |\lfloor J| = 1454$ journal vertices and $|\lfloor E| = 36\,679$ edges. [3] .

2.2. A spectral approach

A wide variety of methods have been proposed in the literature to detect communities in social networks (Barber, 2007; Blondel, Guillaume, Lambiotte, & Lefebvre, 2008; Girvan & Newman, 2002; Lambiotte, Delvenne, & Barahona, 2008; Newman, 2013; Palla, Derényi, Farkas, & Vicsek, 2005; Radicchi, Castellano, Cecconi, Loreto, & Parisi, 2004; Rosvall & Bergstrom, 2008; Traag, Waltman, & van Eck, 2018). Among them, spectral techniques (Newman, 2013; Ng, Jordan, & Weiss, 2002) revolve around a change in representation based on the eigenvectors of a suitable matrix (like the Laplacian of the graph), which enhances the inherent structure of the data, making potential clusters more evident and easy to detect than in the original space (Fortunato, 2010). We will now recall some basic concepts of spectral graph theory which are needed for our analysis. We refer the reader to the very good review paper by Von Luxburg (2007) for a more complete overview on spectral clustering.

Let us start with the problem of partitioning into two disjoint sets A and B the vertex set $\lfloor V$ of a generic graph $\lfloor G = (\lfloor V, \lfloor E)$. The total weight of the edges that have to be removed in order to break the original graph into two separate connected components with vertex set A and B is called the *cut* of the bipartition $\{A, B\}$:

$$cut(A, B) = \sum_{u \in A,\, v \in B} w(u, v) \tag{1}$$

The cut of a bipartition inversely accounts for the degree of dissimilarity between the two vertex sets; along with this interpretation, the optimal bipartitioning of a graph is the one that minimizes this cut value. However, it has been largely pointed out that the minimum-cut criteria often leads in practice to unsatisfactory solutions which simply isolate one or a small group of individual vertices from the rest of the graph. This is an undesirable effect, since in most clustering problems we expect clusters to be reasonably large. Starting from this consideration, Shi and Malik (2000) propose an alternative, unbiased measure, named the *normalized cut*, obtained by normalizing the cut with respect to the total graph edge weight:

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)} \tag{2}$$

where $assoc(A, V) = \sum_{u \in A, t \in V} w(u, t)$ is the total edge weight from nodes in A to all nodes in the graph. Unfortunately, the great downside of the normalized-cut criteria is that the balancing condition introduced in the new formulation of equation (2) makes the original, easily tractable min-cut problem hard to solve. In fact, it has been demonstrated that the discrete optimization problem of finding the minimum normalized cut of a graph is NP- complete (Shi & Malik, 2000). This is where spectral clustering techniques come in, providing a solution to a relaxed version of the norm-cut criteria. Indeed relaxing the discrete costraint on the solution, thus casting the norm-cut problem in the real value domain, an approximate solution z can be found efficiently solving the following generalized eigensystem:

$$(D - W)z = D\lambda z \tag{3}$$

where D is a diagonal matrix with diagonal entries equal to each vertex degree $d_i = \sum_j w(i, j)$, and W is the symmetrical, weighted adjacency matrix with entries $W(i, j) = w_{i,j}$ (we also refer to the matrix $L = D - W$ as the *unnormalized graph Laplacian*). It has been shown that the normalized cut problem has a real-valued solution in the eigenvector of the generalized eigensystem (Eq. (3)) associated to its second smallest eigenvalue.

In Section 2.1 we have modeled our publication collection as a bipartite graph between scholars and journals, which permits us to cast the problem of simultaneous clustering of scholars and journals into a *bipartite* graph partitioning problem. To solve this particular partitioning problem we will now follow the methodology presented in Dhillon (2001) , with some minor differences which will be discussed in the remainder of this section. The co-clustering algorithm proposed in Dhillon (2001), in fact, extends the classic spectral approach which we have briefly outlined, used to find partitions in one- mode networks, to the different case of bipartite graphs. As for one-mode graphs, also in the bipartite case we look for the second

[3] The number of edges is lower than the number of publications collected since the same scholar may have published more than one paper in the same journal.

smallest eigenvalue of the generalized eigenvalue problem $Lz = \lambda Dz$ which provides a relaxation to the normalized cut problem. In the bipartite case, however, matrices L and D present a block structure which can be written as follows:

$$L = \begin{bmatrix} D_1 & A \\ A^T & D_2 \end{bmatrix} D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \tag{4}$$

where D_1 and D_2 are diagonal matrices such that $D_1(i,i) = \sum_j a_{ij}$ and $D_2(j,j) = \sum_i a_{ij}$.

Letting $z = \begin{bmatrix} x, & y \end{bmatrix}^T$ and $u = D_1^{1/2}x$, $v = D_2^{1/2}y$, Dhillon (2001) shows how solving the generalized eigenvalue problem $Lz = \lambda Dz$ is equivalent to performing a singular value decomposition (SVD) of the normalized matrix

$$A_n = D^{-1/2}AD^{-1/2} \tag{5}$$

where u and v are respectively the left and right singular vectors, and $\sigma = (1 - \lambda)$ is the corresponding singular value. This means that, instead of looking for the second smallest eigenvalue λ_2 of $Lz = \lambda Dz$ and its relative z_2 eigenvector, one can (more easily) compute the second largest singular value σ_2 of A_n and its associated left and right singular vectors u_2 and v_2. The same reasonings apply for the eigenvector corresponding to the third smallest eigenvalue of Eq. (3), which can be shown to optimally partition the two subgraphs already induced by the second eigenvalue. Hence, using the eigenvector with the next smallest eigenvalue, one can iteratively subdivide the previous graph partition, exponentially increasing the number of detectable clusters - if we consider the smallest l eigenvalues, the original graph can be split up to 2^l subgraphs.

Now that we have cast the partitioning problem into a SVD framework, the duality in the relation between scholars and journals is more evident since co-clustering reduces to finding the left and right singular vectors of the appropriately scaled adjacency matrix A_n. The singular value decomposition of A_n maps the two node set of the scholar- journal graph into an homogeneous set of points in the low-dimensional space induced by the eigenvectors associated to the largest non-trivial singular values of A_n (Fig. 1a). The spatial coordinates of scholars and journals according to the new basis can be represented as a real valued matrix Z, whose columns correspond to the concatenation of the left and right singular vectors, conveniently "unnormalized":

$$Z = [z_2, z_3, \ldots z_{l+1}] = \begin{bmatrix} D_1^{-1/2}U \\ D_2^{-1/2}V \end{bmatrix} \tag{6}$$

and where $U = [u_2, u_3, \ldots u_{l+1}]$ and $V = [v_2, v_3, \ldots v_{l+1}]$.

The singular vectors of the normalized publication matrix A_n form an alternative, lower-dimensional basis which best represents similarities and differences between the nodes of the scholar-journal graph. Fig. 1a shows a clear clustering structure for scholars and journals, which can now be detected much more easily in the new representation. Our final step is now to apply a suitable clustering algorithm to extract a partition of the set of points from the SVD space. Noting that, as observed by Von Luxburg (2007), there seems to be no principled constraint on the clustering algorithm to use in this step, we have opted for a Spherical K-Means (Dhillon & Modha, 2001) instead of the classical K- Means, for two reasons. The first reason relies on the particular shape of the points visualized in Fig. 1a: since K-Means attempts to group samples into cluster of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares, it poorly captures elongated clusters like the ones represented in Fig. 1a. Second, and more related to our specific application' semantics, our goal is to co-cluster authors and journals on the basis of how publications *distribute* across journals, regardless of the total number of their publications, i.e., whether a scholar displays a more or less intense publishing activity. Such goal is better captured by a "spherical" variant of K-Means, since (i) all vectors are normalized and (ii) cosine dissimilarity $d(x,y) = 1 - cos(x,y)$ is used instead of the Euclidean distance measure. Indeed, if, using a standard K-means, two authors having a large difference in terms of absolute number of published papers would be distant, this may not anymore be the case with a Spherical K-Means, as the distance depends on the proportion of papers published in the various journals.[4] The same arguments apply for the classification of journals, which are in this way characterized by the interaction of authors around them, rather than by the total number of articles published in them.

In Fig. 1a we highlighted with different colors the $K = 5$ different clusters detected by the Spherical K-Means algorithm,[5] which induces a simultaneous partition of both scholars (circle markers) and journals (triangle markers). The $K = 5$ "concept vectors" (Dhillon & Modha, 2001), i.e. the centroids of the clusters normalized to have unit Euclidean norm, are plotted in red. Observing Fig. 1a it seems somewhat intuitive to look for 5 clusters, but a more principled reason behind this particular choice of the number of clusters K will be discussed in Section 2.3. Note that using Spherical K-means on the transformed dataset is equivalent to using classical K-Means based on Euclidean distance when we project our observations and centroids to the unit sphere at the end of each K- Means step (Fig. 1b).

[4] For an example, in the original space an author having published 10 papers on journal A and 5 on journal B would be at zero cosine distance with respect to a second author having published 2 papers on journal A and 1 on journal B.

[5] We considered as co-clustering results the best output in terms of inertia out of 20 different runs of the Spherical K-Means algorithm, initializing each run with a different centroid seed.

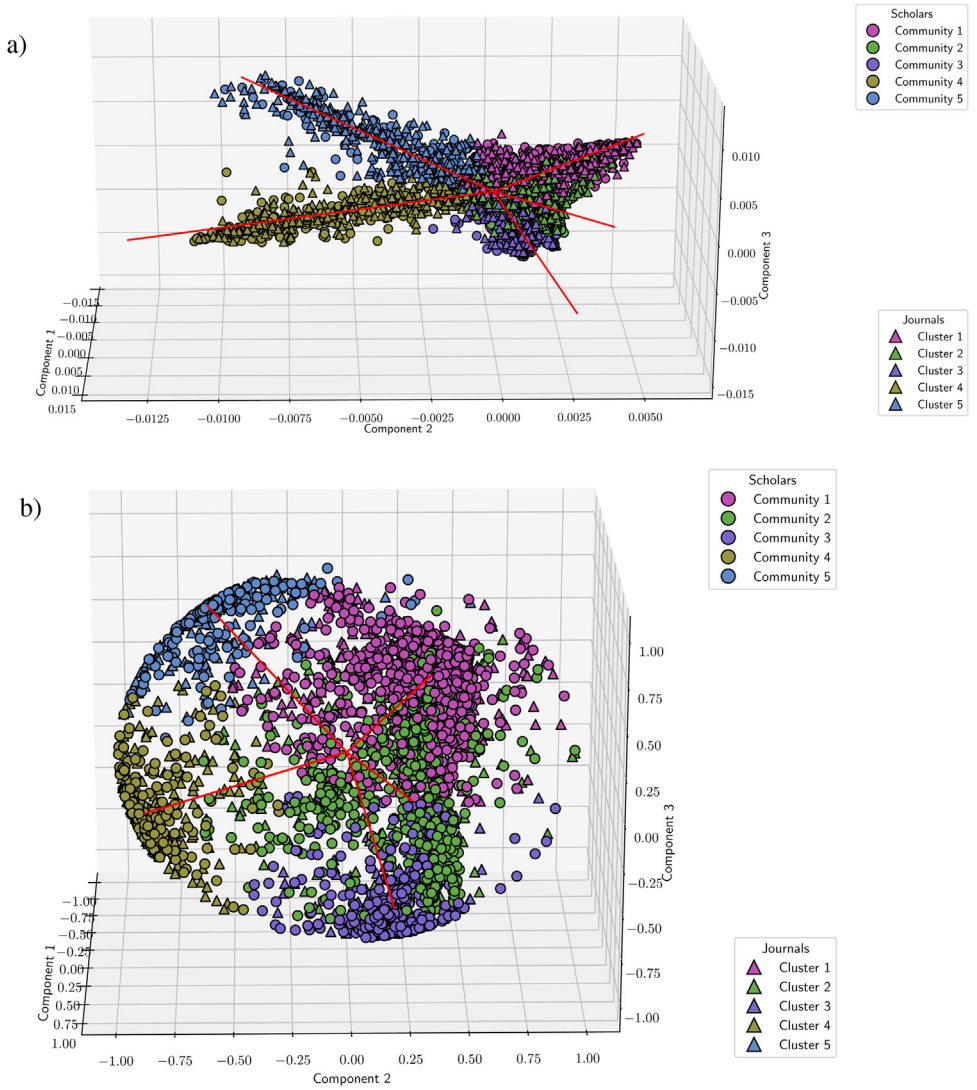


Fig. 1. Singular value decomposition of the normalized matrix A_n.

As a result we obtained a simultaneous partitioning of authors and journals into pairwise clusters. The effectiveness of the discussed co- clustering methodology is visually highlighted in the original basis with Fig. 2, thanks to a suitable graph layout where nodes are located along the x-axis not at random, but according to the different co-clusters they were assigned to. Even though this may not be the optimal layout[6], the graph visualization in Fig. 2 highlights how the majority of edge connections lies within scholars and journals from the same cluster, intuitively confirming the quality of the resulting co-clustering. Some statistics describing the results of the co-clustering algorithm are also reported in Fig. 2. Note that the total weight of edges connecting a scholar and a journal within the same co- cluster amounts to 67723, approximately 85% of the total number of 79 796 publications considered in the ICT dataset.

2.3. Choosing the number of clusters

The methodology proposed so far requires as an explicit input a given number of communities, i.e., the number of clusters K to give to the K-Means clustering algorithm. On one side, we might resort to the pragmatic approach of using a "reasonable" value K, depending on the specific application. In our specific case, for instance, a reasonable choice for K might consist in resorting to a number of communities equal to (or at least comparable with) that currently used in the

[6] To our knowledge, there is no graph layout specifically designed for bipartite graphs which minimizes edge crossing while preserving a spatially evident bipartite structure.

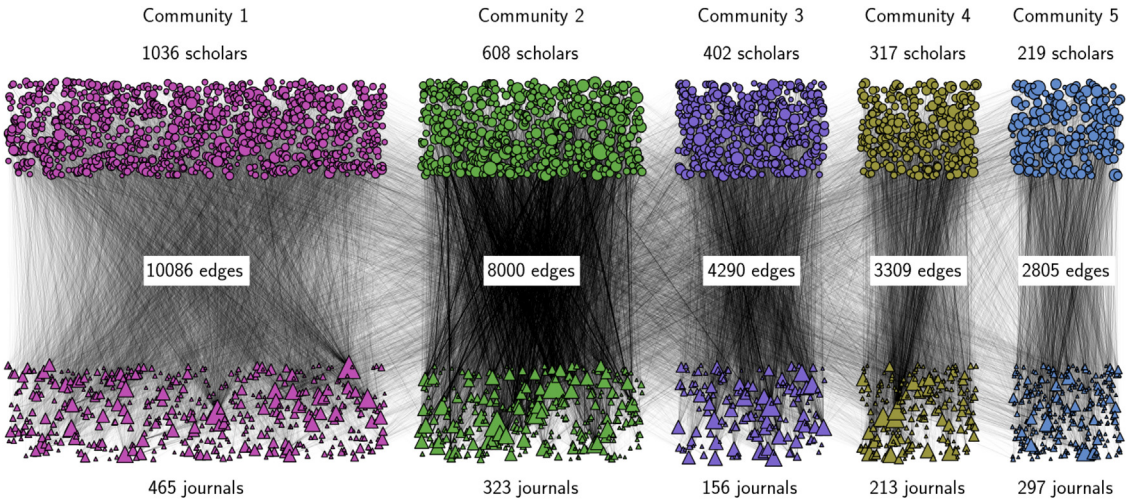


Fig. 2. Two-dimensional representation of the ICT bipartite graph. The position of author and journal nodes along the horizontal axis is set according to the co-clusters they are assigned to, in order to avoid edge crossing and to highlight the effectiveness of the co-clustering method. The size of nodes is proportional to their degree centrality, and the edge thickness is proportional to their associated weight.

Italian academic system – in our specific use-cases, the number of SDSs included in the datasets (8 for ICT and 11 for the Microbiology&Genetics field). This context-based approach would indeed meet our initial motivation for this work, i.e. to assess the classification currently used by the Italian academic system via its comparison with a data-driven benchmark, rather than establishing an exact number of communities.

On the other side, whenever the target number of clusters K is not a priori known (or reasonably guessed), it appears appropriate to resort on an objective methodology specifically devised to identify the optimal configuration (number of clusters) in an unsupervised manner, i.e., via a technique which relies solely on features inherent to the dataset and does not exploit context information. However, to the best of our knowledge, there is no general agreement on a single most effective measure to identify the optimal clustering (Arbelaitz, Gurrutxaga, Muguerza, Pérez, & Perona, 2013; Halkidi, Batistakis, & Vazirgiannis, 2001; Saitta, Raphael, & Smith, 2008). Therefore, in this work we have exploited two different metrics frequently used in literature to help detect communities or evaluate the quality of clustering: the Davies–Bouldin index, based on the geometry of the dataset, and modularity, which looks at the degree properties of the induced graph.

Davies–Bouldin index

The Davies–Bouldin index was proposed for the first time by Davies and Bouldin (1979). To evaluate the quality of a clustering scheme, the Davies–Bouldin index takes into account:

- the scatter S_i within each cluster, which should be as low as possible;
- the separation $M_{i,j}$ between each pair of clusters C_i and C_j, which should be as large as possible.

If we focus on each pair of clusters C_i and C_j within a partition, we can reflect the similarity between the clusters with the following ratio:

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \tag{7}$$

The Davies–Bouldin index is then defined as the average similarity between each cluster and its most similar one:

$$DB \equiv \frac{1}{N}\sum_{i=1}^{N} D_i \quad D_i \equiv \max_{j \neq i} R_{i,j} \tag{8}$$

According to its definition, a lower Davies–Bouldin index indicates a better clustering scheme. Each cluster contribute to the average Davies–Bouldin value with the term D_i, i.e. with their worst-case scenario in terms of similarity to other clusters; the different values taken by each term D_i give an interesting insight on whether all clusters are equally well-defined or not, an aspect which is not expressed by the synthetic DB index itself.

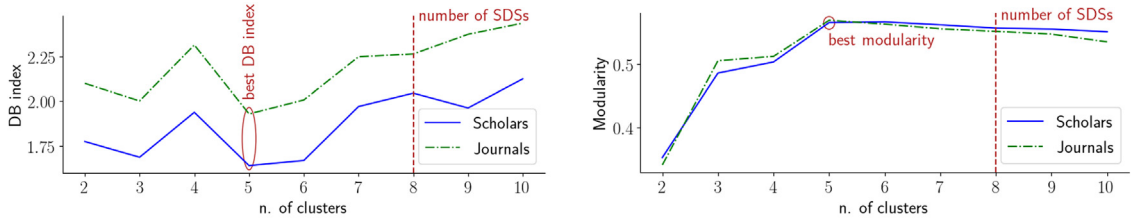


Fig. 3. Values of the Davies–Bouldin index (suplot a) and Barber's modularity (subplot b) for scholar clustering (solid blue line) and journal clustering (dash-dotted green line) for the ICT field, calculated over different numbers of clusters.

To obtain meaningful results, a suitable distance metric should be selected when computing the Davies–Bouldin index, in order to best reflect the kind of similarity we are trying to capture with our clustering scheme. Many distance metrics can be used to define S_i and $M_{i,j}$, depending on the problem at hands; in most cases the Euclidean distance metric is used:

$$S_i = \left(\frac{1}{|C_i|} \sum_{j=1}^{|C_i|} \left| X_j - A_i \right|^2 \right)^{1/2} \qquad M_{i,j} = \left|\left| A_i - A_j \right|\right|_2 = \left(\sum_{k=1}^{K} \left| a_{k,i} - a_{k,j} \right|^2 \right)^{\frac{1}{2}} \tag{9}$$

where A_i is the centroid of cluster C_i, and $a_{k,i}$ is its $k-th$ element; Euclidean distance, however, may not always be the best measure for determining clusters. For the same reasons already discussed for the choice of Spherical K-Means as the clustering scheme, we adapted the classic Davies–Bouldin formulation based on Euclidean distance to a variant based on cosine distance [7] :

$$S_i = \frac{1}{|C_i|} \sum_{j=1}^{|C_i|} \left(1 - \cos(X_j, A_i)\right) \qquad M_{i,j} = \left(1 - \cos(A_i, A_j)\right) \tag{10}$$

Modularity

The concept of modularity of a partition was first introduced by Newman and Girvan (2004). For a given division of network nodes into disjoint communities, modularity is a quality function that measures the density of edges lying inside communities with respect to the expected ratio if edges were randomly distributed, usually according to the actual degree of each node (Blondel et al., 2008; Newman, 2004a). In what follows we will resort to Barber's definition of modularity (Barber, 2007), which extends such concept to the bipartite case:

$$Q = \frac{1}{m} \sum_{ij} \left[\tilde{A}_{ij} - \frac{k_i d_j}{m} \right] \delta(c_i, c_j) \tag{11}$$

where $\tilde{A}$ represents the edge weight between nodes i and j belonging to the two disjoint node sets of the bipartite graph, $k_i = \sum_j \tilde{A}_{ij}$ and $d_j = \sum_i \tilde{A}_{ij}$ are the weighted degrees of nodes i and j, $m = \sum_{ij} \tilde{A}_{ij}$ is the sum of all of the edge weights in the graph, c_i indicate the community of node i and $\delta(\cdot, \cdot)$ is the Kronecker delta function. According to Eq. (11), for each pair of nodes in the same co-cluster, modularity accounts for the difference between the actual edge weight $\tilde{A}_{ij}$ and the expected edge weight $\frac{k_i d_j}{m}$ between nodes. Modularity values fall in the range $[-1, 1]$, and is positive whenever the fraction of edges falling within the same community is higher than expected on the basis of chance. Modularity close to $Q = 1$ indicates a strong community structure, while $Q = 0$ implies that within-community edges are no better than in the case of a randomly composed graph. Good modularity values typically lie in the range between 0.3 to 0.7 (Newman & Girvan, 2004).

Number of ICT communities

We ran the clustering algorithm proposed in Section 2 for different values of K, and then compared the different results according to the two evaluation criteria to identify the choice leading to the best clustering. To avoid the bias introduced by the differences in the total amount of publications among scholars or journals, we defined a Davies–Bouldin index based on cosine similarity. We then calculated the Davies–Bouldin index for the scholar clustering schema over the original publication matrix A, and used the transposed matrix A^T to evaluate journal clustering. Fig. 3a plots the values of the Davies–Bouldin index versus the number of clusters. In the same vein, we computed Barber's modularity for two cases: on the publication

[7] Note that, to obtain meaningful results, the cosine-based Davies–Bouldin index has to be calculated on unit length vectors, and all centroids must be scaled again after computation.

Table 1
List of the scientific disciplinary sectors related to the ICT field, along with the number of associated scholars in the publication dataset, according to the Italian classification system.

SDS code	SDS description	No. of scholars
INF/01	Informatics	738
ING-INF/01	Electronic Engineering	341
ING-INF/02	Electromagnetic Fields	166
ING-INF/03	Telecommunications	328
ING-INF/04	Systems And Control Engineering	267
ING-INF/05	Information Processing Systems	501
ING-INF/06	Electronic And Informatics Bioengineering	119
ING-INF/07	Electrical And Electronic Measurement	122
	Total	**2582**

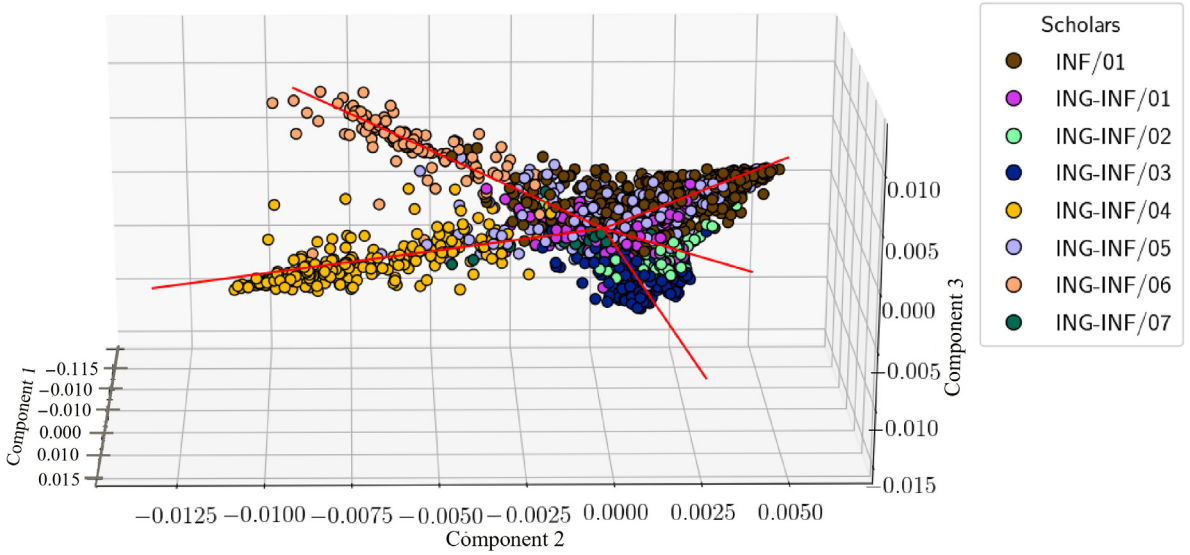


Fig. 4. Representation of scholars for the ICT dataset in the two-dimensional space associated to the two largest non-trivial singular values of A_n. Each node color represents the scholar's scientific disciplinary sector.

matrix normalized along rows (i.e., with respect to scholars) and on the publication matrix normalized along columns (i.e., with respect to journals). Both results are reported in Fig. 3b.

The Davies–Bouldin index shows a similar trend both for scholar and journal clustering, reaching a global minimum for $K = 5$ clusters.[8] For the particular case of ICT, the solution suggested by the Davies–Bouldin index is also confirmed by modularity, which shows a maximum for the same number of co-clusters. Note that such similarity does not occur in general, and that the two metrics will lead to different results for the Microbiology&Genetics case study (see Section 4).

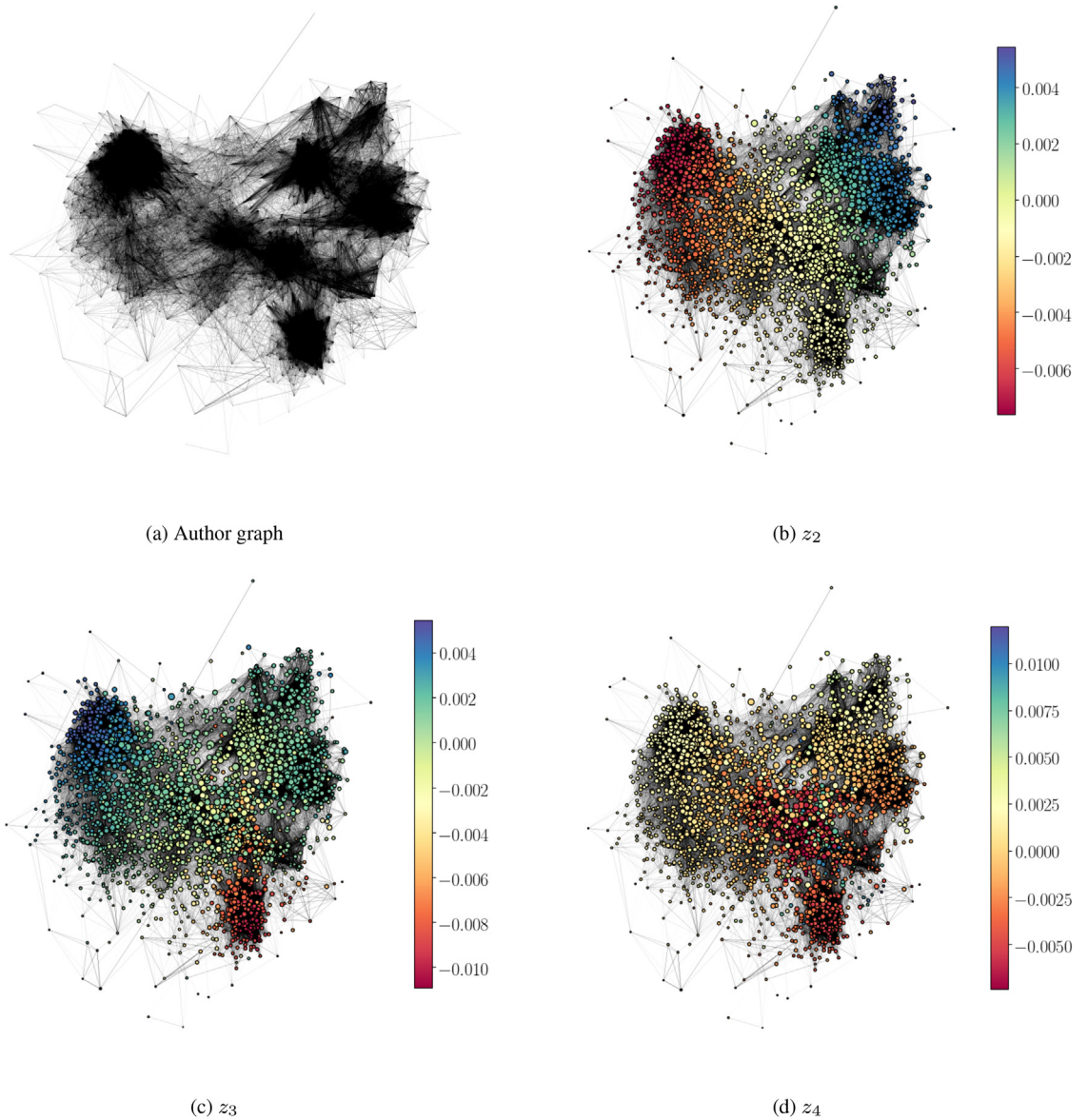
The Davies–Bouldin and modularity criteria will be used in the next sections to validate the quality of our clustering against the official classification system. Nonetheless, results obtained for different values of the number of clusters will be presented in Sections 3 and 4, since the choice made here does not constrain the proposed methodology from applying other types of suitable internal evaluation metrics.

3. ICT analysis

In the present section we will use the data-driven classification of scholars resulting from the application of the methodology presented in Section 2 to assess the quality of the Italian official academic classification in the field of Information and Communication Technology (ICT). In the ICT dataset we collected publications authored by scholars from a set of eight SDSs which covers the vast majority of scholars working in the ICT field (see Appendix). The distribution of scholars present in the ICT dataset across the eight SDSs is reported in Table 1. In Fig. 4 we provide the same SVD representation used in Section 2 for scholar nodes, this time colored according to their scientific disciplinary sector.

In what follows, we are going to present the results of the SDS assessment with the aid of an alternative, visually intuitive graph representation of the scholar communities detected in Section 2. In particular, being the identification of scientific

[8] This explains the choice of using $\lceil log_2(5) \rceil = 3$ singular values of A_n in Section 2.2.



(a) Author graph (b) z_2

(c) z_3 (d) z_4

Fig. 5. (Plot a) Two- dimensional representation of the author graph for the ICT field according to the Fruchterman–Reingold layout. For visualization purposes, we only plotted edges with a weight higher than 0.25. The size of nodes is proportional to their degree centrality, and the thickness of edges is proportional to their associated weight. (Plot b, c, and d) ICT author graph colored according to the value of the three smallest eigenvalues for $Lz = \lambda Dz$. Hot and cold colors represent negative and positive values, respectively.

communities the main focus of this section, we resort to the one-mode projection of the bipartite publication graph G onto the single set of scholar nodes via the cosine similarity metric. The result of this projection is an *author graph* composed of 2582 author nodes and 670 285 edges, weighted according to cosine similarity between the scholars' pubication records (i.e., the rows of the publication matrix A). A two-dimensional representation of the ICT author graph is given in Fig. 5. To minimize edge crossing, the graph nodes were located according to the Fruchterman–Reingold algorithm [9] (Fruchterman & Reingold, 1991), a well-known force-directed layout algorithm also referred to as the "spring layout", since nodes are located in space in order to minimize the energy of a system where the nodes are represented as steel rings and the edges as springs between them. Even though it plays no role in the community detection process (we have drawn our data-driven

[9] An implementation of the Fruchterman-Reingold algorithm is provided by the *spring_layout* function of the *NetworkX* Python package (Hagberg, Swart, & S Chult, 2008).

classification directly from the bipartite graph model),[10] this layout provides a bibliometric map where it is possible to locate the particular position of each author's scholarly activity and its distance from other scholars, giving a simple bird's eye view on the underlying structure of scientific research and visually highlighting tightly-knit modules which largely correspond to the communities detected by the algorithm. In subplots Fig. 5 b, c and d the same author graph is presented in the form of a heatmap where author nodes are colored according to the corresponding elements of z_2, z_3 and $z4$.

To investigate the nature of the detected research structure without resorting to any a-priori knowledge or intuition, we leveraged the co- clustering results obtained in Section 2. We attempted to label each community with a set of representative keywords using the popular text-mining tf-idf scheme on the set of journal titles associated by the co-clustering scheme to that community. The tf-idf metric, short for term frequency-inverse document frequency, was first proposed by Salton and Buckley (1987) and is frequently used in information retrieval to score the importance of words in a document, based on how frequently they appear across a document collection. To get a high score according to tf- idf, a word should display a high frequency in a given document – hence the "term frequency" - but a low frequency in the whole collection of documents, in order to filter out common terms – hence the "inverse document frequency". We ran the tf-idf scheme over a suitably constructed 5-document collection, built concatenating the titles of all journals in the same cluster into a single document.[11] The top ten words for each cluster are reported in Table 2, along with their tf-idf score. In what follows, in order to improve clarity and make the interpretation of co-clustering easier, in tables and figures each pair of scholar and journal communities will be labelled with a selected triple of significant, most representative words.

We now compare the quality of the official SDS classification against the data-driven community scheme. We report in Fig. 6, respectively in the bottom and top plots, a graphical representation of the Davies–Bouldin analysis carried out for the two classification systems. In (8) we defined the Davies–Bouldin index as the average, over all the clusters $i = 1, \ldots, K$, of the similarity R_{i,j^*} between each cluster i and its most similar one j^*. For the sake of analysis, the bar plots in Fig. 6 allow to investigate how clusters differently contribute with their worst-case similarity term $D_i = R_{i,j^*}$ to the average DB value. The average Davies–Bouldin index value is marked with a dashed red line: $DB \approx 1.64$ for data-driven communities and $DB \approx 2.09$ for scientific disciplinary sectors. The Davies–Bouldin analysis suggests that our data-driven methodology based on a network analysis approach leads not only to an overall better clustering result (lower DB value), but also to a more balanced clustering schema (uniform D_i values) compared to the S.D.S. classification currently used in the Italian academic system.

It is interesting to examine how the two classification systems are interrelated. We report in the form of an alluvial graph (Fig. 7) how scholars in different scientific disciplinary sectors redistribute themselves into the 5 data-driven communities (numerical values are provided in Table 3). It can be easily observed that:

- there are three scientific disciplinary sectors, namely *ING-INF/03 - Telecommunications*, *ING-INF/04 - Systems And Control Engineering*, *ING- INF/06 - Electronic And Informatics Bioengineering*, which are basically in a one-to- one correspondence with three specific communities (**remote sensing & telecom**, **control & robotics**, **biomedics**);
- on the contrary, the data-driven classification displays a single, large **knowledge & theoretical logic** community which comprises the great majority of both the *INF/01 - Informatics* and *ING-INF/05 - Information Processing Systems* scientific disciplinary sectors;
- the same applies for the *ING-INF/01 - Electronic Engineering*, *ING-INF/02 - Electromagnetic Fields* and *ING-INF/07 - Electrical And Electronic Measurement* scientific disciplinary sectors, which merge into a single **instrumentation & propagation** community.

A natural question may arise, that is whether or not choosing a higher number of communities as a parameter for the co-clustering algorithm could unveil some of the smaller scientific disciplinary sector categories which are merged together in the 5-community configuration. To address this question, Fig. 8 visually shows how scholars would be reallocated according to the proposed methodology into an increasing number of clusters, unveiling their community structure at different resolution levels. In Fig. 9 we report the same alluvial diagram with a different coloring, to give a visual intuition on how scholars from different scientific disciplinary sectors contribute to the composition of the communities. We provided here only the confusion matrix for the 9-community configuration (Table 4), but all confusion matrices and tf-idf keywords are available in the Supplementary Material. In particular:

- The **remote sensing & telecom**, **control & robotics** and **biomedics** communities, detected in the 5-community configuration, remain essentially unchanged as we increase the number of communities to detect, suggesting that the *ING-INF/03 - Telecommunications*, *ING-INF/04 - Systems And Control Engineering* and *ING-INF/06 - Electronic And Informatics Bioengineering* scientific disciplinary sectors have a clear and robust cluster structure;

[10] A unified approach to mapping and clustering of bibliometric networks, both based on the concept of modularity, is proposed in Waltman, van Eck, and Noyons (2010).

[11] To make the inverse document frequency term more robust against "noisy" publications, in the tf-idf computation we did not consider journals whose papers were published by less than five different scholars from the same co-cluster

Table 2
Top 10 words for journal clusters in the ICT field according to the tf-idf score. Note that, to keep the labeling process totally automatic, data was on purpose not cleaned from synonyms and derived words (e.g., automatic/automatica, medicine/medical), acronyms (e.g. ifac, plos) or non-significant words (e.g., notes, section, papersonline, annual).

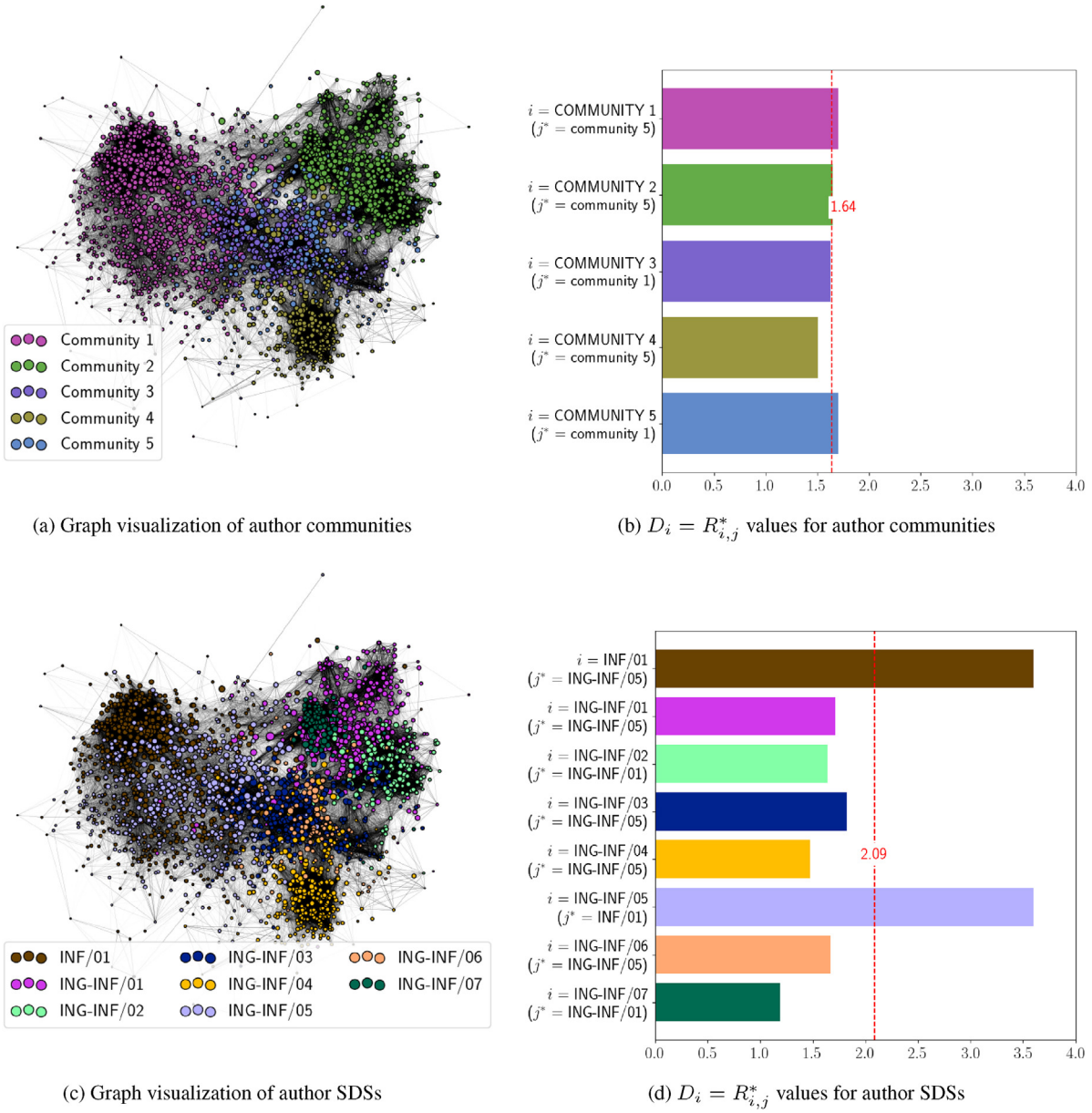
Community 1 Word	TF-IDF score
knowledge	0.015857
theoretical	0.014364
notes	0.013691
logic	0.012384
intelligence	0.011789
pattern	0.010877
fundamenta	0.009348
informaticae	0.009348
software	0.008236
recognition	0.008082

Community 2 **Word**	**TF-IDF score**
instrumentation	0.028343
propagation	0.021387
antennas	0.021387
microwave	0.020758
electron	0.014972
section	0.014376
physics	0.013902
devices	0.013470
measurement	0.012237
instruments	0.011792

Community 3 **Word**	**TF-IDF score**
remote	0.040865
geoscience	0.028530
sensing	0.023265
communications	0.016235
networking	0.013753
wireless	0.012764
vehicular	0.012676
areas	0.009772
satellite	0.005877
observations	0.005706

Community 4 **Word**	**TF-IDF score**
control	0.047702
automatic	0.039202
automatica	0.037065
robotics	0.027734
ifac	0.013033
papersonline	0.013033
robust	0.010322
nonlinear	0.009794
mechatronics	0.009540
fusion	0.009349

Community 5 **Word**	**TF-IDF score**
medicine	0.077507
medical	0.021522
biology	0.021300
biomedical	0.015019
diabetes	0.014893
physiology	0.014393
neuroscience	0.011079
annual	0.010419
rehabilitation	0.010171
plos	0.009444



(a) Graph visualization of author communities

(b) $D_i = R^*_{i,j}$ values for author communities

(c) Graph visualization of author SDSs

(d) $D_i = R^*_{i,j}$ values for author SDSs

Fig. 6. Comparison of Davies–Bouldin analysis calculated on the 5-community schema and on the official scientific disciplinary sector classification for the ICT field. The bar plots represent the value of D_i for the i-th community (plot b) and for the i-th scientific disciplinary sector (plot d); the Davies–Bouldin index corresponds to the average value of D_i, and is marked by a dashed red line. The author graph visualization for the two clustering systems (plot a and plot b) has been added to help interpretation of results.

- Even though not evident from the 5-community classification represented in Fig. 7, scholars from *ING-INF/02 - Electromagnetic Fields* and *ING-INF/07 - Electrical And Electronic Measurement* both have a distinguishable identity which becomes clear when we look at the community structure of the scholar network at a higher resolution. In fact, moving from 5 to 6 communities, scholars from these two scientific disciplinary sectors split almost perfectly into two parallel flows which remain separated as the number of communities increase;
- *ING-INF/01 - Electronic Engineering* scholars split into a larger group which flows along with *ING-INF/02 - Electromagnetic Fields* scholars, and into a smaller group which flows along with *ING-INF/07 - Electrical And Electronic Measurement* scholars. Moving from 8 to 9 communities, however, a large share of *ING-INF/01 - Electronic Engineering* scholars branch off to form an additional community with almost no participation from other scientific disciplinary sectors;
- Scholars from *INF/01 - Informatics* and *ING-INF/05 - Information Processing Systems* initially merge into the **knowledge & theoretical logic** community, which then starts to split into smaller communities related to research on **pattern recognition** and **pervasive multimedia**. Differently from the previous case, looking at the scholar network at a higher resolution

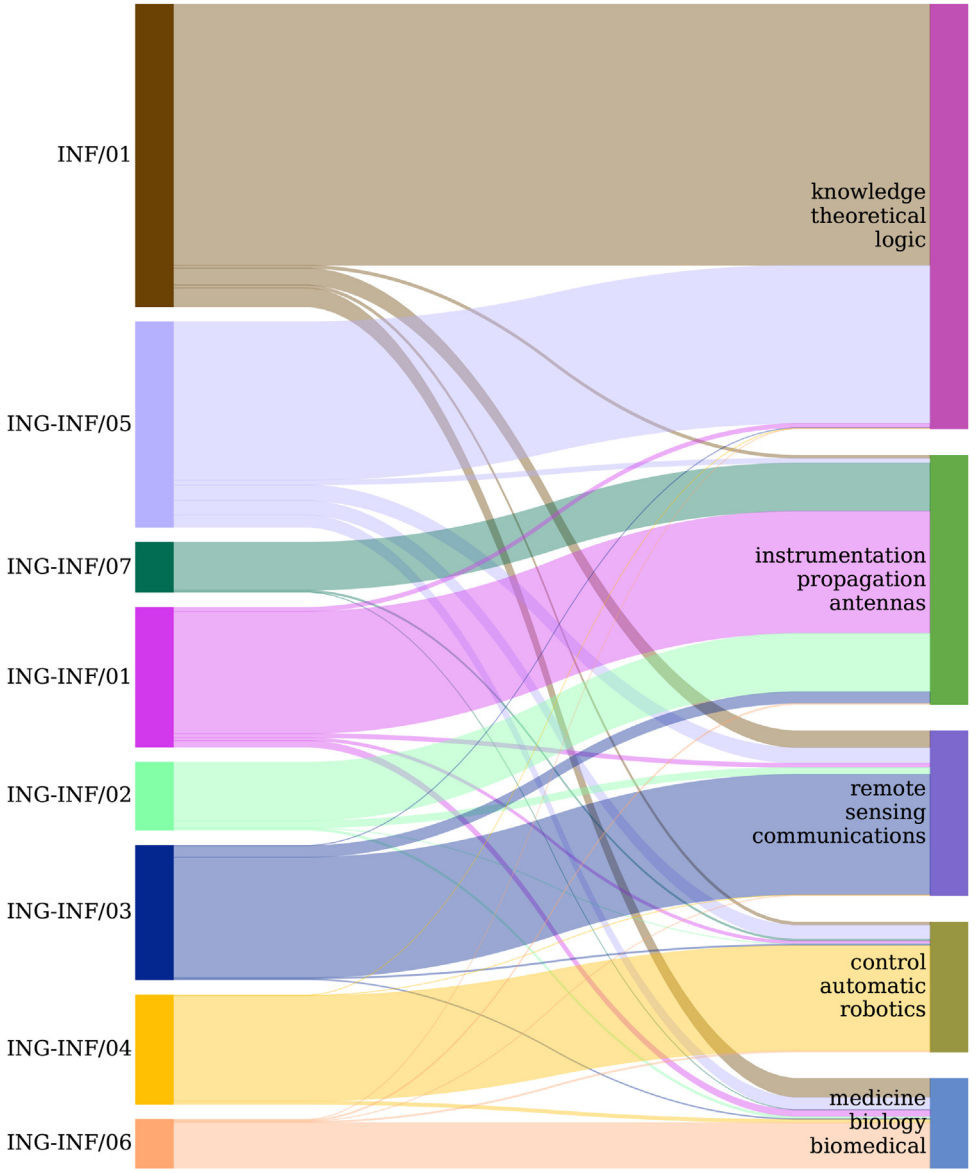


Fig. 7. The alluvial diagram represented in the figure shows how scholars from the various ICT scientific disciplinary sectors (on the left) flow together or apart into the data-driven communities (on the right), revealing similarities and differences between the SDS classification and the detected network structure. The width of the streams are proportional to the number of scholars flowing from each SDS to the different data-driven communities.

Table 3
Confusion matrix between the official classification into scientific disciplinary sectors and the 5- community classification detected with network analysis methodologies for the ICT field. Entries are highlighted in gradual shades of gray according to the percentage in brackets, which represents how scholars from a specific scientific disciplinary sector distribute across the different communities. To improve clarity, communities are labelled with a selected triple of significant keywords.

	Data-driven communities				
	Knowledge theoretical logic	Instrumentation propagation antennas	Remote sensing communications	Control automatic robotics	Medicine biology biomedical
INF/01	637 (86%)	7 (1%)	41 (6%)	7 (1%)	46 (6%)
ING-INF/05	386 (77%)	12 (2%)	38 (8%)	35 (7%)	30 (6%)
ING-INF/07	0 (0%)	117 (96%)	0 (0%)	4 (3%)	1 (1%)
ING-INF/01	10 (3%)	298 (87%)	10 (3%)	7 (2%)	16 (5%)
ING-INF/02	0 (0%)	143 (86%)	17 (10%)	1 (1%)	5 (3%)
ING-INF/03	1 (0%)	28 (9%)	294 (90%)	3 (1%)	2 (1%)
ING-INF/04	1 (0%)	0 (0%)	1 (0%)	257 (96%)	8 (3%)
ING-INF/06	1 (1%)	3 (3%)	1 (1%)	3 (3%)	111 (93%)

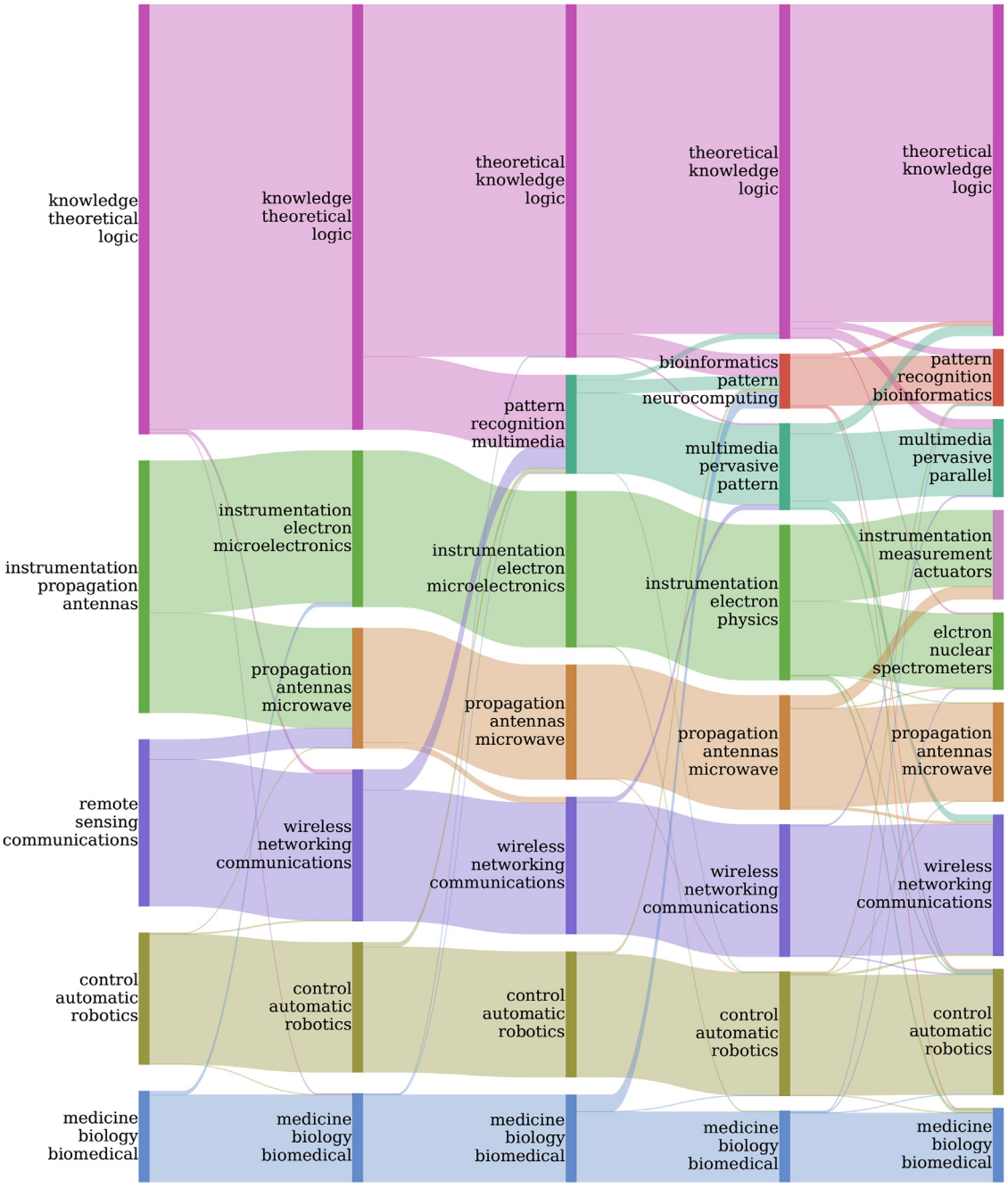


Fig. 8. An alluvial diagram showing how the proposed methodology redistribute the ICT scholars into a series of classifications characterized by an increasing number of clusters (from 5 to 9 clusters). All clusters are labelled according to the tf-idf schema. At each step, a new color is assigned to the novel detected cluster.

does not help to isolate the two scientific disciplinary sectors, which on the contrary appear to contribute equally to more specific research subfields. The entangled nature of research in *INF/01 - Informatics* and *ING-INF/05 - Information Processing Systems* is somewhat consistent with what is suggested by the Davies–Bouldin coefficient calculated for the scientific disciplinary sector classification system (Fig. 6).

4. Microbiology & genetics analysis

Despite the large number of scholars, the ICT field analyzed in the previous section is relatively narrow in terms of macro-sectors tackled - indeed seven out of the eight considered SDS come from the macro-sector "information engineering"

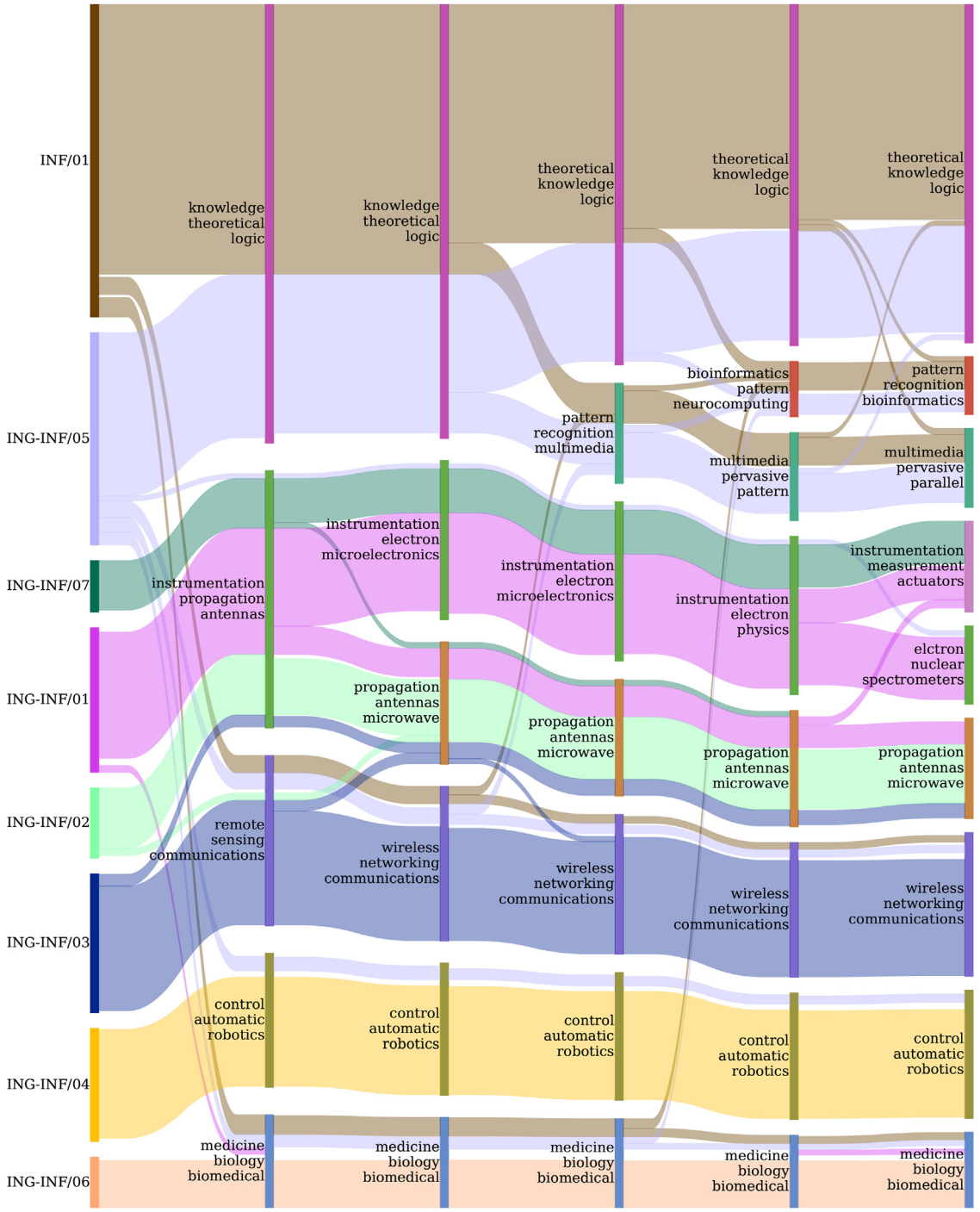


Fig. 9. The same alluvial diagram of Fig. 8, showing how scholars in the ICT field redistribute themselves starting from the scientific disciplinary sector official classification into an increasing number of data-driven categories (from 5 to 9 clusters). Flows between the series of classifications are colored according to the composition of scholars in terms of scientific disciplinary sectors. To simplify the reading of the graph, only flows accounting for more than 10 scholars were plotted.

(ING-INF), whereas only the "informatics" DSD, namely INF/01, comes from a different macro-sector. Quite interestingly, the major overlap was found among SDSs in different macro-sectors.

This initial finding pushed us to analyze the much more complex case of a topic area that, in what follows, we will hereafter refer to as Microbiology&Genetics. After consultation with experts in the field, we specifically selected as starting point the eleven SDSs reported inTable 5. Note that such 11 scientific disciplinary sectors belong to 4 widely different macrosectors,

Table 4
Confusion matrix between the official classification into scientific disciplinary sectors and the 9- community classification detected with network analysis methodologies for the ICT field. Entries are highlighted in gradual shades of gray according to the percentage in brackets, which represents how scholars from a specific scientific disciplinary sector distribute across the different communities. To improve clarity, communities are labelled with a selected triple of significant keywords.

	Data-driven communities								
	Theoretical knowledge logic	Pattern recognition bioinformatics	Multimedia pervasive parallel	Instrumentation measurement actuators	Elcetron nuclear spectrometers	Propagation antennas microwave	Wireless networking communications	Control automatic robotics	Medicine biology biomedical
INF/01	522 (71%)	79 (11%)	81 (11%)	0 (0%)	9 (1%)	1 (0%)	23 (3%)	5 (1%)	18 (2%)
ING-INF/05	270 (54%)	55 (11%)	96 (19%)	0 (0%)	14 (3%)	1 (0%)	27 (5%)	24 (5%)	14 (3%)
ING-INF/07	0 (0%)	0 (0%)	0 (0%)	106 (87%)	3 (2%)	7 (6%)	0 (0%)	5 (4%)	1 (1%)
ING-INF/01	5 (1%)	0 (0%)	5 (1%)	99 (29%)	149 (44%)	53 (16%)	3 (1%)	8 (2%)	19 (6%)
ING-INF/02	0 (0%)	0 (0%)	0 (0%)	8 (5%)	5 (3%)	140 (84%)	5 (3%)	1 (1%)	7 (4%)
ING-INF/03	0 (0%)	1 (0%)	5 (2%)	1 (0%)	3 (1%)	35 (11%)	281 (86%)	0 (0%)	2 (1%)
ING-INF/04	1 (0%)	2 (1%)	0 (0%)	0 (0%)	1 (0%)	0 (0%)	0 (0%)	257 (96%)	6 (2%)
ING-INF/06	1 (1%)	0 (0%)	0 (0%)	1 (1%)	1 (1%)	1 (1%)	1 (1%)	3 (3%)	111 (93%)

Table 5
List of the scientific disciplinary sectors related to the field of Microbiology&Genetics, along with the number of associated scholars in the publication dataset, according to the Italian classification system.

SDS code	SDS description	No. of scholars
AGR/16	Agriculture Microbiology	147
BIO/11	Molecular Biology	245
BIO/13	Experimental Biology	293
BIO/15	Pharmaceutical Biology	69
BIO/18	Genetics	190
BIO/19	General Microbiology	121
CHIM/10	Food Chemistry	72
CHIM/11	Chemistry and Biotechnology of Fermentation	41
MED/03	Medical Genetics	154
MED/04	Experimental Medicine and Pathophysiology	489
MED/07	Microbiology and Clinical Microbiology	290
	Total	**2111**

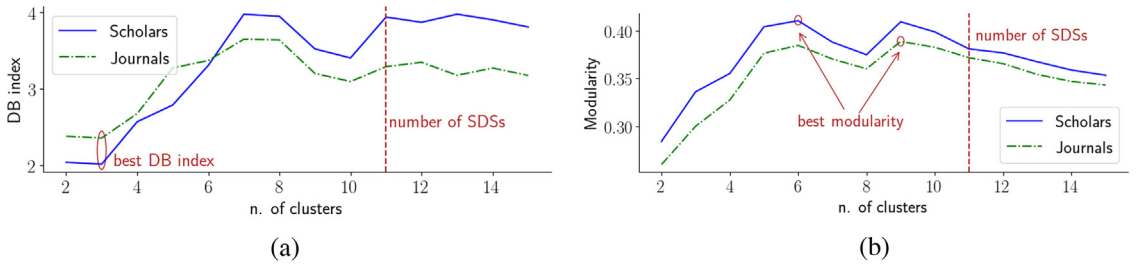


Fig. 10. Values of the Davies–Bouldin index (suplot a) and Barber's modularity (subplot b) for scholar clustering (solid blue line) and journal clustering (dash-dotted green line) for the Microbiology & Genetics field, calculated over different numbers of clusters.

and specifically biology (BIO), medicine (MED), chemistry (CHIM) and even agriculture (AGR). We then followed up exactly as in the ICT case, by preparing a publication dataset comprising the journal papers authored by Italian scholars associated to one of such 11 scientific disciplinary sectors. The final Microbiology&Genetics dataset, cleaned as detailed in the Appendix, covers 61 950 papers published in 2099 journals by 2111 scholars The relative bipartite graph is composed of $n = 2111$ scholar nodes, $m = 2099$ journal nodes and 56 891 edges.

Following the same steps as in Section 3, we applied the co-clustering algorithm presented in Section 2 for different choices of the number of co-clusters, detecting from 2 up to 15 communities (i.e. up to a number of communities equal to the number of the "official" sectors taken into consideration). As reported in Fig. 10a, for the case of Microbiology&Genetics the Davies–Bouldin criterion suggests that the optimal choice in terms of intra-cluster scatter and inter-cluster separation would be to select as little as just 3 communities. On the contrary, modularity favours higher values as the optimal number of communities, showing two local maxima at 6 and 9 communities. In what follows, we will focus on the 6-community configuration, which places halfway between the high-level representation suggested by the Davies–Bouldin index and the original number of 11 SDSs, thus representing in our opinion a more interesting and viable solution in the considered

context. The spatial distribution of scholar and journal nodes in the 3-dimensional SVD basis[12] is reported in Fig. 11. In particular, subplot Fig. 11 a shows both scholar and journal nodes, colored according to the community they have been associated to by the Spherical K-Means run in this lower- dimensional SVD space. Instead, Fig. 11 c plots only scholars nodes, colored according to the "official" scientific disciplinary sector classification. The pairwise clusters of scholars and journals are also represented in the original bipartite graph form in Fig. 12, together with the composition of co-clusters in terms of number of scholars, number of journals, and number of intra-community edges (i.e. edges linking two nodes in the same co-cluster). The total edge weight that falls inside the same co-clusters amounts to 60 932, approximately 60% of the 100 817 publications considered in the Microbiology&Genetics dataset.

Repeating the very same fully automated approach previously described in Section 3, we have labeled each scholar community with a set of representative keywords automatically extracted from the journal titles associated to the same co-cluster, i.e., without resorting to any a- priori knowledge on the communities. The top ten words detected for each cluster are reported in Table 6, along with their tf-idf score.

In this section we present the results of the assessment, based on the same Davies–Bouldin analysis of Section 3, of the scientific disciplinary sector classification, this time for the Microbiology&Genetics field. Again we resort, for visualization purposes, to a two-dimensional, force-directed representation of the one-mode projection of the bipartite publication graph onto the single set of scholar nodes via cosine similarity (Fig. 13). The Microbiology&Genetics author graph is composed by 2111 author nodes and 1 259 922 edges. In subplots Fig. 13 b, c and d the same author graph is colored according to the corresponding elements of vectors z_2, z_3 and z_4 used by the co- clustering algorithm to identify communities.

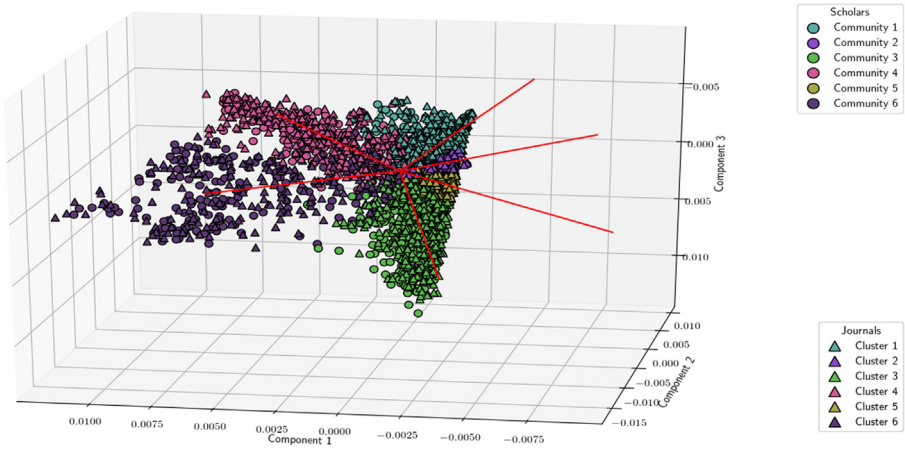
To assess the quality of the official SDS classification for the Microbiology&Genetics area against the data-driven 6-community scheme, we provide in Fig. 14 (respectively in the bottom and top plots) a graphical representation of the Davies–Bouldin analysis carried out for both classification systems. In the case of the Microbiology&Genetics field, the decrement in the Davies–Bouldin index value *DB* (from ≈4.63 of the 11 official SDSs to ≈3.31 for the six clusters found by our data-driven analysis) is even more evident than in the previous ICT case, thus suggesting that the current official classification in 11 SDSs yields a quite loose clustering - the lower the DB index, the better the clustering is considered to be. The bar graph in subplot Fig. 14 d, representing the worst-case similarity term $D_i = R_{i,j^*}$ for each sector, shows that some sectors contribute more than others to the average of the *DB* value; as for the ICT case, major overlap occurs between sectors from different macroareas (especially among the biology sectors BIO-11, BIO-13, and BIO-18, as well as between biology and medicine sectors, specifically BIO-13 and MED-04).

To examine how the two classification systems are interrelated, we report in the form of an alluvial graph (Fig. 15) how scholars from different scientific disciplinary sectors distribute into the 6 data- driven communities (numerical values are provided in Table 7). Fig. 15 suggests that:

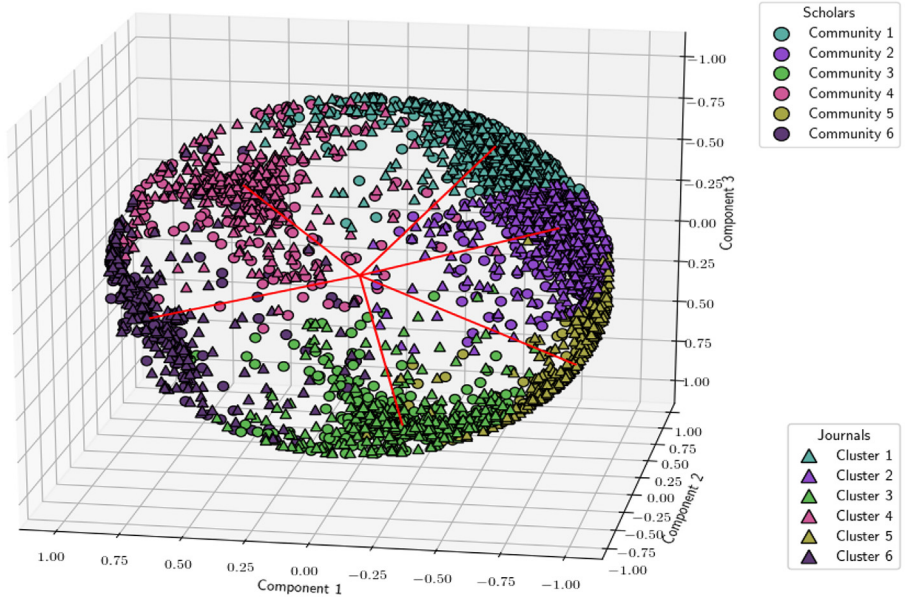
- there is no scientific disciplinary sector which is individually identified by one of the data-driven communities;
- on the contrary, all six data-driven communities aggregate various scientific disciplinary sectors from different macro-areas:
  - (i) the **natural products** community (keywords: food, natural, product) basically accounts for *CHIM/10 - Food Chemistry* and *BIO/15 - Pharmaceutical Biology*,
  - (ii) the **microbiology** community (keywords: microbiology, food, technology) comprises the vast majority of scholars from *AGR/16 - Agriculture Microbiology* and *CHIM/11 - Chemistry and Biotechnology of Fermentation*, and a large share of scholars from *BIO/19 - General Microbiology*;
  - (iii) the **virology** community (keywords: virology, antimicrobial, microbiology) includes the rest of *BIO/19 - General Microbiology* and basically all scholars from *MED/07 - Microbiology and Clinical Microbiology*;
  - (iv) a larger **genomics** community (keywords: genetics, mutation, genomics) consists of the great majority of *BIO/18 - Genetics* and *MED/03 - Medical Genetics* scholars and a large share of *BIO/13 - Experimental Biology* and *BIO/11 - Molecular Biology*;
  - (v) the **oncology** community (keywords: oncogene, endocrinology, cancer) is composed of *MED/04 - Experimental Medicine and Pathophysiology* scholars for one half, and brings together the remainder of *MED/03 - Medical Genetics*, *BIO/13 - Experimental Biology* and *BIO/11 - Molecular Biology* for the other half;
  - (vi) apart from some small contributions from other sectors, the **immunology** community (keywords: immunology, leukocyte, blood) corresponds to the remaining third of *MED/04 - Experimental Medicine and Pathophysiology*;
- in this 6-community configuration, *BIO/19 - General Microbiology*, *BIO/13 - Experimental Biology*, *BIO/11 - Molecular Biology* and *MED/04 - Experimental Medicine and Pathophysiology* are the scientific disciplinary sectors which make a substantial contribution to the composition of different communities.

In Figs. 16 and 17 we take into consideration also the community structure that we would obtain for the Microbiology&Genetics area when the number of co-clusters varies from 3, as suggested by the Davies–Bouldin analysis, up to 11 communities, equal to the number of SDSs. The alluvial diagram in Fig. 17 shows how scholars would be reallocated accord-

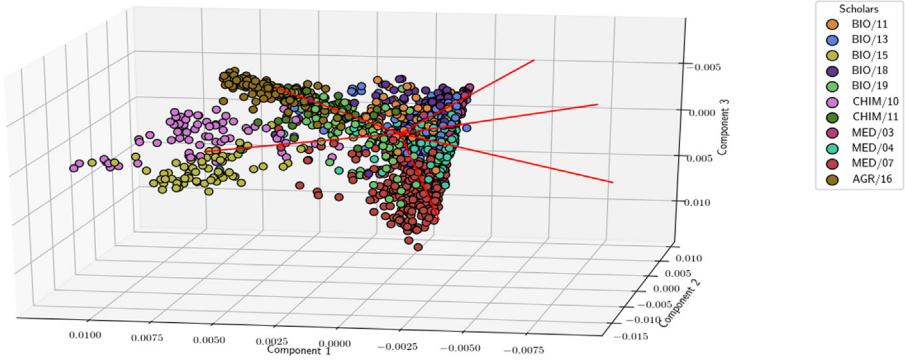
[12] As for the 5 communities of the ICT case, also this 6-communities configuration is based on the $3 = \lceil log_2(6) \rceil$ smallest singular values of the normalized publication matrix A_n.



(a) Detected communities



(b) Detected communities - projection onto the unit sphere



(c) Scientific disciplinary sectors

Fig. 11. Representation of scholars and journals for the Microbiology&Genetics dataset in the two-dimensional space associated to the two largest non-trivial singular values of A_n. Scholars and journals are marked respectively with circles and triangles, and red lines/crosses represent the concept vectors.

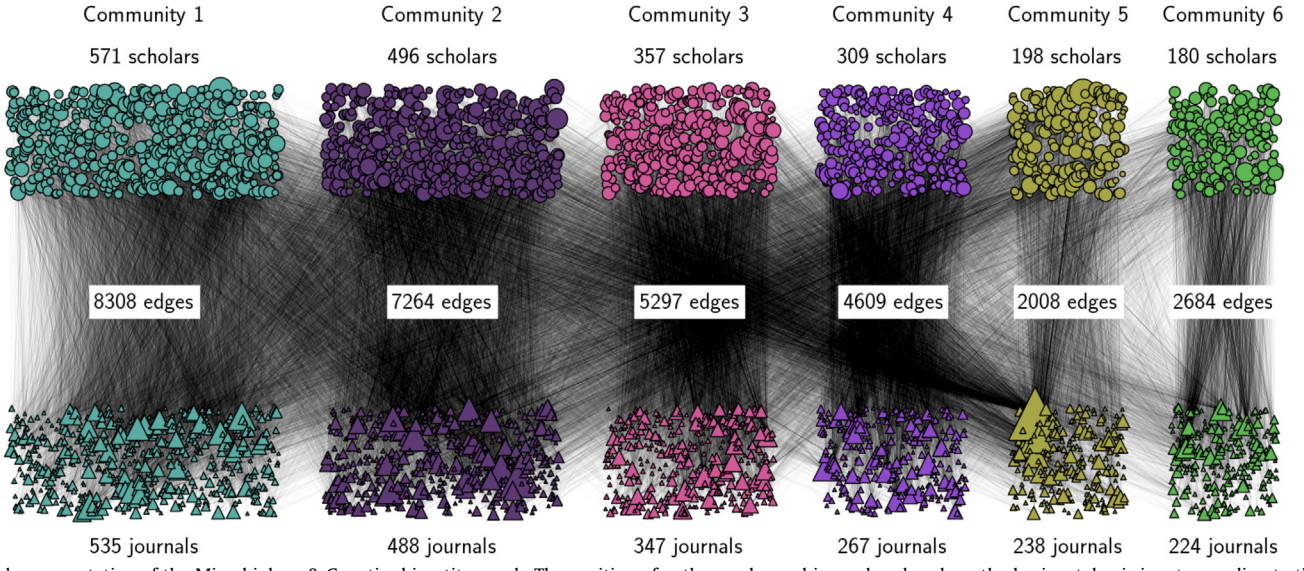


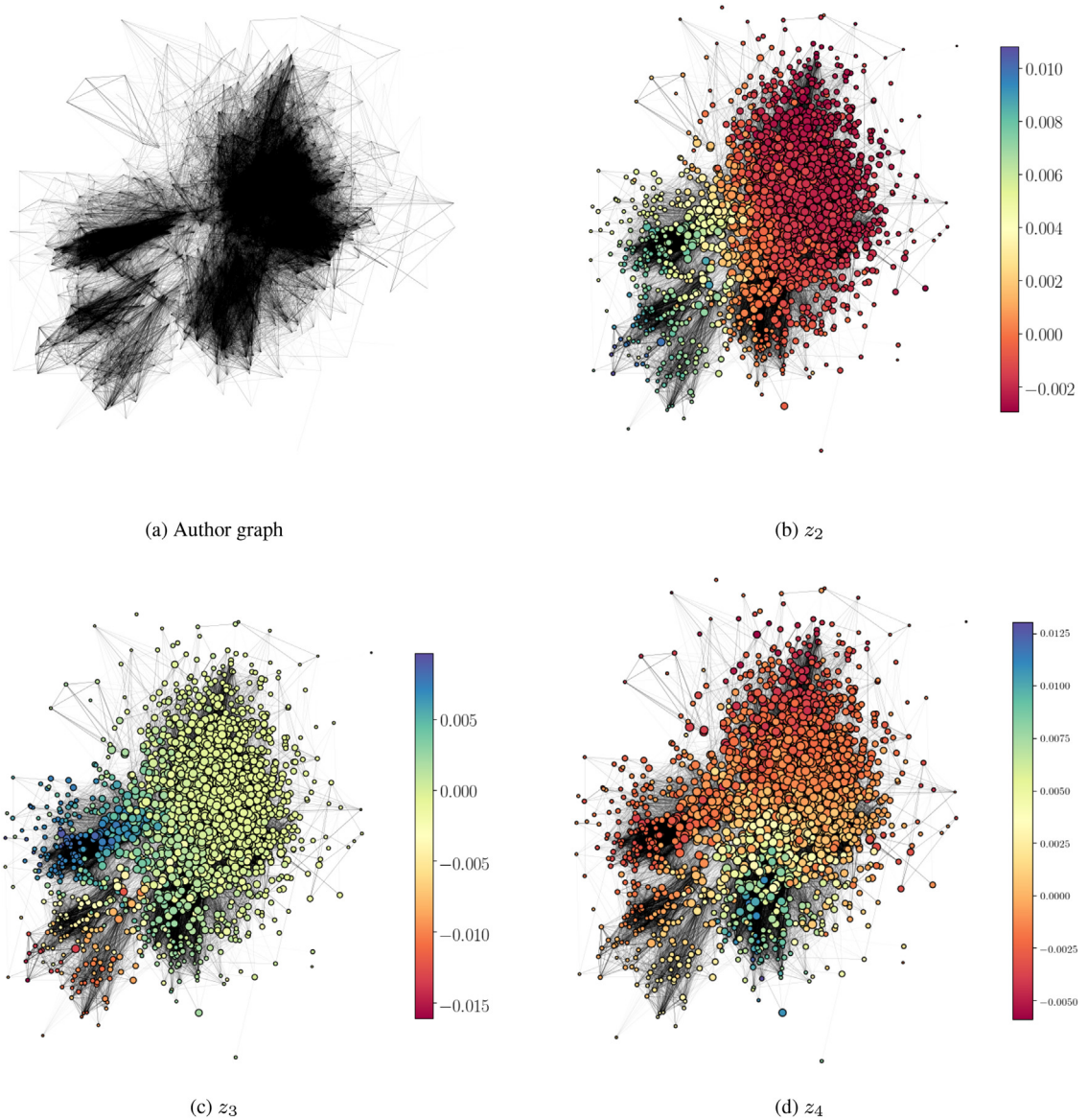
Fig. 12. Two-dimensional representation of the Microbiology & Genetics bipartite graph. The position of author nodes and journal nodes along the horizontal axis is set according to the co-clusters they are assigned to, in order to avoid edge crossing and to highlight the effectiveness of the co-clustering method. The size of nodes is proportional to their degree centrality, and the thickness of edges is proportional to the associated weight.

Table 6
Top 10 words for journal clusters in the Microbiology&Genetics field according to the tf-idf score. Note that, to keep the labeling process totally automatic, data was on purpose not cleaned from synonyms and derived words (e.g., microbiology/microbiologica, product/products), acronyms (e.g. plos, fems) or non-significant words (e.g., one, national).

Community 1	
Word	**TF-IDF score**
Genetics	0.058666
Mutation	0.018699
Medical	0.018603
Genomics	0.011393
Neurology	0.010559
Bioinformatics	0.010185
Embo	0.009135
Acids	0.008652
Reproduction	0.008436
Biophysical	0.008241
Community 2	
Word	**TF-IDF score**
Oncogene	0.020958
Oncotarget	0.015894
National	0.014947
Death	0.012312
Thrombosis	0.010129
Endocrinology	0.010048
Cancer	0.009666
Radical	0.009388
Free	0.009388
Leukemia	0.008812
Community 3	
Word	**TF-IDF score**
Virology	0.055982
Antimicrobial	0.054937
Microbiology	0.048358
Infection	0.048195
Infectious	0.043921
Chemotherapy	0.033685
Agents	0.028357
Microbiologica	0.019797
Aids	0.018186
Diseases	0.014879
Community 4	
Word	**TF-IDF score**
Microbiology	0.142977
Food	0.074441
Technology	0.024565
Dairy	0.017624
Alimentari	0.015613
Industries	0.015613
Fems	0.013370
Environmental	0.011545
Yeast	0.011432
Microbial	0.011033
Community 5	
Word	**TF-IDF score**
Immunology	0.060093
One	0.051408
Leukocyte	0.026205
Blood	0.024213
PLOS	0.019887
Hepatology	0.018527
Immunotherapy	0.011850
Oncoimmunology	0.011684
Immunological	0.010181
Neuroimmunology	0.008179

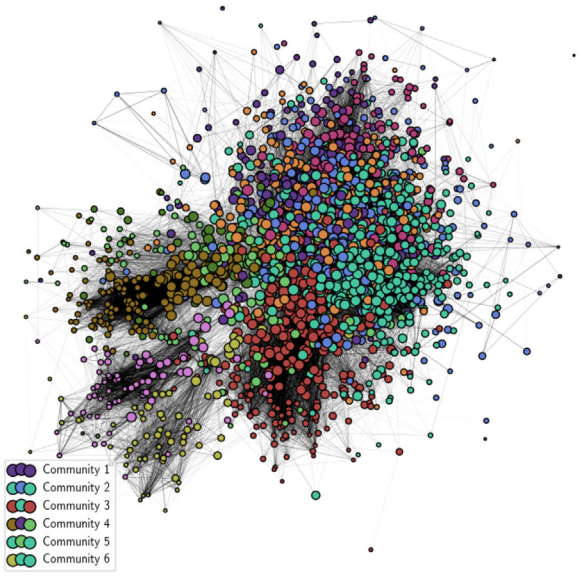
Table 6 (*Continued*)

Community 6 **Word**	**TF-IDF score**
Food	0.074666
Natural	0.070565
Product	0.054298
Agricultural	0.050231
Chromatography	0.029069
Products	0.017547
Oil	0.014083
Separation	0.013782
Fitoterapia	0.013330
Phytochemistry	0.012577

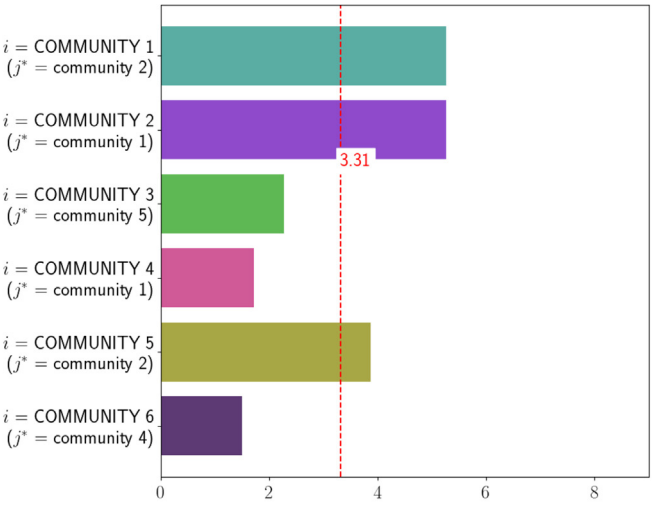


(a) Author graph (b) z_2 (c) z_3 (d) z_4

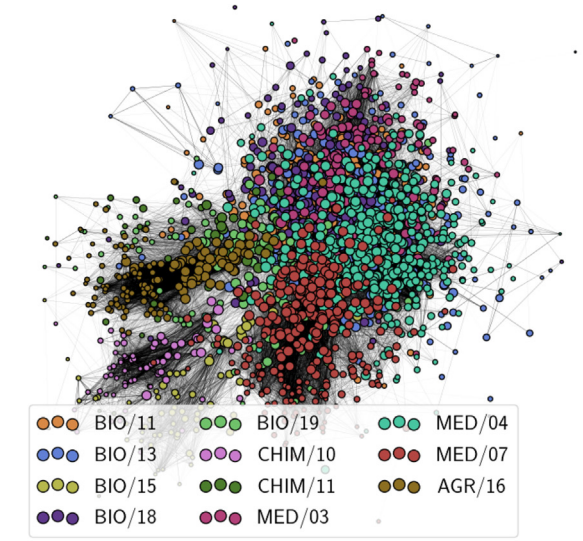
Fig. 13. (Plot a) Two- dimensional representation of the author graph for the Microbiology&Genetics field according to the Fruchterman-Reingold layout. For visualization purposes, we only plotted edges with a weight higher than 0.25. The size of nodes is proportional to their degree centrality, and the thickness of edges is proportional to their associated weight. (Plot b, c and d) Microbiology&Genetics author graph colored according to the value of the three smallest eigenvalues for $Lz = \lambda Dz$. Hot and cold colors represent negative and positive values, respectively.



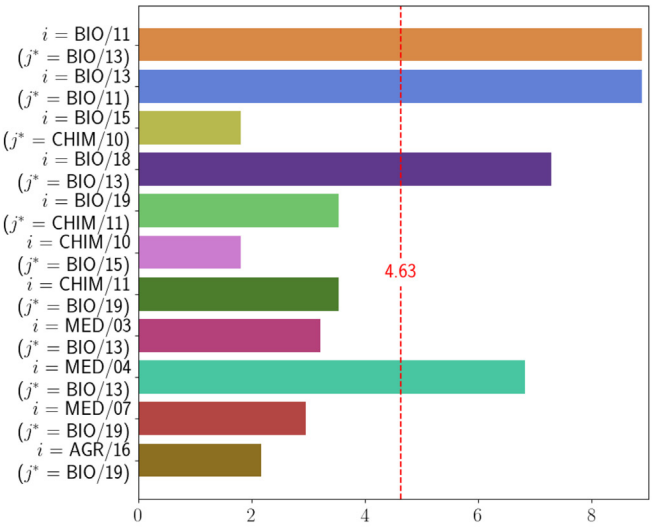
(a) Graph visualization of author communities



(b) $D_i = R^*_{i,j}$ values for author communities



(c) Graph visualization of author SDSs



(d) $D_i = R^*_{i,j}$ values for author SDSs

Fig. 14. Comparison of Davies–Bouldin analysis calculated on the 6-community schema and on the official scientific disciplinary sector classification for the Microbiology & Genetics field. The bar plot represents the value of D_i for each community (plot b) and for each scientific disciplinary sector (plot d); the Davies–Bouldin index corresponds to the average value of D_i, marked by a dashed red line. A graph visualization for the two clustering systems (plot a and plot b) has been added.

ing to the proposed methodology into an increasing number of clusters. The same applis for Fig. 17, where however flows of scholars from a community to the other are colored according to their scientific disciplinary sectors, to make the contribution of each sector evident. We provided here only the confusion matrix for the 11-community configuration (Table 8), but all confusion matrices and tf-idf keywords are available in the Supplementary Material. In particular:

- The **natural products** community detected in the 3-community configuration remains essentially the same as the number of communities increase, suggesting a clear, well-distinguishable research trend for *CHIM/10 - Food Chemistry* and *BIO/15 - Pharmaceutical Biology* scholars (only in the 8- community schema a part of scholars flows into a new, more specific,

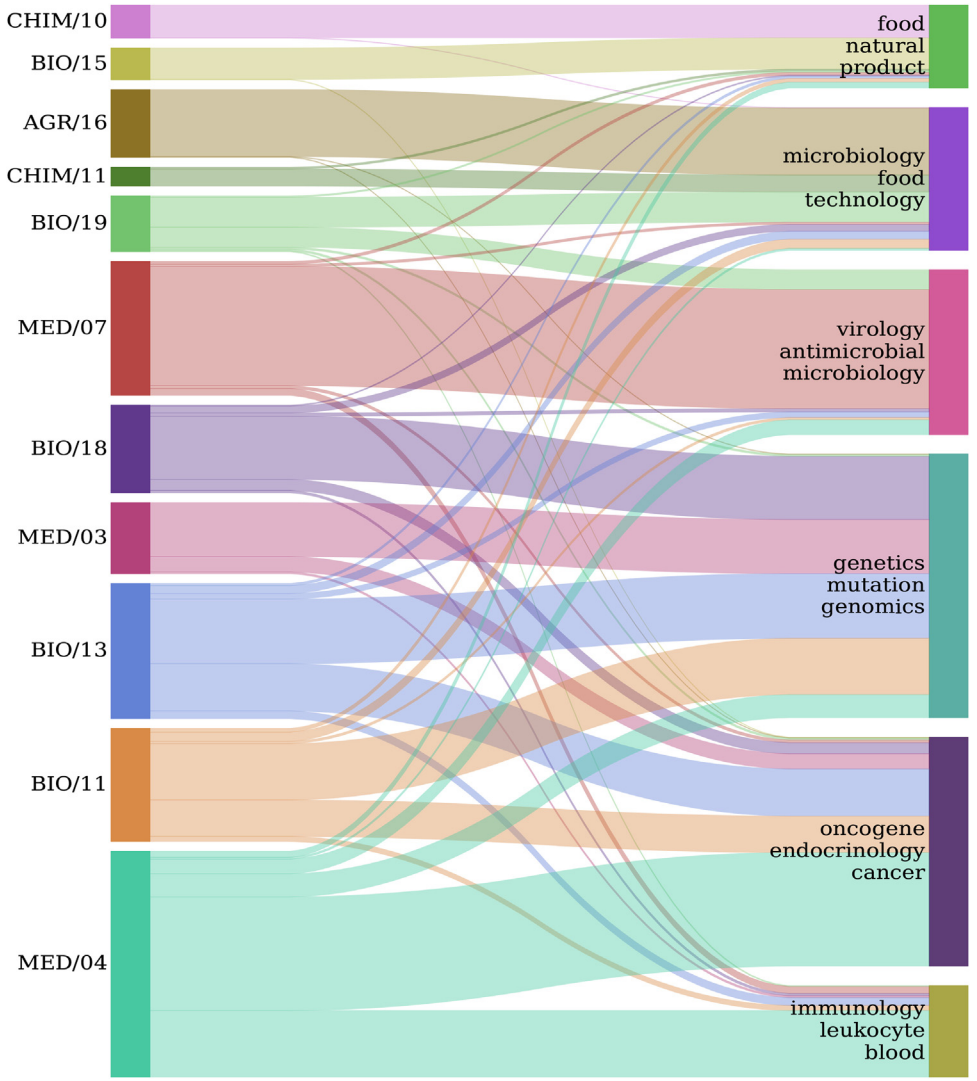


Fig. 15. An alluvial diagram, with scientific disciplinary sectors on the left and data-driven communities on the right, reveals similarities and differences of the SDS classification for the Microbiology & Genetics field with respect to the detected network structure.

Table 7
Confusion matrix between the official classification into scientific disciplinary sectors and the 6- community classification detected with network analysis methodologies for the Microbiology & Genetics field. Entries are highlighted in gradual shades of gray according to the percentage in brackets, which represents how scholars from a specific scientific disciplinary sector distribute across the different communities. To improve clarity, communities are labelled with a selected triple of significant keywords.

	Data-driven communities					
	Food natural product	Microbiology food technology	Virology antimicrobial microbiology	Genetics mutation genomics	Oncogene endocrinology cancer	Immunology leukocyte blood
CHIM/10	71 (99%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
BIO/15	68 (99%)	0 (0%)	0 (0%)	0 (0%)	1 (1%)	0 (0%)
AGR/16	0 (0%)	145 (99%)	0 (0%)	1 (1%)	1 (1%)	0 (0%)
CHIM/11	4 (10%)	37 (90%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
BIO/19	3 (2%)	65 (54%)	43 (36%)	4 (3%)	4 (3%)	2 (2%)
MED/07	6 (2%)	5 (2%)	258 (89%)	0 (0%)	6 (2%)	15 (5%)
BIO/18	2 (1%)	15 (8%)	7 (4%)	137 (72%)	24 (13%)	5 (3%)
MED/03	0 (0%)	0 (0%)	0 (0%)	117 (76%)	33 (21%)	4 (3%)
BIO/13	5 (2%)	17 (6%)	12 (4%)	140 (48%)	102 (35%)	17 (6%)
BIO/11	8 (3%)	20 (8%)	5 (2%)	122 (50%)	79 (32%)	11 (4%)
MED/04	13 (3%)	4 (1%)	32 (7%)	50 (10%)	246 (50%)	144 (29%)

Table 8
Confusion matrix between the official classification into scientific disciplinary sectors and the 11- community classification detected with network analysis methodologies for the Microbiology & Genetics field. Entries are highlighted in gradual shades of gray according to the percentage in brackets, which represents how scholars from a specific scientific disciplinary sector distribute across the different communities. To improve clarity, communities are labelled with a selected triple of significant keywords.

	Data-driven communities										
	Food natural product	Food microbiology technology	Microbiology biotechnology bacteriology	Antimicrobial virology infection	Plant aquatic marine	Genetics medical human	Gerontology ageing rejuvenation	Oncogene cell endocrinology	Cancer blood	Medicinal bioorganic immunology	Neuroimmun. biomaterials materials
CHIM/10	70 (97%)	1 (1%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
BIO/15	68 (99%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1%)	0 (0%)	0 (0%)
AGR/16	0 (0%)	113 (77%)	32 (22%)	0 (0%)	0 (0%)	1 (1%)	0 (0%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)
CHIM/11	4 (10%)	2 (5%)	32 (78%)	0 (0%)	3 (7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
BIO/19	1 (1%)	9 (7%)	53 (44%)	30 (25%)	9 (7%)	0 (0%)	1 (1%)	3 (2%)	1 (1%)	14 (12%)	0 (0%)
MED/07	6 (2%)	2 (1%)	3 (1%)	230 (79%)	0 (0%)	2 (1%)	2 (1%)	0 (0%)	7 (2%)	26 (9%)	12 (4%)
BIO/18	1 (1%)	3 (2%)	9 (5%)	3 (2%)	20 (11%)	36 (19%)	43 (23%)	63 (33%)	7 (4%)	4 (2%)	1 (1%)
MED/03	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	115 (75%)	21 (14%)	9 (6%)	7 (5%)	0 (0%)	2 (1%)
BIO/13	3 (1%)	0 (0%)	10 (3%)	4 (1%)	16 (5%)	53 (18%)	51 (17%)	107 (37%)	34 (12%)	8 (3%)	7 (2%)
BIO/11	0 (0%)	0 (0%)	14 (6%)	1 (0%)	50 (20%)	10 (4%)	28 (11%)	111 (45%)	25 (10%)	6 (2%)	0 (0%)
MED/04	8 (2%)	0 (0%)	3 (1%)	18 (4%)	5 (1%)	10 (2%)	48 (10%)	154 (31%)	175 (36%)	53 (11%)	15 (3%)

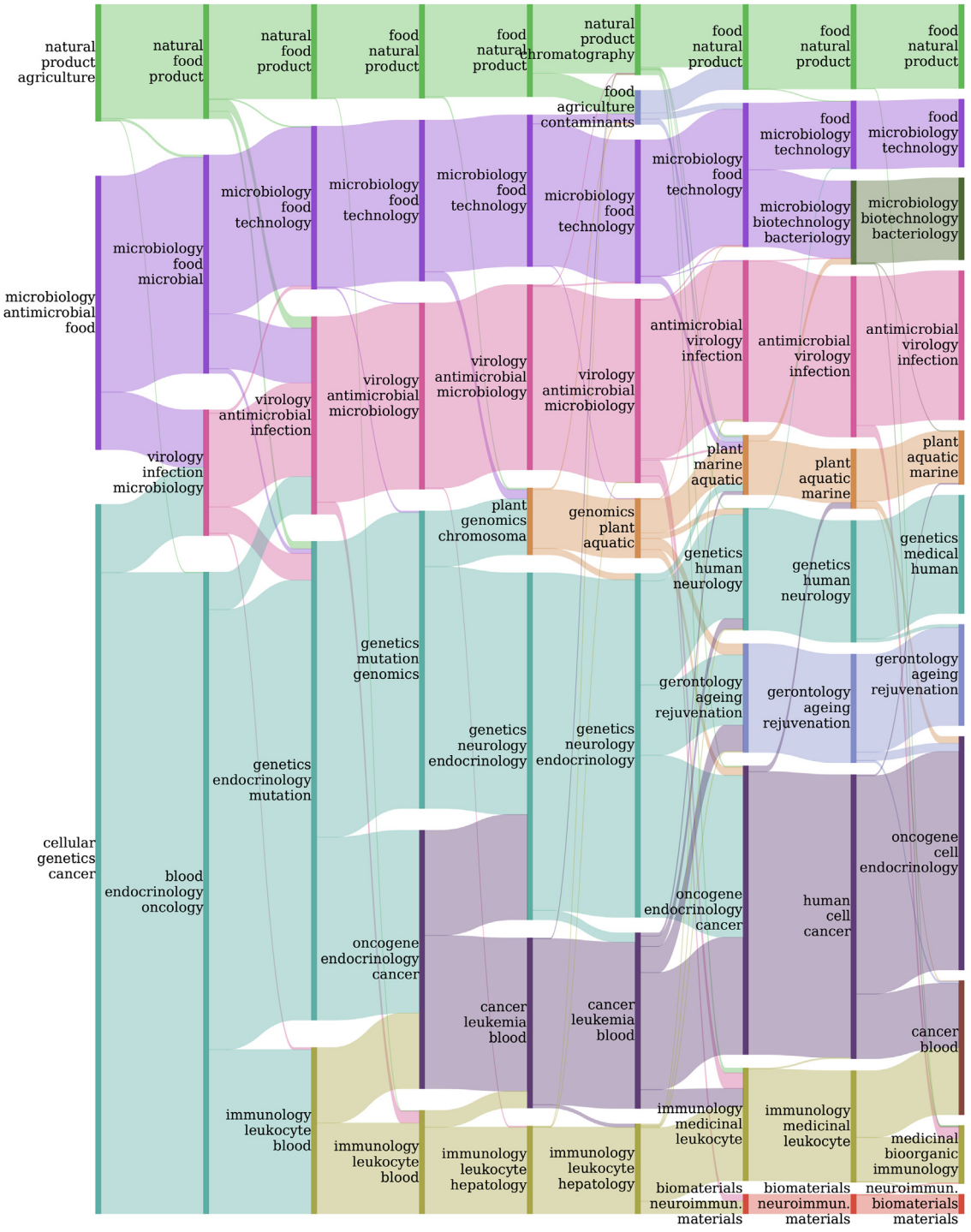


Fig. 16. An alluvial diagram showing how the proposed methodology redistribute scholars in the Microbiology & Genetics field into an increasing number of clusters (from 3 to 11 clusters). All clusters are labelled according to the tf-idf schema. At each step, a new color is assigned to the novel detected cluster.

smaller community which is however no more detected by the co-clustering algorithm in the 9-, 10-, 11- community schema);

- Moving from 3 to 5 detected communities, scholars from the **microbiology** community split into two stable, parallel flows: a first research trend in microbiology seems to be related to **food technology**, involving the majority of scholars from *AGR/16 - Agriculture Microbiology*, *CHIM/11 - Chemistry and Biotechnology of Fermentation* and a significant part of *BIO/19 - General Microbiology* scholars, whereas the latter microbiology research trend, related to **infectious diseases**

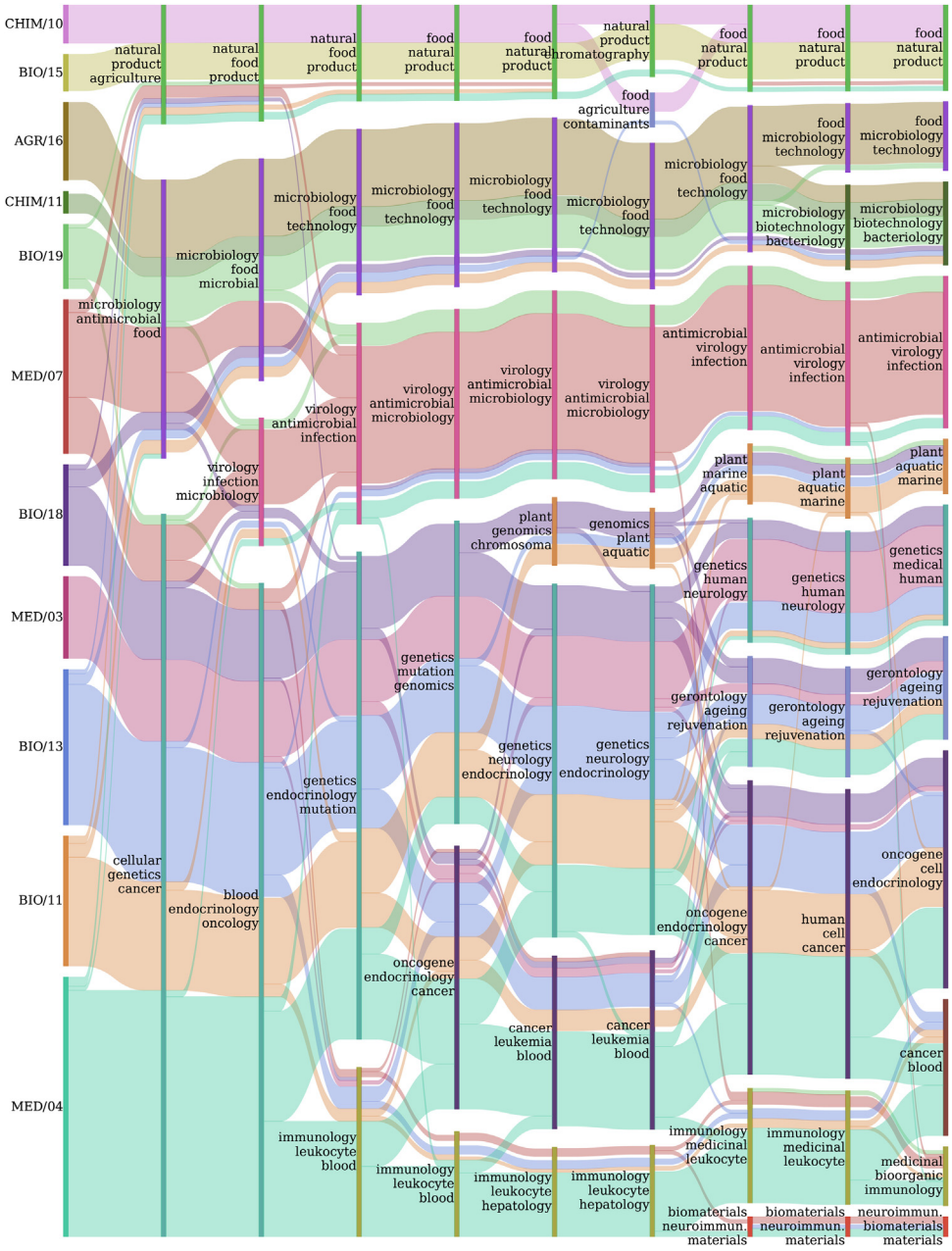


Fig. 17. The same alluvial diagram of Fig. 16, where flows between classifications at different resolution levels are colored according to the scientific disciplinary sectors of scholars composing the communities. The graph shows how scholars in the Microbiology & Genetics field redistribute themselves starting from the scientific disciplinary sector official classification into an increasing number of data-driven categories (from 3 to 11 clusters). To simplify the reading of the graph, only flows accounting for more than 5 scholars were plotted.

gathers back basically the whole *MED/07 - Microbiology and Clinical Microbiology* scientific disciplinary sector (along with a smaller contribution of *BIO/19 - General Microbiology* and *MED/04 - Experimental Medicine and Pathophysiology* scholars), suggesting that this sector has a more clear and robust cluster structure which was not directly evident in the 3-community schema;

- Differently from the previous case, even if the initial **biology** community also splits into two main flows in the 6, 7 and 8-community schema (respectively related to **genetics** and **hemato-oncology**), scholars from all sectors involved (*BIO/18 - Genetics*, *MED/03 - Medical Genetics*, *BIO/13 - Experimental Biology*, *BIO/11 - Molecular Biology* and *MED/04 - Experimental Medicine and Pathophysiology*) still contribute substantially to both research trends. This behaviour is still evident moving to even higher resolution schemas, when the genetics and oncology flows further split into smaller communities (e.g. the

genomics or ageing community). As noted also in the ICT analysis in Section 3, this result was somewhat already suggested by the Davies–Bouldin bar plot in Fig. 14.

- Finally, starting from the 5-community resolution schema, the remaining half of *MED/04 - Experimental Medicine and Pathophysiology* scholars (the ones not associated with the **genetics** and **hemato-oncology** trends) follow an indipendent research flow related to **immunology**.

5. Discussion and conclusions

5.1. Main results

To the best of our knowledge, in the scientometrics literature this paper is the first to apply co-clustering to a bipartite bibliometric network between authors and journals. Unlike most bibliometric networks, which are defined at the level of researchers or journals, the proposed methodology has the great advantage of taking into account both aspects at the same time: within the framework outlined in Section 2, one can indeed investigate at the same time the structural properties of the author network, in which nodes are scientists linked by "shared" journals, and the structural properties of the journal network, in which nodes are journal linked by "shared" authors. Moreover, our methodology permits to detect community of researchers (respectively, journals) in a purely unsupervised manner, and without relying on labels (e.g. keywords). As such it may be used as a basis for constructing a data-driven classification system for research areas and/or communities of scholars.

Our approach was further used to challenge the official classification adopted by the Italian academic system, one of the very few European countries which provides an explicit classification of research areas, further requiring every faculty member to be registered in one and only one of such "scientific disciplinary sectors" (SDS). Such system is extensively used since decades in Italy for career promotions and research productivity assessment, call for university positions, research grants, and so on. We have specifically targeted two broad areas for which we assembled for the purpose two novel datasets collecting publications authored by Italian faculty members: an "ICT" (Information and Communication Technology) dataset comprising publications authored by faculty members in 8 SDS (from ICT engineering and from Computer Science), and a "microbiology and Genetics" dataset, comprising publications authored by faculty members in 11 SDS from four macro-areas (Biology, Medicine, Chemistry and one targeted Agriculture sector).

The results of our analysis shows evidence of a significant overlapping among some of the considered SDSs, therefore showing that, at least in some cases, the current governmental subdivision among SDSs does not reflect a meaningful difference in publication habits. Interestingly, in at least two cases, two ore more SDSs were so much intertwined that separation was not emerging even when significantly increasing the number of clusters. Moreover, in both the considered datasets, it is worth to note that most of the overlaps occurs among sectors classified in different macro-areas, thus also questioning the validity of a rigid and hierarchic classification approach.

5.2. Limitations and further research

We are aware that the approach followed in this paper may have some limitations. Firstly, it has been argued that classifications defined at the level of individual publications provide a better representation of science than journal- based taxonomies (Klavans & Boyack, 2017; Perianes- Rodriguez & Ruiz-Castillo, 2017). In particular, document-based classifications have the advantage of naturally handling multidisciplinarity, in terms of single publications which, otherwise, would not fit well in a journal-based taxonomy (Ruiz-Castillo & Waltman, 2015; Wang & Waltman, 2016). From this point of view, the use of journals rather than individual publications may be regarded as a limitation of our approach: whereas scholars are usually experts in a particular area, the same does not always apply to journals, which may not be field-specific and may cover more than one discipline (Abramo, D'Angelo, & Costa, 2012; Leydesdorff, 2007; Leydesdorff & Rafols, 2011; Wang & Waltman, 2016). At present, the proposed methodology does not directly provide information on the degree of multidisciplinarity of journals, which will be addressed in future work.

It should also be mentioned that the Davies–Bouldin index is not specifically designed to handle overlapping clusters or high-dimensional data (Saitta et al., 2008), so the detection of the optimal number of co-clusters can be certainly improved resorting to other types of internal evaluation metrics which may prove to be more suitable, especially in the presence of overlap. Modularity, on the other hand, may favour the detection of large communities rather than smaller subgroups, due to the so-called "resolution limit" (Fortunato & Barthelemy, 2007). This was indeed one of the reasons that prompted us to include results for different choices of the number of clusters in sections 3 and 4. Another reason lies in the fact that, from the point of view of a research management application, there would be practical issues to take into account when choosing the number of clusters, e.g. cluster dimensions. This work was therefore more focused on *how* communities are identified (data- driven vs a-priori), more than on *how many* communities.

Finally, the choice made here of focusing only on journal papers was actually due the lack of a mechanism to group together conference papers over the years, as in publication databases like Scopus different editions of the same conference are usually associated with different identifiers. Extending the analysis to the case of conference papers will certainly be considered in future work.

Nonetheless, we believe that our approach may be exploited and/or adapted to challenge a number of very interesting future research issues, among which:

- *Multidisciplinarity* the proposed methodology may be extended in order to provide an estimate of the degree of journal multidisciplinarity;
- *Comparative analysis with other countries* it would be interesting to extend the scope of the analysis to the scholar-journal networks of other countries, to find out whether research trends are in some way influenced by the characteristics of each national academic system;
- *Time evolution* another promising direction for future work is to study how the identified research groups are evolving in time;
- *Labeling of research areas* Labeling of research area can be improved, on the one hand collecting more information (for instance, from publication abstracts) and, on the other hand, testing other text mining techniques.

Author contributions

Chiara Carusi: Concieved and designed the analysis; Collected the data; Contributed data or analysis tools; Performedthe analysis; Wrote the paper.

Giuseppe Bianchi: Concieved and designed the analysis; Wrote the paper.

Acknowledgements

The authors are grateful to Prof. Stefania Stefani for the suggestion of the SDS to use as starting point for the construction of the "Microbiology & Genetics" dataset and the relevant analysis, and to Prof. Sergio Palazzo and Prof. Francesco de Natale for the initial discussions on the "internal" structure of the SDS we belong to, a discussion which has pushed us to gather more insights on the much broader ICT community, thus stimulating this work. Moreover, we are also grateful to the anonymous reviewers for the many valuable suggestions which in particular prompted us to extend and revise the clustering evaluation process.

Appendix

The publication datasets employed in this study are based on articles published over a sixteen year period (2000–2016) by scholars working in Italian universities respectively in the Information and Communication Technology and Microbiology&Genetics fields. Each dataset was built collecting information about such publications from the Scopus[13] database; for each paper, we collected information on its title, on its author(s), on the journal it was published in and on its publication date. In this section we describe how the datasets used in the present paper were collected, processed and cleaned.

Data Collection

To properly collect the publication dataset, the first step was building a registry of scholars working in Italian universities. We focused our analysis on ICT and on the Microbiology&Genetics field, so we restricted our search to professors and researchers associated with one of the following scientific disciplinary sectors:

- INF/01 - Informatics
- ING-INF/01 - Electronic Engineering
- ING-INF/02 - Electromagnetic Fields
- ING-INF/03 - Telecommunications
- ING-INF/04 - Systems And Control Engineering
- ING-INF/05 - Information Processing Systems
- ING-INF/06 - Electronic And Informatics Bioengineering
- ING-INF/07 - Electrical And Electronic Measurement

in the case of the ICT field, and with one of the following scientific disciplinary sectors:

- AGR/16 - Agriculture Microbiology
- BIO/11 - Molecular Biology
- BIO/13 - Experimental Biology
- BIO/15 - Pharmaceutical Biology
- BIO/18 - Genetics
- BIO/19 - General Microbiology

[13] https://www.scopus.com/

Table 9
Number of scholars associated to each scientific disciplinary sector according to the Italian classification system.

Information and Communication Technology	
SDS code	**n. of scholars**
INF/01	855
ING-INF/01	357
ING-INF/02	170
ING-INF/03	345
ING-INF/04	290
ING-INF/05	745
ING-INF/06	122
ING-INF/07	130
Total	**3014**

Microbiology &Genetics	
SDS code	**n. of scholars**
AGR/16	149
BIO/11	255
BIO/13	298
BIO/15	70
BIO/18	192
BIO/19	122
CHIM/10	74
CHIM/11	43
MED/03	160
MED/04	496
MED/07	303
Total	**2162**

- CHIM/10 - Food Chemistry
- CHIM/11 - Chemistry and Biotechnology of Fermentation
- MED/03 - Medical Genetics
- MED/04 - Experimental Medicine and Pathophysiology
- MED/07 - Microbiology and Clinical Microbiology

in the case of the Microbiology & Genetics field. We collected from the Cineca database[14] the full name of professors and researchers working in Italian universities, together with their scientific disciplinary sector (SDS) and other kind of information, such as their role and affiliation.[15] A total number of 3014 and 2162 scholars were retrieved respectively for the two research areas, distributed across the scientific disciplinary sectors as reported in Table 9.

The second step was reconstructing the publication record of each scholar in our registry. Bibliographical information was retrieved automatically from the Scopus database, one of the largest abstract and citation database of peer-reviewed research literature. To collect each scholar's publications profile we used the Scopus APIs provided by the Elsevier Developer Portal,[16] together with the *pyscopus* Python wrapper[17] . The data collection process was based on three kind of queries sent subsequently to the Scopus database: author search, author retrieval and publication retrieval.

Author search Scopus assigns to each author indexed in the database an author profile together with a unique identifier (*Scopus Author ID*). Each author profile provides information about the author's current and previous affiliations, number of publications and relative bibliographic data, references, and details on the number of citations each published document has received. Unfortunately, no information was available in the Cineca Database about the Scopus Author ID associated to each scholar. Without a common identifier between the Cineca records and the Scopus author profiles, we are left with the scholar full name as the only link between the two databases. Moreover, for the same Cineca scholar there may be more than one author profile in the Scopus database. As explained by Moed and Halevi (2014), Scopus associates publications to authors taking into account many parameters such as name variants, co-authorship, subject areas and publications history. However, every time a publication cannot be clearly ascribed to an author, the Scopus system creates an additional author profile. Even though a feedback system is provided to authors to indicate whether there have been errors in the publication attribution process, there are still many cases where more than one author account has been created for the same person.

We built a mapping between scholars and one or more Scopus Author IDs by making use of the *author search* Scopus API, querying the Scopus database for all authors whose full name matched one of the scholar full names in the Cineca list[18] .

[14] A complete list of Italian faculty members is available at http://cercauniversita.cineca.it/php5/docenti/cerca.php
[15] Data is up-to- date as of the end of December 2016.
[16] The data was downloaded from Scopus API via http://api.elsevier.com and http://www.scopus.com between August 11 and 20, 2017 for the ICT field and between September 15 and 22, 2017 for the Microbiology & Genetics field.
[17] http://zhiyzuo.github.io/python-scopus/
[18] Please note that in this section when we use the word "author" we refer to Scopus author profiles.

Table 10
Distribution of homonymous and synonymous names after the author search procedure.

Information and communication technology **N. of Scopus profiles found** for the same scholar	**Scholar count**
0	36
1	2074
2	525
3	190
4	67
5	34
6	22
7	25
8	10
9	6
10	4
11	4
13	2
14	3
17	3
25	1
26	3
31	1
37	1
42	1
76	1
228	1
Total	**3014**

Microbiology &Genetics **No. of Scopus profiles found** for the same scholar	**Scholar count**
0	31
1	1591
2	309
3	103
4	56
5	29
6	12
7	8
8	4
9	2
10	3
11	1
12	2
13	2
14	1
15	1
16	2
25	1
30	1
35	1
50	1
299	1
Total	**2162**

Predictably, querying the database by scholar names introduced some ambiguity in the results of the matching procedure, since (i) two different scholars can share the same name and (ii) there can be two or more Scopus profiles, and, therefore, two or more author names, referring to the same scholar. Borrowing the definitions of Calero et al. (2006), we will refer to the former cases as *homonymous* names, and to the latter ones as *synonymous* names. To solve the problem with homonymous names, Calero et al. (2006) used a combination of author name and main organization: in our case, unfortunately, affiliation information was not always correct or up-to-date on Scopus database, so it could not be used to restrict our query. The only additional information we used in our query, other than the author name, was the Scopus *subject area* field. We restricted our search to Scopus authors whose subject area list contained either "COMP - Computer Science" or "ENGI - Engineering" when searching for ICT scholars, or "BIOC - Biochemistry, Genetics and Molecular Biology" when searching for scholars working in the Microbiology and Genetics field.

On the other hand, to keep track of all synonymous names, we followed a conservative approach: in case of a scholar with middle names, we looked for all authors whose full name had at least one of the scholar names in it. As expected, such an approach resulted in taking into consideration also many homonymous names. Cases of homonymous names were

detected and removed only at the end of the entire data collection process, leveraging the information gathered on authors' publications, as explained in more detail later in the *Data Cleaning* paragraph.

The author search procedure for the ICT field resulted in a dataset of 5265 different Scopus Author IDs potentially associated with the 3014 scholars in the Cineca list; for the Microbiology & Genetics field, the procedure found 3666 different Scopus Author IDs starting from the 2136 scholars in the Cineca list. Table 10 shows the distribution of homonymous and synonymous cases. Note that for some scholars no Scopus profile was found.

Author Retrieval Once the scholar-Scopus Author IDs mapping was built, we could resort to the author retrieval Scopus API to retrieve additional information about each Scopus author profile. From the author retrieval API we collected the author name, their affiliation and their subject areas.

Publication Retrieval The list of publications for each Scopus Author ID was retrieved using the *Scopus search API*. For each publication we kept track of its title, the type of publication (Article, Conference Paper, Book...), the source name (name of the journal/conference/book) and the publication date. Note that some Scopus Author IDs had no publication associated. In our analysis we considered only papers published in academic journals between 2000 and 2016, so results were filtered by publication type and by year of publication. We also excluded publications with no journal title available.

Data Cleaning

After applying the publication type and year filter, the resulting dataset included 72 277 journal papers for the ICT field and 79 312 journal papers for the Microbiology & Genetics field. The dataset at this point still presented many cases of homonymous and synonymous names: there were 3926 Scopus profiles against 2789 scholars for the ICT field, and 3327 Scopus profiles against 2134 for the Microbiology & Genetics field. A data cleaning step was therefore necessary, in order to clear cases of homonymies while preserving synonymous profiles. Due to the size of the dataset, it was not feasible to manually inspect all cases of multiple Scopus profiles, so we attempted to automatically detect as many ambiguities as possible.

Exploring the dataset, we found that most of homonymy cases were due to the presence of scholars sharing the same name, but working in very different research areas - since a Scopus author profile can be associated with more than one area of interest, filtering query results by the subject area field was not always effective enough to avoid homonymies. Starting from this observation, the main assumption underlying the cleaning process is that these outliers usually publish their papers in very different journals compared to ICT scholars (or, respectively, to scholars working in the Microbiology & Genetics field). Thank to the bibliometric information collected from the Scopus database, we could associate each author to a "publication record", defined as a vector whose elements are equal to the number of papers published in each journal. The idea behind the outlier detection procedure is that we can assess the "similarity" between two scholars looking at the similarity between their publication records, according to some similarity metric. In our case, cosine similarity was the most suitable metric as it measures where authors publish their papers, but disregarding how many papers they publish (see discussion in Section 2.2).

For each scientific disciplinary sector, we started with a "reliable subset" of Scopus profiles which could be considered as properly assigned to scholars. Authors were added to this initial subset whenever (i) the Scopus profile was the unique result in the author search query and/or (ii) every time the city in the current affiliation field of the Scopus profile was equal to the one in the Cineca database. We regarded as outliers all authors with a little similarity with the other authors in the starting subset from the same scientific disciplinary sector (we empirically picked a threshold value equal to 0.005 by inspection of the dataset)[19] . After having automatically discarded Scopus IDs accounting for less than 5% of the scholar total publication count, the remaining ambiguous cases were manually inspected.

For practical reasons, we added a last filtering step where (i) we discarded journals with papers published by less than five different scholars and then (ii) we considered only scholars with at least five publications on this final set of journals. The final version of the dataset, used for the analysis presented in this paper, covered 47 718 papers published in 1454 journals by 2582 scholars for Information&Communication Technology, and 61 950 papers published in 2099 journals by 2111 scholars for the Microbiology&Genetics area.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.joi.2019.01.004.

References

Abramo, G., & D'Angelo, C. A. (2011). National-scale research performance assessment at the individual level. *Scientometrics*, *86*(2), 347–364.

Abramo, G., D'Angelo, C. A., & Costa, F. (2012). Identifying interdisciplinarity through the disciplinary classification of coauthors of scientific publications. *Journal of the Association for Information Science and Technology*, *63*(11), 2206–2222.

Abramo, G., D'Angelo, C. A., & Murgia, G. (2013). The collaboration behaviors of scientists in Italy: A field level analysis. *Journal of Informetrics*, *7*(2), 442–454.

[19] In all cases where scholars were left with no author associated, the author with the best similarity was kept.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*(1), 243–256.
Archambault, É, Beauchesne, O. H., & Caruso, J. (2011). Towards a multilingual, comprehensive and open scientific journal ontology. In *Proceedings of the 13th international conference of the international society for scientometrics and informetrics* (pp. 66–77).
Barabâsi, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, *311*(3-4), 590–614.
Barber, M. J. (2007). Modularity and community detection in bipartite networks. *Physical Review E*, *76*(6), 066102.
Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.
Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351–374.
Calero, C., Buter, R., Cabello Valdés, C., & Noyons, E. (2006). How to identify research groups using publication analysis: An example in the field of nanotechnology. *Scientometrics*, *66*(2), 365–376.
Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2), 224–227.
Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 269–274.
Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, *42*(1-2), 143–175.
van Eck, N., & Waltman, L. (2009). Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.
van Eck, N. J., & Waltman, L. (2017). Citation-based clustering of publications using citnetexplorer and vosviewer. *Scientometrics*, *111*(2), 1053–1070.
Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3-5), 75–174.
Fortunato, S., & Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, *104*(1), 36–41.
Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force- directed placement. *Software: Practice and experience*, *21*(11), 1129–1164.
Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826.
Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.
Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using networkx*. Los Alamos, NM (United States): Los Alamos National Lab.(LANL). Tech. rep.,
Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, *17*(2-3), 107–145.
Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the Association for Information Science and Technology*, *60*(3), 455–476.
Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, *68*(4), 984–998.
Lambiotte, R., Delvenne, J. C., & Barahona, M. (2008). *Laplacian dynamics and multiscale modular structure in networks.* , arXiv preprint arXiv:08121770.
Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the Association for Information Science and Technology*, *58*(9), 1303–1319.
Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the isi subject categories. *Journal of the Association for Information Science and Technology*, *60*(2), 348–362.
Leydesdorff, L., & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, *5*(1), 87–100.
Minguillo, D. (2010). Toward a new way of mapping scientific fields: Authors' competence for publishing in scholarly journals. *Journal of the Association for Information Science and Technology*, *61*(4), 772–786.
Moed, H. F., & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. *Scientometrics*, *101*(3), 1987–2001.
Moya-Anegón, F., Vargas- Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera- Álvarez, E., & Munoz-Fernández, F. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, *61*(1), 129–145.
Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404–409.
Newman, M. E. (2004a). Analysis of weighted networks. *Physical Review E*, *70*(5), 056131.
Newman, M. E. (2004b). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5200–5205.
Newman, M. E. (2013). Spectral methods for community detection and graph partitioning. *Physical Review E*, *88*(4), 042822.
Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113.
Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 849–856.
Noyons, E. C., & Calero- Medina, C. (2009). Applying bibliometric mapping in a high level science policy context. *Scientometrics*, *79*(2), 261–275.
Noyons, E. C. M., et al. (1999). *Bibliometric mapping as a science policy and research management tool.* Centrum voor Wetenschaps-en, Faculty of Science, Leiden University. PhD thesis.
Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *nature*, *435*(7043), 814.
Perianes- Rodriguez, A., & Ruiz-Castillo, J. (2017). A comparison of the web of science and publication-level classification systems of science. *Journal of Informetrics*, *11*(1), 32–45.
Perianes- Rodríguez, A., Olmeda-Gómez, C., & Moya-Anegón, F. (2010). Detecting, identifying and visualizing research groups in co-authorship networks. *Scientometrics*, *82*(2), 307–319.
Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(9), 2658–2663.
Rodriguez, M. A., & Pepe, A. (2008). On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Informetrics*, *2*(3), 195–201.
Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123.
Ruiz- Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, *9*(1), 102–117.
Saitta, S., Raphael, B., & Smith, I. F. (2008). A comprehensive validity index for clustering. *Intelligent Data Analysis*, *12*(6), 529–548.
Salton, G., & Buckley, C. (1987). *Term weighting approaches in automatic text retrieval.* Cornell University. Tech. rep.
Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905.
Šubelj, L., van Eck, N. J., & Waltman, L. (2016). Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PLOS One*, *11*(4), e0154404.
Traag, V., Waltman, L., & van Eck, N. J. (2018). *From louvain to leiden: guaranteeing well- connected communities.* , arXiv preprint arXiv:181008473.
Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416.
Waltman, L., & Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the Association for Information Science and Technology*, *63*(12), 2378–2392.
Waltman, L., van Eck, N. J., & Noyons, E. C. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629–635.
Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of web of science and scopus. *Journal of Informetrics*, *10*(2), 347–364.