# A new clustering method to explore the dynamics of research communities

Jordan Cambe, Sébastian Grauwin, Patrick Flandrin, Pablo Jensen

**Noname manuscript No.**
(will be inserted by the editor)

# A new clustering method to explore the dynamics of research communities

**Jordan Cambe · Sebastian Grauwin · Patrick Flandrin · Pablo Jensen.**

Received: date / Accepted: date

**Abstract** Description of temporal networks and detection of dynamic communities have been hot topics of research for the last decade. However, no consensual answers to these challenges have been found due to the complexity of the task. Static communities are not well defined objects, and adding a temporal dimension makes the description even more difficult. In this article, we propose a coherent temporal clustering method to explore the dynamics of research communities: the Best Combination of Local Communities (BCLC). Our method aims at finding a good balance between two contradictory objectives: closely following the short-term evolution by finding optimal partitions at each time step, on the one hand, and temporal smoothing, which privileges historical continuity, on the other hand. We test our algorithm on two bibliographic data sets by comparing their mesoscale dynamic description to those derived from a (static) simple clustering algorithm applied over the whole

Jordan Cambe
Univ Lyon, ENS de Lyon, UCB Lyon 1, CNRS, Laboratoire de Physique, F-69342 Lyon, France
Institut Rhônalpin des Systemes Complexes, IXXI, F-69342 Lyon, France

Sebastian Grauwin
Univ Lyon, ENS de Lyon, UCB Lyon 1, CNRS, Laboratoire de Physique, F-69342 Lyon, France
Institut Rhônalpin des Systemes Complexes, IXXI, F-69342 Lyon, France

Patrick Flandrin
Univ Lyon, ENS de Lyon, UCB Lyon 1, CNRS, Laboratoire de Physique, F-69342 Lyon, France

Pablo Jensen
Univ Lyon, ENS de Lyon, UCB Lyon 1, CNRS, Laboratoire de Physique, F-69342 Lyon, France
Institut Rhônalpin des Systemes Complexes, IXXI, F-69342 Lyon, France
E-mail: pablo.jensen@ens-lyon.fr

data set. We show that our clustering algorithm can reveal more complex dynamics than the simple approach and reach a good agreement with expert's knowledge.

**Keywords** dynamic community detection · research communities · temporal networks · bibliographic networks

## 1 Introduction

Networks are a convenient way to represent real-world complex systems, such as social interactions [29],[31], metabolic interactions [4], the Internet/world wide web [32], transportation systems [7,2], etc. For several systems it is interesting to find and describe areas of the network which are more densely connected, i.e. the communities of the network. In 20 years of complex networks history extensive work was conducted on community detection in static - non evolving - networks, see [30,3,9] and the review [10] for an overview on community detection in static graphs.

However, many networks have a temporal dimension and need a dynamic mesoscopic description at risk of non-negligible information losses if studied as static networks. Therefore the description of large temporal graphs has been a hot topic of research for the last decade, see the reviews [20] and [19] for a description of temporal networks. Most recently the detection of dynamic communities, that is communities on temporal networks, has become one of the main interests in network science, as temporal networks require to adapt the methods of static community detection. So far no consensual method was found and around 60 methods have been proposed to try to detect dynamic communities evolving with temporal networks. A total of four published reviews try to classify and summarize them [1], [18], [25] and [33].

In [33], these methods are classified into 3 main categories: (a) *instant optimal*, (b) *temporal trade-off* and (c) *cross-time*. Methods in (a) aim to detect clusters at different times t, i.e. for many snapshots of the temporal network. As these clusters are only dependent on the state of the network at time t, it is then necessary to match the communities at different t with some similarity measures, e.g. Jaccard based [27,23,16], core-node [36]. Methods in category (b) define clusters at t depending on current and past states of the network. Clusters are incrementally temporally smooth. However such methods are subject to drift as clusters are added up to each other locally. There is no compromise between temporal smoothness and 'optimal' partition at time $t$, see for example [34,17,13,12]. Finally, in category (c) clusters at $t$ depend on both past and future states of the network, see [8,28,26,11]. Clusters are temporally smooth and not subject to drift, but they do not respect causality as communities at $t$ are determined using network's information at $t+n$, i.e. communities at time $t$ can change depending on what comes next, which makes these methods inappropriate for use on-the-fly.

In this article, we present a new tool to achieve a mesoscopic description of dynamic networks, which tries to find a good compromise between 'global' and

'local' methods. The main originality of our method is to achieve a *coherent* temporal clustering procedure, which relies on a single partition methodology, namely modularity. We apply our method on two data sets of scientific articles, to show how it can describe the emergence and evolution of scientific disciplines. The main difficulties for meta-community detection methods are twofold: Finding the right temporal smoothing and quantifying the 'stability' of communities. How to know whether changes between snapshots are due to structural evolution of the community or only to algorithm instability, as many static community detection methods (used at each time $t$) are noisy (see [33] for a description of pros and cons of each clustering category). Here, we propose an algorithm which aims to find a good balance between temporal inertia (smoothness) and 'optimal' partition at any given time $t$. We compare this method to the most basic approach, which optimizes the modularity of the aggregated network using the Louvain algorithm [3]. The latter can be assimilated to a category (c) method in [33]. We then describe the methods to analyze differences between partitions: mutual information (MI) measures and bipartite network (BN) representations. We show that MI based measures are interesting but give a limited amount of information on how different two partitions are, whereas bipartite network representation allows to see how streams split between partitions. We used the methods on two bibliographic data sets: (1) the scientific publications of a scientific institution, ENS Lyon and (2) publications related to the emergence of a new mathematical tool, the 'wavelets'. We show that the global approach represents a good approximation when the dynamics is simple, i.e. when there are mainly parallel streams without much interaction, as in the ENS Lyon case. However, when the dynamics is more complex (and interesting), i.e. when many communities are born, die, split or merge, one needs a more sophisticated approach. Using the wavelets example, we show that our algorithm leads to temporal streams that compare well to a description by a domain expert of the history of this field.

## 2 Data sets

In this section, we present the specificity of each data set and the motivations for studying them. Key statistical data are summarized in table 1.

All the data sets were extracted from the ISI Web of Knowledge Core Collection database[1]. The bibliographic records were parsed and analyzed using Bibliotools, a Python-based open-source software and the historical streams figures were generated using the web-based visualisation platform BiblioMaps [14,24,15]. Bibliotools and its extension BiblioMaps were developed by one of us and are available online, as all the data analysis presented in this paper[2].

[1] *http://apps.isiknowledge.com/*

[2] *http://www.sebastian-grauwin.com/bibliomaps/*

Table 1: Statistics on data sets investigated in the paper. *Type* is the type of organization data come from. *Period* is the period over which spans the data set. $N$ is the number of publications in the data set. $N_{BC}$ is the number of articles in the BC table. $\rho_{links}$ is the density of links in the BC network. $< d >= (N_{BC} - 1) * \rho_{links}$, it indicates the average number of publications a given publication shares references with. $< w >$ indicates the average link weight. $Q$ is the modularity of the network using global partitioning (GA).

| data set | Type | Period | N | $N_{BC}$ | $\rho_{links}$ | $< d >$ | $< w >$ | $Q_{GA}$ |
|---|---|---|---|---|---|---|---|---|
| Wavelets | Thematic | 1963-2012 | 6,582 | 5,568 | 0.0065 | 35.98 | 0.000719 | 0.677 |
| ENS-Lyon | Institution | 1988-2017 | 16,679 | 14,389 | 0.0019 | 27.04 | 0.000175 | 0.919 |

### 2.1 ENS-Lyon Publications data set

The ENS-Lyon Publications data set contains all publications produced by researchers affiliated to the *École Normale Supérieure de Lyon* in natural science fields. It spans the 1988-2017 period and contains 16,679 publications. As for many scientific institutions, its publication records is highly structured by disciplinary academic departments. Here, we compare our temporal clustering methods to a partition that clusters articles according to their authors' laboratories (reference partition, $P_{REF}$).

### 2.2 Wavelets Publications data set

The Wavelets Publications data set contains all publications related to wavelets and spans from 1910 to 2012 (however the period before 1960 contains only a few publications). This data set contains 6,582 publications, corresponding to all the publications of a list of 83 key actors in the field of wavelets selected by expert advice and bibliographic searches (for more details, see [27]). The study of this data set represents a difficult task because it emerged from the collaboration of several research fields, constituted by many entangled subfields. Based on the knowledge of one of the authors (PF), a field's expert, we built manually a temporal partition drawing the history of wavelets. We refer to this partition as $P_{REF}$ and compare our automatically generated partitions to this partition of reference. We acknowledge that this partition is not an absolute ground truth as it relies on the subjectivity of an expert. However, we assume that this reference gives a reasonable picture of the field's evolution.

## 3 Methods

We start by presenting the three building blocks used in the algorithms we want to compare: how we define and partition a Bibliographic Coupling (BC) network, how we match clusters from successive time periods and finally how we define historical streams (temporal meta-clusters). We then present two

standard methods to compare partitions: Normalized Mutual Information and Bipartite Networks.

### 3.1 Building Block 1: Bibliographic Coupling partitioning

Given a set of publications on a given period, a Bibliographic Coupling (BC) network can be defined based on the relative overlap between the references of each pair of publications. More specifically, we compute Kessler's similarities $\omega_{ij} = R_{ij}/\sqrt{R_i R_j}$, where $R_{ij}$ is the number of shared references between publications $i$ and $j$ and $R_i$ is the number of references of publication $i$. In the BC network, each publication corresponds to a node and two publications $i$ and $j$ share a link of weight $\omega_{ij}$. If they don't share any reference, they are not linked ($\omega_{ij} = 0$); if they have an identical set of references, their connexion has a maximal weight ($\omega_{ij} = 1$). Here, we consider that the link between two publications is only meaningful if they share at least two references and we impose $\omega_{ij} = 0$ if they share only one reference, to avoid meaningless links between articles [14].

We use weighted links to reinforce the dense (in terms of links per publication) regions of the BC networks. This reinforcement facilitates the partition of the network into meaningful groups of cohesive publications, or communities. We measure the quality of the partition with the *modularity* Q (eq. 1), a quantity that roughly compares the weight of the edges inside the communities to the expected weight of these edges if the network were randomly produced:

$$Q = \frac{1}{2\Omega} \sum_{i,j} \left[ \omega_{i,j} - \frac{\omega_i \omega_j}{2\Omega} \right] \delta\left(c_i, c_j\right), \tag{1}$$

where $\omega_i = \sum_j \omega_{ij}$ is the sum of the weights of the edges linked to node i, $c_i$ and $c_j$ are the communities containing respectively nodes $i$ and $j$, $\delta$ is the Kronecker function ($\delta(u, v)$ is 1 if $u = v$ and 0 otherwise) and $\Omega = \frac{1}{2}\sum_{ij} \omega_{ij}$ is the total weight of edges. We compute the graph partition using the efficient heuristic algorithm presented in [3].

### 3.2 Building Block 2: Matching communities from successive time periods

Given the sets of communities $\{C_1^t, ...C_{k_t}^t\}$ in each time window $t$, the problem at hand is to identify a set of relevant historical communities, or *streams*, that correspond to a chain of communities from successive time periods (at most one per period). In order to decide which community of a given period should be added to a chain of communities from previous periods, we need to use some measure of the similarity between communities from different time periods. A standard measure is the Jaccard index, which computes the proportion of shared nodes between clusters of successive and overlapping periods (see e.g. [6,27]). One drawback of this measure is the need to use overlapping

periods which implies that there is no bijection between the publications and the streams (a given publication can belong to several streams).

Here, we take advantage of the fact that links can be computed between nodes from different time periods (publications from different periods can have common references). We could for example define a similarity measure between two clusters $C_a$ and $C_b$ from different periods either by the total sum of the links between pairs of publications from these clusters $\Omega_{a,b} = \sum_{i \in C_a, j \in C_b} \omega_{i,j}$ or by a normalized version of this sum $\omega_{a,b} = \Omega_{a,b}/|C_a||C_b|$, which is comprised between 0 and 1. While these two measures may appear quite intuitive, each of them has some drawbacks as well: using $\Omega_{a,b}$ may bias the construction of the streams by linking two 'large' (in terms of publications) but rather dissimilar (in terms of shared references) clusters. On the opposite, using $\omega_{a,b}$ may create some biases by linking two very similar clusters of very different sizes rather than the two clusters that have the second-best similarity and have similar sizes. To be coherent with our construction of clusters (maximizing the modularities within each time period), we propose here to use a modularity-based concept to match clusters from successive time periods. The similarity measure we use is thus

$$\delta Q = \Omega_{a,b} - \Omega_a \Omega_b / 2\Omega_{A,B} \tag{2}$$

where $\Omega_{A,B} = \sum_{i \in A, j \in B} \omega_{i,j}$ is the total sum of the links between pairs of publications from these two periods. $\delta Q$ thus corresponds to the change in the modularity of the BC network built from the two periods $A$ and $B$ when, starting from partitions defined on each period, clusters $a$ and $b$ are merged.

**Matching Algorithm**

Only compare pairs of communities $(a, b)$ with a minimum similarity $\omega_{a,b} > \Theta = 10^{-6}$.
Define the best match of each cluster by the one maximizing $\delta Q$.
**for** *each temporal window* **do**
  Define the **predecessor** of each cluster as its best match from the previous time period.
  Define the **successor** of each cluster as its best match from the next time period.
**end**
Two clusters are said to be *paired* if they are predecessors / successors of each other.
If a cluster is not the successor of its predecessor, we have a *split*.
If a cluster is not the predecessor of its successor, we have a *merge*.
Streams are defined as chains of paired clusters.

An illustration of this algorithm is given in Figure 1.

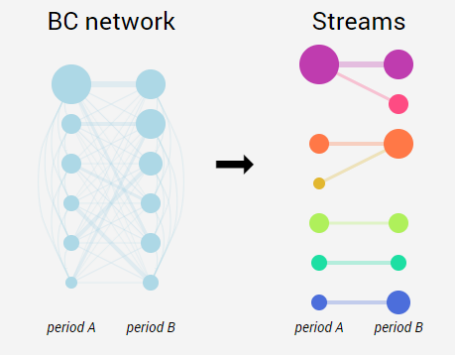



Fig. 1: Matching clusters from successive time periods. We start with the BC network built from clusters of publications detected independently in periods A and B. For each cluster $a$ of period A and cluster $b$ of period B, we compute the modularity change $\delta Q_{ab}$ obtained by grouping these two clusters in the 2-periods BC network. The 'successor' of cluster $a$ is defined as the cluster $b$ from period B maximizing this quantity and the 'predecessor' of cluster $b$ is defined as the cluster $a$ from period B maximizing this quantity. In the stream visualisation, 'paired' clusters (successor / predecessor of each other) are represented on the same y-position, and we only show the BC links between successors or predecessors, which highlight dynamical events such as merge and splits. In these figures, the size of the nodes are proportional to the number of publications in the corresponding clusters and the thickness of the links represent the average weights of the BC links between publications from two clusters.

### 3.3 Building Block 3 : Defining historical streams

We compare the results of two types of algorithms which build historical communities, or *streams*, starting from publications data sets. The first method is 'global', as it considers the whole data set to compute the communities. The second is 'local', as it starts from successive windows of $\Delta T$ years and starts by building a mesoscopic description adapted to that specific window. Hereafter, we present the results from the two main variants and refer the reader to the Annex A for more detailed results.

#### *3.3.1 Global Algorithm (GA)*

The Global Algorithm builds a *global* BC network by taking into account all the publications in the data set. Streams are defined as time evolution of these (static) communities maximizing the global modularity. Since we are working in a single (large) time period, this approach does not yield any dynamical

events such as splitting / merging of communities, but it provides a simple reference.

### *3.3.2 Best-Combination Local Communities (BCLC)*

This *local* algorithm starts by running, for each time period, $N$ independent runs (we used $N = 100$) of the Louvain algorithm. Because of the noise inherent to the Louvain algorithm, the best modularity partitions in each time period are not necessarily the ones that best match each other across successive time periods. We thus optimize the inter-period combination by the following algorithm:

**BCLC Algorithm**

Compute the Bibliographic Coupling Graph ;
Split the data set into temporal windows $\Delta t$ ;
**for** *each of the* $N = 100$ *partitions of the first period* **do**
  Run the matching algorithm with each of the $N = 100$ partitions of the second period to define the 2-periods streams;
**end**
Among the $N * N$ defined streams, select the ones maximizing the modularity of the BC network on the first 2 periods, as defined in Eq. 2;
Define the 'best combination' partitions of the first 2 periods as those corresponding to those streams;
**for** *each pair of successive temporal windows* $A$ *and* $B$*, starting from the second one* **do**
  **for** *each of the* $N = 100$ *partitions of the period* $B$ **do**
    Run the matching algorithm between these partitions and the 'best combination' partition of period $A$ (known from a previous step) to define 2-periods streams;
  **end**
  Among the $N$ defined streams, select the ones maximizing the modularity of the BC network on periods $A$ and $B$;
  Define the 'best combination' partition of period $B$ as the one corresponding to those streams
**end**

Note that maximizing a global indicator over the $T$ periods with $N$ runs would take too long as there would be $N^T$ possibilities to explore. For this reason, we choose the best combinations between the first two periods ($N^2$ checks) and then one period at a time ($N(T-2)$ checks).

This algorithm returns temporal streams we call *BCLC-streams*. These streams still maximize the modularity at each time $t$ while using some cross-time information to improve the global modularity. Figure 2 is an illustration of different runs of this algorithm. Choosing the value of the period $T$ is a

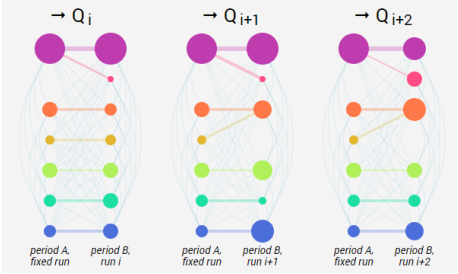



Fig. 2: Choosing the best combination. Given a set of clusters from period A, we perform the matching algorithm between these clusters and N=100 sets of clusters from period B obtained by independent runs of the Louvain algorithm. In each case, we compute the modularity $Q$ obtained by grouping paired clusters in a single clusters in the 2-periods BC network. The 'best combination' is defined as the set of clusters from period B maximizing this quantity.

trade-off. It needs to be long enough so that communities within each period have enough articles to be meaningful and limit clustering variability. But it also needs to be short enough to follow scientific dynamics. For example, the mean cited half-life of scientific articles is close to 7 years [22]. After trying different values, we chose a period $T = 5$ years, which is adapted to the temporality of scientific communities.

## 3.4 Comparing partitions

We conclude this methodological section by introducing two standard ways of comparing partitions.

### *3.4.1 Normalized Mutual Information*

The mutual information (MI) is a widely used measure for comparing community detection algorithms. It is defined as a measure of the statistical independence between two random variables (see eq. 3). In other words, if $H(P_X)$ is the entropy associated with partition $X$ and $H(P_Y)$ is the entropy associated with partition $Y$ (the entropy is a measure of how partitioned is our network, the more communities - here temporal streams - the higher the entropy), then $MI(P_X, P_Y)$ represents the overlap of the two partitions. In layman's terms, it tells us how much we know about the partition $P_X$ when the partition $P_Y$ is given. You may refer to [35,21] for a deeper description on mutual information. In particular, note that the mutual information is a symmetrical measure

$$MI(P_X, P_Y) = H(P_Y) - H(P_Y|P_X) = MI(P_Y, P_X) \tag{3}$$

The MI is defined on $[0, +\infty]$, therefore it is difficult to make sense of it without an upper-bound. There exists different ways to normalize the mutual information. The idea is to take into account the entropies of the partitions we consider to gauge the proportion of mutual information between the partitions. Normalizing by the entropy of one of the partition, e.g. $H(P_X)$ (see eq. 4) measures how much of the partition $P_X$ is included in the partition $P_Y$. We call this normalized mutual information $NMI_X$. If it reaches its maximum value 1, it means that it is possible to retrieve all the information (the partition) of $P_X$ from the partition $P_Y$. However this measure does not take into account the size of the other partition, $P_Y$. A partition $P_Y$ where each node would be its own community would make $NMI_X$ equals to 1 even though both partitions are very different. This measure then needs to be combined with at least another NMI which takes into account the relative size of both partitions (see eq. 5). Here the mutual information is normalized by $\sqrt{H(P_X) * H(P_Y)}$, which shows how much of the two entropies overlap on a scale between 0 and 1. This expresses how similar the partitions are. It is equal to 1 when the partitions are the same. Moreover, this last NMI is symmetrical, so it takes into account both retrieval of $P_X$ from $P_Y$ and retrieval of $P_Y$ from $P_X$.

$$NMI_X(P_X, P_Y) = \frac{MI(P_X, P_Y)}{H(P_X)} \tag{4}$$

$$NMI(P_X, P_Y) = \frac{MI(P_X, P_Y)}{\sqrt{H(P_X) * H(P_Y)}} = NMI(P_Y, P_X) \tag{5}$$

While Mutual information based measures give a value of similarity between two partitions, it is not straightforward to analyze. For example, it does not allow to track where the (dis)similarities come from. To allow in depth comparison, we represent pairs of partitions as bipartite networks.

### *3.4.2 Bipartite Network of streams*

To track and quantify differences between partitions $X$ and $Y$, we compute a bipartite network where the $n_X^i \in N_X$ are the first kind of nodes. They represent the streams $s_X^i \in P_X$ (hence $|N_X| = |P_X|$). It follows that the second kind of nodes $n_Y^j \in N_Y$ represent the streams $s_Y^j \in P_Y$. A weighted directed edge is drawn between $n_X^i$ and $n_Y^j$ only if their corresponding streams $s_{GA}^i$ and $s_{BCLC}^j$ share articles. For a given pair of nodes ($n_X^i$,$n_Y^j$) the weights of the two edges between them (one in each direction) are defined in eq.6. We quantify differences between streams of two partitions from this graph, quantities are given in table 3.

$$\begin{cases} w_{n_X^i \to n_Y^j} = \dfrac{|s_X^i \cap s_Y^j|}{|s_X^i|} \\ w_{n_Y^j \to n_X^i} = \dfrac{|s_Y^j \cap s_X^i|}{|s_Y^j|} \end{cases} \tag{6}$$

Table 2: $|P_X|$ is the number of streams in partition $X$. $H(P_X)$ is the entropy of partition $X$. $MI(P_X, P_Y)$ is the mutual information between the partitions $X$ and $Y$. $NMI_{REF}$ is the mutual information MI normalized by $H(P_{REF})$. $NMI(GA, BCLC)$ is the symmetrical normalized mutual information (normalized by $\sqrt{H(GA) * H(BCLC)}$).

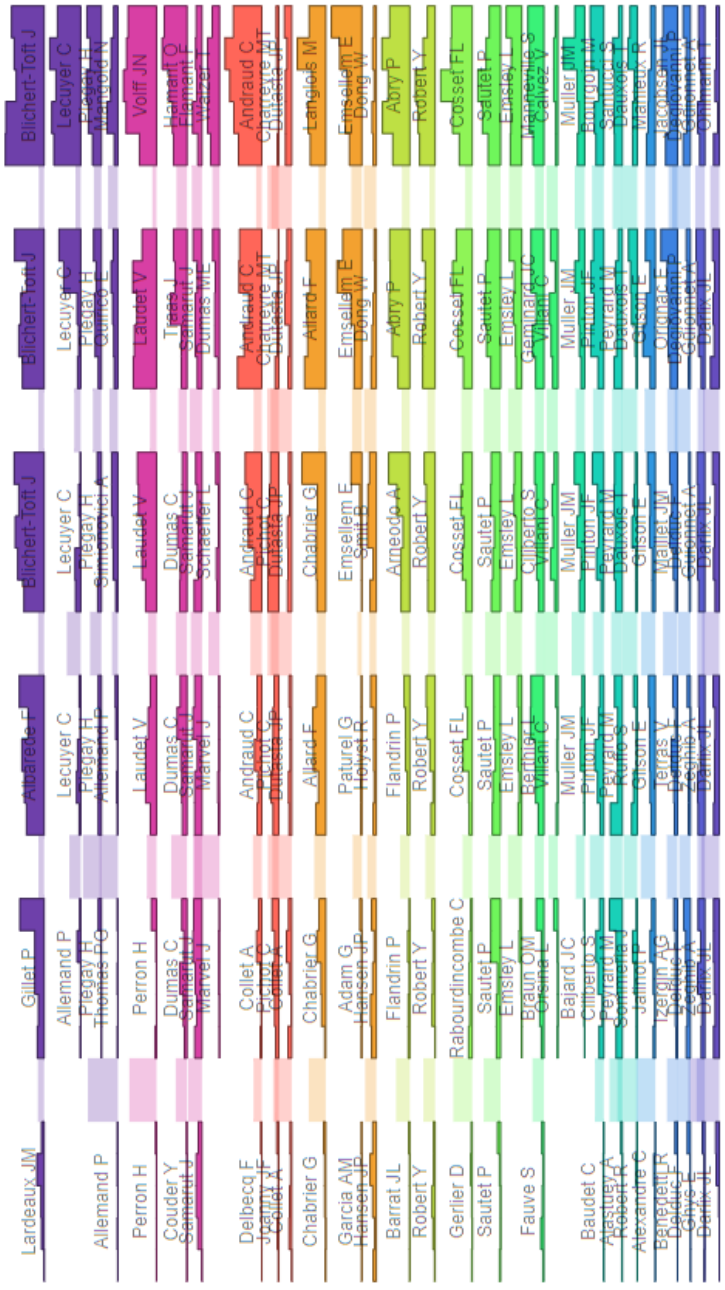
| Measures | ENS-Lyon | Wavelets |
|---|---|---|
| $\|P_{GA}\|$ | 57 | 27 |
| $\|P_{BCLC}\|$ | 97 | 36 |
| $\|P_{REF}\|$ | 17 | 36 |
| $H(P_{GA})$ | 3.63 | 2.80 |
| $H(P_{BCLC})$ | 4.00 | 2.97 |
| $H(P_{REF})$ | 2.37 | 3.18 |
| $MI(GA, REF)$ | 1.93 | 2.03 |
| $MI(BCLC, REF)$ | 1.93 | 2.49 |
| $MI(GA, BCLC)$ | 3.10 | 1.90 |
| $NMI_{REF}(GA, REF)$ | 0.82 | 0.64 |
| $NMI_{REF}(BCLC, REF)$ | 0.81 | 0.80 |
| $NMI(GA, BCLC)$ | 0.81 | 0.64 |

## 4 Results

It is difficult to represent the richness of the information conveyed by streams in paper figures. To be able to attribute scientific meaning to each of the streams, and characterize them through their main authors, references, keywords... an interactive stream visualization is available at
*http://www.sebastian-grauwin.com/streams/BCstreams.html*.

### 4.1 General features

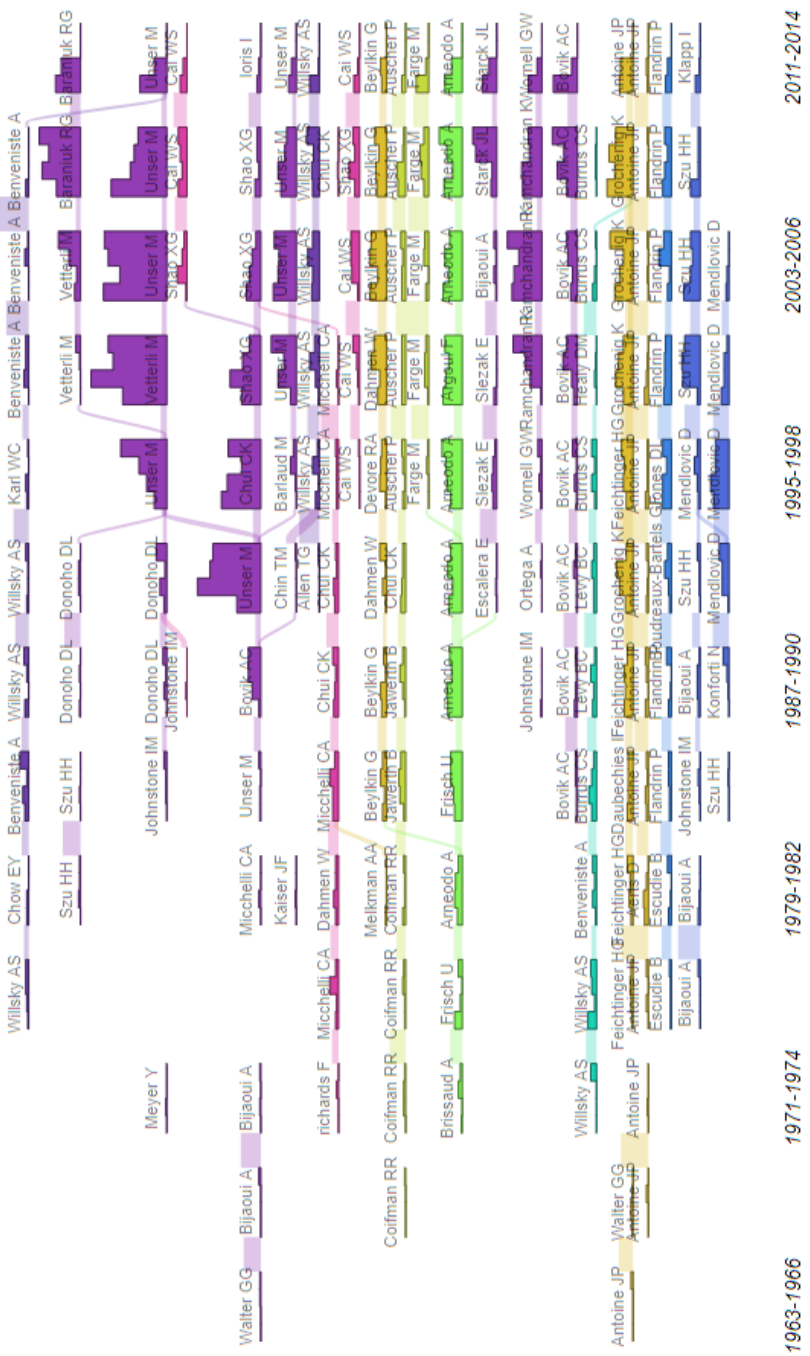
As illustrated in Figures 3 and 4, the global method cannot lead to a rich dynamics description. By construction, GA streams are well separated from each other (Figure 3a) and show only a few links in Figure 4a, which could be interpreted as splits or merges of subfields. On the opposite, BCLC streams lead to a more dynamical history for both data sets. There are only a few links Figure 3b, because different streams correspond to different scientific (sub)disciplines, which are known to be only marginally connected. However, our method rightly spots teams that split to focus on different research topics (as streams 'Blichert-Toft' and 'Lecuyer', $6^{th}$ and $7^{th}$ from the top). Similarly, many splits and merges occur in Figure 4b. To analyze these differences, we compare the partitions for each data set using the two measures defined above: Mutual information (Table 2) and community similarity from the bipartite network representation.



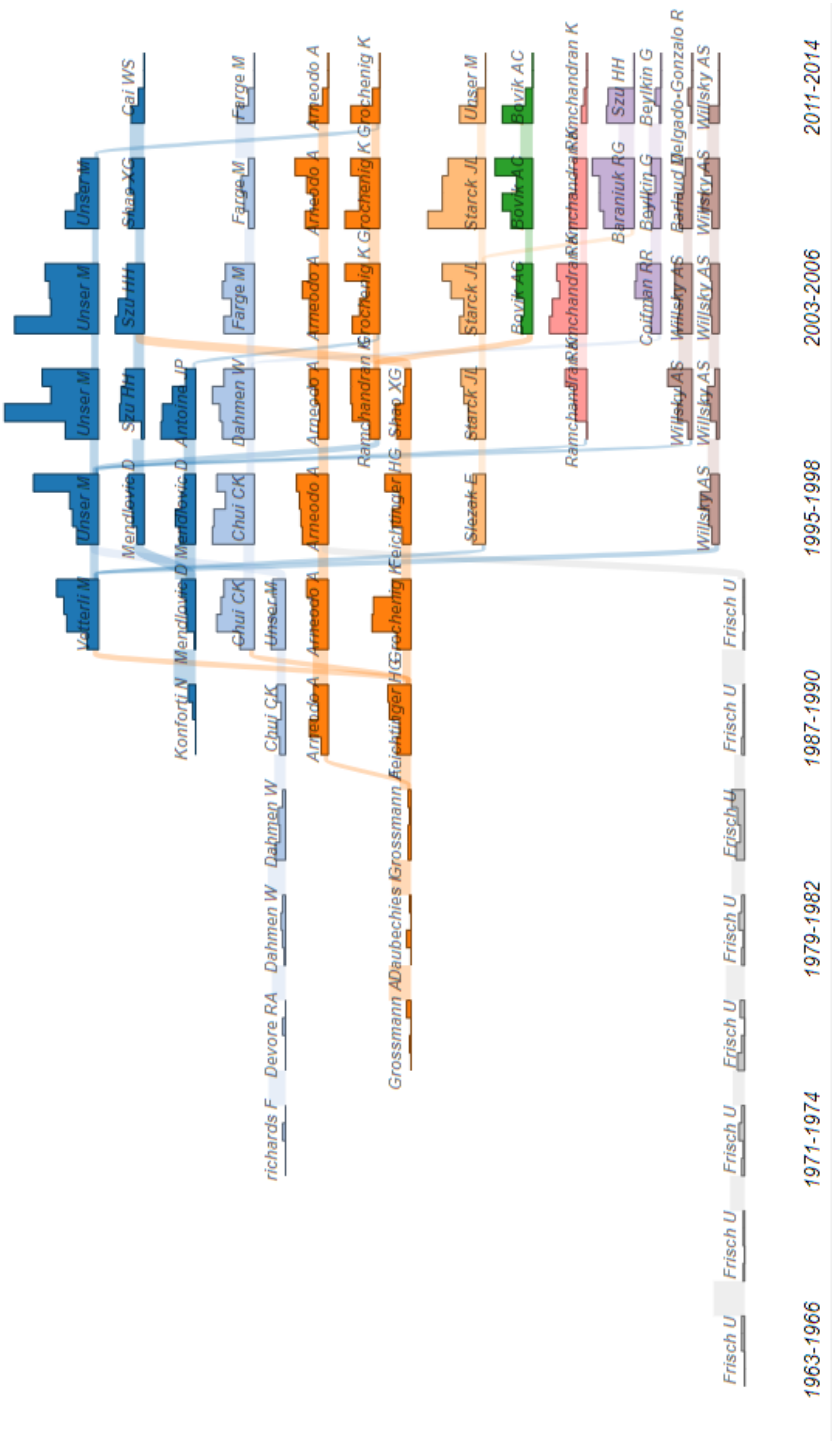

(a) Historical streams computed from the ENS Lyon natural sciences publications. Streams were determined using the global algorithm (GA).


(b) Historical streams computed from the ENS Lyon natural sciences publications. Streams were computed using our local method (BCLC).

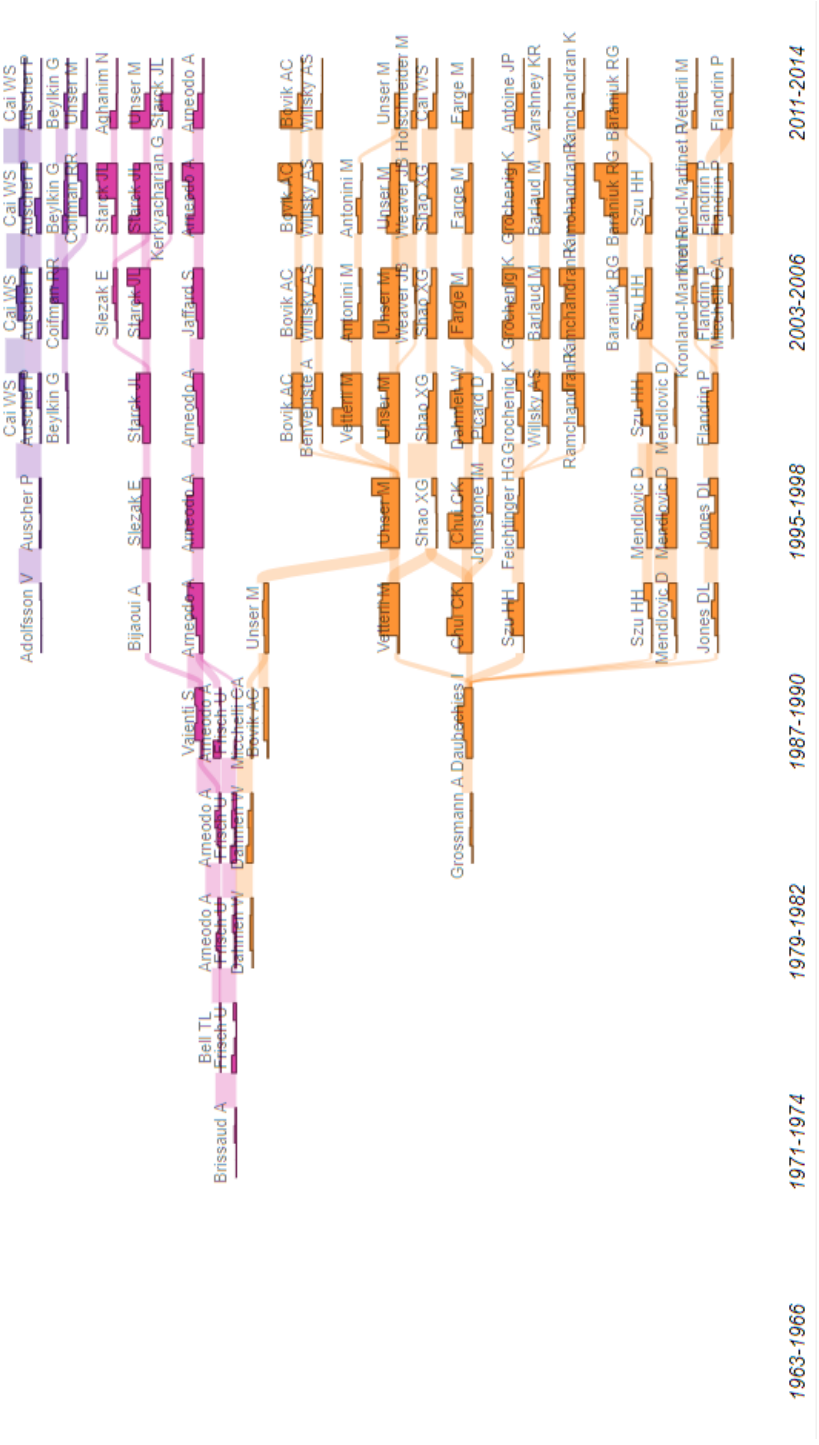
Fig. 3: Labels on each stream correspond to the most frequent author name in that stream during a given period. Streams with the same color have close research topics (here the proximity of streams to each other is computed from the weight of BC network links between clusters of a same period). Bar height is proportional to the number of publications in a given year. Links between streams show the streams that are preceding/following each other.




(a) Historical streams computed from the wavelets field of research publications. Streams determined using the global algorithm (GA).




(b) Historical streams computed from the wavelets field of research publications. Streams were computed using our local method (BCLC).




(c) Historical streams computed from the wavelets field of research publications. Streams are the reference streams determined manually.

Fig. 4: Labels on each stream correspond to the most frequent author name in that stream during a given period. Streams with the same color have close research topics (here the proximity of streams to each other is computed from the weight of BC network links between clusters of a same period). Bar height is proportional to the number of publications in a given year. Links between streams show the streams that are preceding/following each other.

### 4.2 Results on ENS Lyon data set

Table 2 shows the highly different number of streams of each partition : 57 streams for the global method, 97 for the local one and only 17 for the reference partition (the 17 laboratories of the ENS Lyon). The high values of $NMI_{REF}(GA, REF)$ (0.82) and $NMI_{REF}(BCLC, REF)$ (0.81) suggest that the extra streams in both $P_{GA}$ and $P_{BCLC}$ are mostly hierarchical subdivisions of the laboratory streams from $P_{REF}$. A partition being a subdivision of another does not result in a decrease of MI between them. The MI decreases only if communities of a partition need to be mixed to become communities of another. These results suggest that $P_{GA}$ and $P_{BCLC}$ are merely a smaller-scale division of $P_{REF}$. Similarly, the high value of $NMI(GA, BCLC)$ (0.81) suggests that $P_{BCLC}$ and $P_{GA}$ convey the same information.

The measures from Table 3 confirm this analysis. $\overline{1^{st}E}(GA, REF)$ shows that streams from $P_{GA}$ share on average $86 \pm 17\%$ of their articles with a stream from $P_{REF}$ and an average of $3.37 \pm 1.76$ streams from $P_{GA}$ are needed to retrieve 80% of streams from $P_{REF}$. Similar observations can be made for $P_{BCLC}$. Moreover, $\overline{Sum_{80}}(GA, BCLC)$ shows that it takes on average two streams from $P_{BCLC}$ to reach 80% of streams from $P_{GA}$.

Figure 5 shows a part of the bipartite network between $P_{GA}$ (left) and $P_{BCLC}$ (right) on the ENS Lyon publications data set. The part of the network is centered on nine streams from $P_{GA}$ equivalent to 17 streams from $P_{BCLC}$. It suggests that streams from $P_{REF}$ are not a mix of different streams from $P_{GA}$ or $P_{BCLC}$. They are rather unions of (almost) entire streams, which means that GA and BCLC yield almost the same partitions, but at different scales.

### 4.3 Results on Wavelets data set

#### *4.3.1 Overall comparison*

Describing the history of the wavelets research field is a complicated task as it was born from the collaboration of multiple fields and sub-fields. The values from table 2 show that, even though partitions have a similar number of streams (27 for $P_{GA}$ and 36 for $P_{BCLC}$), there are significant differences between the local and global method. In this case, $NMI(BCLC, REF)$ is significantly higher than $NMI(GA, REF)$ (0.82 vs. 0.68). Moreover $NMI(GA, BCLC)$ is rather low (0.64) which suggests that differences do not only arise from differences of scale. We visualize some of these differences in section 4.3.2. From Table 3 we see that most similar streams between $P_{GA}$ and $P_{REF}$ share $75\% \pm 20\%$ of articles on average, whereas the corresponding figure for $P_{BCLC}$ and $P_{REF}$ is $87\% \pm 17\%$.

#### *4.3.2 Examples of major differences*

We now show some major differences between $P_{GA}$ and $P_{BCLC}$ for the wavelets data set. From Figure 6, we can see two types of differences between partitions:

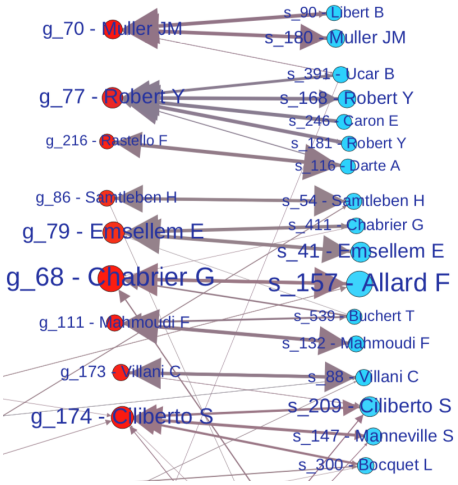



Fig. 5: Part of the bipartite network representation of ENS Lyon data set. This network shows the links between temporal communities from $P_{GA}$ (left in red) and $P_{BCLC}$ (right in blue). On each node is given the stream ID and the most frequent author name of the temporal community. Size of nodes accounts for the size of the streams, each stream contains at least 20 articles.

scale differences (e.g. g_9 with s_51 and s_166) as in the ENS Lyon case; and more significant differences, when fractions of $P_{GA}$ streams have to be combined to retrieve $P_{BCLC}$ streams (for example, the group of streams around g_5 and g_10). Interestingly, g_7 combines scale and mixing differences. Looking at the BN representation of these $P_{BCLC}$ streams with corresponding $P_{REF}$ streams (Figure 7), we see that our $P_{BCLC}$ description is quite similar to the reference description. There are more 'stream-to-stream' equivalences, represented by the double arrow on each side of the edge linking streams. Note that, though $P_{BCLC}$ is closer to $P_{REF}$, there are still scale differences (e.g. s_21, s_111) and mixing differences (e.g. s_85, s_52).

To understand the origin of the better match of $P_{BCLC}$ to the reference, it is instructive to inspect some of the differences between the local ($P_{BCLC}$) and the global partition $P_{GA}$. Let's look first at the difference between the global method stream g_7, which corresponds to a merger of four local streams, among which s_21 and s_53 (see Figure 6). Figure 4b shows that these streams do not belong to the same time period. Stream s_53 corresponds to the bottom stream (labelled 'Frisch'), and represents early works on wavelets, from 1963 to 1994, focusing on multi fractal analysis and turbulence. The second stream, s_21 (1987 - 2014, 6th from the top, labelled 'Arneodo'), addressed similar issues in a first period and then, since the early 90's, enlarged the subject matter to include mathematical formalization, together with new applications beyond turbulence, such as genome characterization. Our method appropriately distinguishes these two streams, which correspond to different subfields. The second

Table 3: In this table each cell contains two lines. Each measure $M(X,Y)$ is made on edges. The first line correspond to $M$ measured on edges from $n_X$ to $n_Y$ and the second line corresponds to $M$ being measured on edges from $n_Y$ to $n_X$. So, the first row in $\overline{1^{st}E}(X,Y)$ is the average proportion of articles $n_X$ shares with $n_Y$ $\pm$ its standard deviation. The second row is the average proportion of articles $n_Y$ shares with $n_X$ $\pm$ its standard deviation. For instance, for the ENS-Lyon, this means that streams of $P_{GA}$ share on average 86% of their articles with their most similar stream in $P_{REF}$, whereas streams from $P_{REF}$ only share on average 49% of their articles with their most similar stream in $P_{GA}$. $\overline{Sum_{80}}(X,Y)$ is the average number of streams from $P_Y$ it takes to retrieve 80% of the streams' articles from $P_X$. For example in the case of the Wavelet data set, on average $1.88 \pm 0.96$ streams from $P_{BCLC}$ are needed to retrieve 80% of a stream from $P_{GA}$.

| Measures | ENS-Lyon | Wavelets |
|---|---|---|
| $\overline{1^{st}E}(GA, REF)$ | $0.86 \pm 0.17$ | $0.75 \pm 0.20$ |
| | $0.49 \pm 0.20$ | $0.81 \pm 0.17$ |
| $\overline{Sum_{80}}(GA, REF)$ | $1.26 \pm 0.54$ | $1.88 \pm 0.93$ |
| | $3.37 \pm 1.76$ | $1.5 \pm 0.73$ |
| $\overline{1^{st}E}(BCLC, REF)$ | $0.89 \pm 0.14$ | $0.87 \pm 0.17$ |
| | $0.49 \pm 0.26$ | $0.87 \pm 0.15$ |
| $\overline{Sum_{80}}(BCLC, REF)$ | $1.23 \pm 0.44$ | $1.26 \pm 0.50$ |
| | $4.87 \pm 3.35$ | $1.31 \pm 0.57$ |
| $\overline{1^{st}E}(GA, BCLC)$ | $0.74 \pm 0.23$ | $0.72 \pm 0.23$ |
| | $0.85 \pm 0.16$ | $0.83 \pm 0.19$ |
| $\overline{Sum_{80}}(GA, BCLC)$ | $1.96 \pm 1.14$ | $1.88 \pm 0.96$ |
| | $1.34 \pm 0.51$ | $1.61 \pm 0.83$ |

difference relates to the evolution of one of the authors' (PF) activities. In the global approach, most PF articles belong to a single cluster that gathers papers in signal representations, and especially time-frequency representations that have been at the heart of his works over the years (third stream starting from the bottom in Figure 4a). This is a good approximation, but a finer description of the subjects addressed by PF during his career include three topics: (a) time-frequency methods per se, (b) relations of these methods with wavelets and (c) wavelet methods related to self-similarity, in domains such as turbulence. Using the interactive stream visualization, it is possible to look for $BCLC$ streams containing PF's publications. One finds three streams, addressing the three topics described above, and corresponding to (the stream label refers to its position in Figure 4b, starting from the top) respectively streams 3, 8 and 6. These two examples suggest that $BCLC$ is able to capture the complexity of a field dynamics', including relevant subfields, while the global approach tends to merge streams that represent different fields of inquiry.

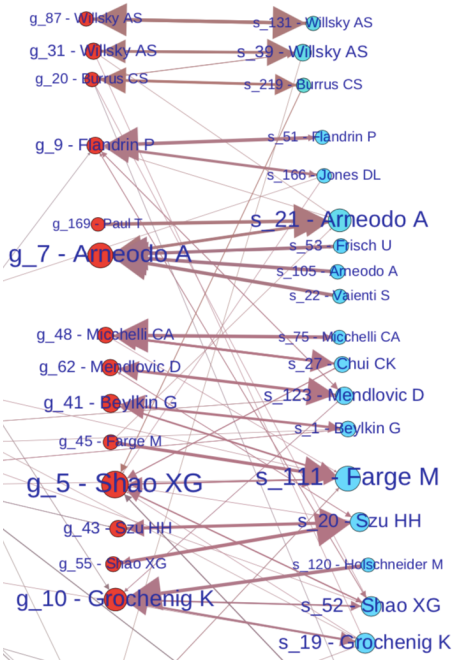



Fig. 6: Part of the bipartite network representation of Wavelets data set between temporal communities from $P_{GA}$ (left in red) and $P_{BCLC}$ (right in blue). Each node is labelled by the stream ID and the most frequent author name of the temporal community. Node size accounts for stream size.

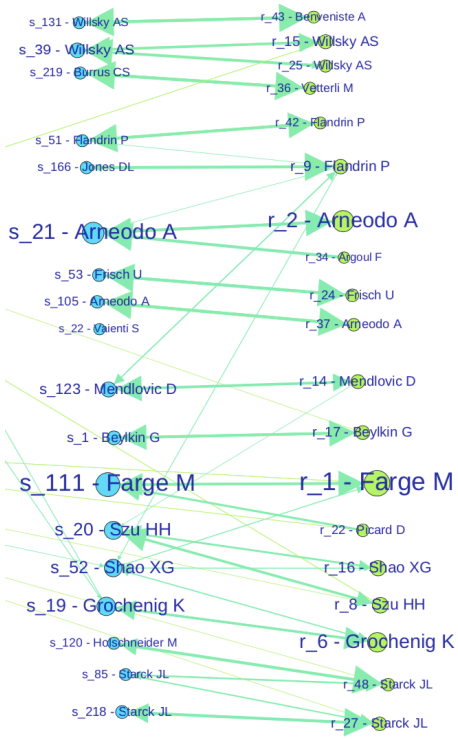



Fig. 7: Part of the bipartite network representation of Wavelets data set between temporal communities from $P_{BCLC}$ (left in blue) and $P_{REF}$ (right in green). Each node is labelled by the stream ID and the most frequent author name of the temporal community. Node size accounts for stream size.

## 5 Discussion

We have presented a coherent approach to create a dynamic mesoscopic description of a temporal network. As the standard method used to create static communities, our method only uses modularity to build the dynamic communities. We have compared our method to the static (global) approach. We first showed that both methods give the same result for networks with well-separated streams (high modularity), as in the case of ENS-Lyon publications. However, when analyzing data sets with more complex dynamics, as for the birth of wavelets (section 4.3.2), our method can generate a more satisfactory dynamics, as compared to an expert reference partition.

Clearly, much more work is needed to develop a standard approach for describing dynamical networks at a mesoscopic scale. The stochastic character of many partitioning algorithms (as Louvain's [3]), and the different scales generated by each method make comparisons difficult. Moreover, the dynamical

character of the communities renders the definition of an acceptable reference partition even trickier than for static networks.

## 6 Declarations

### 6.1 Funding

This work was supported by the ACADEMICS grant of the IDEXLYON, project of the Université de Lyon, PIA operated by ANR-16-IDEX-0005.

### 6.2 Acknowledgements

A previous version of this study has been archived as a preprint [5]

### 6.3 Conflicts of Interest

The authors declare that they have no conflict of interest.

### 6.4 Availability of Data

Data used in this article are available at:

### 6.5 Authors' Contributions

Conceptualization: P. Jensen; Methodology: J. Cambe, S. Grauwin, P. Jensen; Formal analysis and investigation: J. Cambe, S. Grauwin, P. Flandrin, P. Jensen; Writing - original draft preparation: J. Cambe, S. Grauwin, P. Flandrin, P. Jensen; Writing - review and editing: J. Cambe, S. Grauwin, P. Flandrin, P. Jensen; Funding acquisition: P. Jensen; Supervision: P. Jensen.

## References


1. Aynaud, T., Fleury, E., Guillaume, J.L., Wang, Q.: Communities in evolving networks: Definitions, detection, and analysis techniques. In: A. Mukherjee, M. Choudhury, F. Peruani, N. Ganguly, B. Mitra (eds.) Dynamics On and Of Complex Networks, Volume 2: Applications to Time-Varying Dynamical Systems, pp. 159–200. Springer New York, New York, NY (2013)
2. Barthélemy, M.: Spatial networks. Physics Reports **499**(1), 1 – 101 (2011). DOI https://doi.org/10.1016/j.physrep.2010.11.002. URL http://www.sciencedirect.com/science/article/pii/S037015731000308X
3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), P10008 (2008). URL http://stacks.iop.org/1742-5468/2008/i=10/a=P10008
4. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex networks: Structure and dynamics. Physics Reports **424**(4), 175 – 308 (2006). DOI https://doi.org/10.1016/j.physrep.2005.10.009. URL http://www.sciencedirect.com/science/article/pii/S037015730500462X
5. Cambe, J., Grauwin, S., Flandrin, P., Jensen, P.: Exploring and comparing temporal clustering methods (2020). URL https://arxiv.org/abs/2012.01287
6. Claveau, F., Gingras, Y.: Macrodynamics of Economics: A Bibliometric History. History of Political Economy **48**(4), 551–592 (2016). DOI 10.1215/00182702-3687259. URL https://doi.org/10.1215/00182702-3687259
7. Dall'Asta, L., Barrat, A., Barthélemy, M., Vespignani, A.: Vulnerability of weighted networks. Journal of Statistical Mechanics: Theory and Experiment **2006**(04), P04006 (2006). URL http://stacks.iop.org/1742-5468/2006/i=04/a=P04006
8. Duan, D., Li, Y., Jin, Y., Lu, Z.: Community mining on dynamic weighted directed graphs. In: Proceedings of the 1st ACM International Workshop on Complex Networks Meet Information & Knowledge Management, CNIKM '09, pp. 11–18. ACM, New York, NY, USA (2009). DOI 10.1145/1651274.1651278. URL http://doi.acm.org/10.1145/1651274.1651278
9. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. Proceedings of the National Academy of Sciences **104**(1), 36–41 (2007). DOI 10.1073/pnas.0605965104. URL http://www.pnas.org/content/104/1/36
10. Fortunato, S., Hric, D.: Community detection in networks: A user guide. Physics Reports **659**, 1 – 44 (2016). DOI https://doi.org/10.1016/j.physrep.2016.09.002. URL http://www.sciencedirect.com/science/article/pii/S0370157316302964. Community detection in networks: A user guide
11. Ghasemian, A., Zhang, P., Clauset, A., Moore, C., Peel, L.: Detectability thresholds and optimal algorithms for community structure in dynamic networks. Phys. Rev. X **6**, 031005 (2016). DOI 10.1103/PhysRevX.6.031005. URL https://link.aps.org/doi/10.1103/PhysRevX.6.031005
12. Görke, R., Maillard, P., Schumm, A., Staudt, C., Wagner, D.: Dynamic graph clustering combining modularity and smoothness. J. Exp. Algorithmics **18**, 1.5:1.1–1.5:1.29 (2013). DOI 10.1145/2444016.2444021. URL http://doi.acm.org/10.1145/2444016.2444021
13. Görke, R., Maillard, P., Staudt, C., Wagner, D.: Modularity-driven clustering of dynamic graphs. In: P. Festa (ed.) Experimental Algorithms, pp. 436–448. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
14. Grauwin, S., Jensen, P.: Mapping scientific institutions. Scientometrics **89**(3), 943 (2011). DOI 10.1007/s11192-011-0482-y. URL https://doi.org/10.1007/s11192-011-0482-y
15. Grauwin, S., Sperano, I.: Bibliomaps-a software to create web-based interactive maps of science: The case of ux map. Proceedings of the Association for Information Science and Technology **55**(1), 815–816 (2018)
16. Greene, D., Doyle, D., Cunningham, P.: Tracking the evolution of communities in dynamic social networks. In: 2010 International Conference on Advances in Social Networks Analysis and Mining, pp. 176–183 (2010). DOI 10.1109/ASONAM.2010.17
17. Guo, C., Wang, J., Zhang, Z.: Evolutionary community structure discovery in dynamic weighted networks. Physica A: Statistical Mechanics and its Applications

**413**, 565 – 576 (2014). DOI https://doi.org/10.1016/j.physa.2014.07.004. URL http://www.sciencedirect.com/science/article/pii/S037843711400569X
18. Hartmann, T., Kappes, A., Wagner, D.: Clustering evolving networks. In: L. Kliemann, P. Sanders (eds.) Algorithm Engineering: Selected Results and Surveys, pp. 280–329. Springer International Publishing, Cham (2016)
19. Holme, P.: Modern temporal network theory: a colloquium. The European Physical Journal B **88**(9), 234 (2015). DOI 10.1140/epjb/e2015-60657-4. URL https://doi.org/10.1140/epjb/e2015-60657-4
20. Holme, P., Saramäki, J.: Temporal networks. Physics Reports **519**(3), 97 – 125 (2012). DOI https://doi.org/10.1016/j.physrep.2012.03.001. URL http://www.sciencedirect.com/science/article/pii/S0370157312000841. Temporal Networks
21. Kvålseth, T.O.: On normalized mutual information: Measure derivations and properties. Entropy **19**(11), 631 (2017). DOI 10.3390/e19110631. URL http://www.mdpi.com/1099-4300/19/11/631
22. Lariviere, V., Archambault, E., Gingras, Y.: Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). Journal of the American Society for Information Science and Technology **59**(2), 288–296 (2008). DOI 10.1002/asi.20744. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20744
23. Lorenz, P., Wolf, F., Braun, J., Djurdjevac Conrad, N., Hövel, P.: Capturing the dynamics of hashtag-communities. In: C. Cherifi, H. Cherifi, M. Karsai, M. Musolesi (eds.) Complex Networks & Their Applications VI, pp. 401–413. Springer International Publishing, Cham (2018)
24. Lund, K., Jeong, H., Grauwin, S., Jensen, P.: Une carte scientométrique de la recherche en éducation vue par la base de données internationales scopus. Les Sciences de l'education-Pour l'Ere nouvelle **50**(1), 67–84 (2017)
25. Masuda, N., Lambiotte, R.: A Guide to Temporal Networks. WORLD SCIENTIFIC (EUROPE) (2016). DOI 10.1142/q0033. URL https://www.worldscientific.com/doi/abs/10.1142/q0033
26. Matias, C., Miele, V.: Statistical clustering of temporal networks through a dynamic stochastic block model. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **79**(4), 1119–1141 (2016). DOI 10.1111/rssb.12200. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12200
27. Morini, M., Flandrin, P., Fleury, E., Venturini, T., Jensen, P.: Revealing evolutions in dynamical networks (2017). URL https://hal.inria.fr/hal-01558219
28. Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, J.P.: Community structure in time-dependent, multiscale, and multiplex networks. Science **328**(5980), 876–878 (2010). DOI 10.1126/science.1184819. URL http://science.sciencemag.org/content/328/5980/876
29. Newman, M.: The structure and function of complex networks. SIAM Review **45**(2), 167–256 (2003). DOI 10.1137/S003614450342480. URL https://doi.org/10.1137/S003614450342480
30. Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences **103**(23), 8577–8582 (2006). DOI 10.1073/pnas.0601602103. URL http://www.pnas.org/content/103/23/8577
31. Onnela, J.P., Saramäki, J., Hyvönen, J., Szabó, G., de Menezes, M.A., Kaski, K., Barabási, A.L., Kertész, J.: Analysis of a large-scale weighted network of one-to-one human communication. New Journal of Physics **9**(6), 179 (2007). URL http://stacks.iop.org/1367-2630/9/i=6/a=179
32. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scale-free networks. Phys. Rev. Lett. **86**, 3200–3203 (2001). DOI 10.1103/PhysRevLett.86.3200. URL https://link.aps.org/doi/10.1103/PhysRevLett.86.3200
33. Rossetti, G., Cazabet, R.: Community discovery in dynamic networks: A survey. ACM Comput. Surv. **51**(2), 35:1–35:37 (2018). DOI 10.1145/3172867. URL http://doi.acm.org/10.1145/3172867
34. Rossetti, G., Pappalardo, L., Pedreschi, D., Giannotti, F.: Tiles: an online algorithm for community discovery in dynamic social networks. Machine Learning **106**(8), 1213–

1241 (2017). DOI 10.1007/s10994-016-5582-8. URL https://doi.org/10.1007/s10994-016-5582-8
35. Wagner, S., Wagner, D.: Comparing clusterings - an overview. Tech. Rep. 4, Karlsruhe (2007)
36. Wang, Y., Wu, B., Pei, X.: Commtracker: A core-based algorithm of tracking community evolution. In: C. Tang, C.X. Ling, X. Zhou, N.J. Cercone, X. Li (eds.) Advanced Data Mining and Applications, pp. 229–240. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)

# Appendices

## A Dynamics of Scientific Research Communities

We investigated four temporal community detection methods, two global and two local methods. However, as measures from GA and GPA are very close and measures from BMLA and BCLC are also very close, we only presented the GA and BCLC methods in the core of this article. The two other methods (GPA and BMLA) and their measures are described below.

### A.1 Global Projected Algorithm (GPA)

Here, we want to include some dynamics into our global algorithm. We thus start with the set of GA-streams obtained by running the Louvain algorithm [3] on the global BC network. Then, we define BC networks in each period, only keeping the articles sharing at least two references with at least one other article within the period. Removing the "long-term connections only" articles which do not share two or more references with another article in their period results in an average loss of 7.8% of the articles taken into account in the global BC network. For each time period, we define *local* communities by grouping together the publications that are in the same GA-streams, resulting in a set of local projected communities in each period. Finally, we compute historical streams by applying our matching algorithm to the projected communities. Interestingly, the streams that are build from this method do not necessarily correspond to the GA-streams: the predecessors / successors of a cluster may not be subsets of the same GA-stream of this particular cluster, resulting in splits or merges. In practice, a few GA-streams may in effect be cut into into two or more GPA-streams localized in different time periods. This approach thus allows to visualize the evolution of a GA-stream in terms of dynamical events (splits and merges).

### A.2 Best-Modularity Local Algorithm (BMLA)

For each time period, we run $N$ independent runs (we used $N = 100$) of the Louvain algorithm. Because of the noise inherent to the Louvain algorithm,

these partitions may be a bit different, while having similar modularity values (in practice the modularity difference between the partitions of different runs is lower than 0.005). Compared to the BCLC method, we do not try here to choose the partitions of the run best matching the partition from the previous or next period, but keep the partition with the best modularity among the $N$ runs in each time period. BMLA historical streams are then defined by applying the matching algorithm to these 'best-modularity' partitions.

**BMLA Algorithm**

Compute the Bibliographic Coupling Graph ;
Split the data set into temporal windows $\Delta t$ ;
**for** *each temporal window* **do**
  run $N = 100$ Louvain algorithm on the instant network;
  select the instant partition with the highest modularity $Q$;
**end**
Match the most similar communities between successive temporal windows ;
Link the paired communities along time;

This algorithm returns temporal streams we call *BMLA-streams*. These streams maximize the modularity at each time $t$ without considering the global modularity of the whole system.

### A.3 Comparing All Algorithms

Table 4 and Table 5 show there is very little difference between the local algorithms and between the global algorithms, for all measures on both data sets.

Table 4: Similarly to Table 2, $|P_X|$ is the number of streams in partition $X$. $H(P_X)$ is the entropy of partition $X$. $MI(P_X, P_Y)$ is the mutual information between the partitions $X$ and $Y$. $NMI_X$ is the mutual information MI normalized by $H(P_X)$. $NMI(P_X, P_Y)$ is the symmetrical normalized mutual information (normalized by $\sqrt{H(X) * H(Y)}$).

| Measures | ENS-Lyon | Wavelets |
|---|---|---|
| $\|P_{GA}\|$ | 57 | 27 |
| $\|P_{GPA}\|$ | 54 | 30 |
| $\|P_{BCLC}\|$ | 97 | 36 |
| $\|P_{BMLA}\|$ | 103 | 40 |
| $\|P_{REF}\|$ | 17 | 36 |
| $H(P_{GA})$ | 3.63 | 2.80 |
| $H(P_{GPA})$ | 3.63 | 2.83 |
| $H(P_{BCLC})$ | 4.00 | 2.97 |
| $H(P_{BMLA})$ | 4.04 | 3.17 |
| $H(P_{REF})$ | 2.37 | 3.18 |
| $MI(GA, REF)$ | 1.93 | 2.03 |
| $MI(GPA, REF)$ | 1.94 | 2.09 |
| $MI(BCLC, REF)$ | 1.93 | 2.49 |
| $MI(BMLA, REF)$ | 1.94 | 2.47 |
| $MI(GA, BCLC)$ | 3.10 | 1.90 |
| $NMI_{GA}(GA, REF)$ | 0.53 | 0.73 |
| $NMI_{REF}(GA, REF)$ | 0.82 | 0.64 |
| $NMI(GA, REF)$ | 0.66 | 0.68 |
| $NMI_{GPA}(GPA, REF)$ | 0.54 | 0.74 |
| $NMI_{REF}(GPA, REF)$ | 0.82 | 0.66 |
| $NMI(GPA, REF)$ | 0.67 | 0.70 |
| $NMI_{BCLC}(BCLC, REF)$ | 0.48 | 0.84 |
| $NMI_{REF}(BCLC, REF)$ | 0.81 | 0.80 |
| $NMI(BCLC, REF)$ | 0.63 | 0.82 |
| $NMI_{BMLA}(BMLA, REF)$ | 0.48 | 0.78 |
| $NMI_{REF}(BMLA, REF)$ | 0.82 | 0.80 |
| $NMI(BMLA, REF)$ | 0.63 | 0.79 |
| $NMI_{GA}(GA, BCLC)$ | 0.86 | 0.67 |
| $NMI_{BCLC}(GA, BCLC)$ | 0.77 | 0.62 |
| $NMI(GA, BCLC)$ | 0.81 | 0.64 |

Table 5: Similarly to Table 3, In this table each cell contains two lines. Each measure $M(X, Y)$ is made on edges. The first line correspond to $M$ measured on edges from $n_X$ to $n_Y$ and the second line corresponds to $M$ being measured on edges from $n_Y$ to $n_X$. So, the first row in $\overline{1^{st}E}(X, Y)$ is the average proportion of articles $n_X$ shares with $n_Y$ $\pm$ its standard deviation. The second row is the average proportion of articles $n_Y$ shares with $n_X$ $\pm$ its standard deviation. For instance, for the ENS-Lyon, this means that streams of $P_{GA}$ share on average 86% of their articles with their most similar stream in $P_{REF}$, whereas streams from $P_{REF}$ only share on average 49% of their articles with their most similar stream in $P_{GA}$. $\overline{Sum_{80}}(X, Y)$ is the average number of streams from $P_Y$ it takes to retrieve 80% of the streams' articles from $P_X$. For example in the case of the Wavelet data set, on average $1.88 \pm 0.96$ streams from $P_{BCLC}$ are needed to retrieve 80% of a stream from $P_{GA}$.

| Measures | ENS-Lyon | Wavelets |
|---|---|---|
| $\overline{1^{st}E}(GA, REF)$ | $0.86 \pm 0.17$ | $0.75 \pm 0.20$ |
| | $0.49 \pm 0.20$ | $0.81 \pm 0.17$ |
| $\overline{Sum_{80}}(GA, REF)$ | $1.26 \pm 0.54$ | $1.88 \pm 0.93$ |
| | $3.37 \pm 1.76$ | $1.5 \pm 0.73$ |
| $\overline{1^{st}E}(GPA, REF)$ | $0.87 \pm 0.16$ | $0.78 \pm 0.19$ |
| | $0.54 \pm 0.23$ | $0.83 \pm 0.17$ |
| $\overline{Sum_{80}}(GPA, REF)$ | $1.24 \pm 0.5$ | $1.65 \pm 0.84$ |
| | $3.12 \pm 1.61$ | $1.47 \pm 0.72$ |
| $\overline{1^{st}E}(BCLC, REF)$ | $0.89 \pm 0.14$ | $0.87 \pm 0.17$ |
| | $0.49 \pm 0.26$ | $0.87 \pm 0.15$ |
| $\overline{Sum_{80}}(BCLC, REF)$ | $1.23 \pm 0.44$ | $1.26 \pm 0.50$ |
| | $4.87 \pm 3.35$ | $1.31 \pm 0.57$ |
| $\overline{1^{st}E}(BMLA, REF)$ | $0.89 \pm 0.14$ | $0.85 \pm 0.19$ |
| | $0.49 \pm 0.25$ | $0.84 \pm 0.17$ |
| $\overline{Sum_{80}}(BMLA, REF)$ | $1.23 \pm 0.44$ | $1.34 \pm 0.63$ |
| | $5.0 \pm 3.60$ | $1.37 \pm 0.59$ |
| $\overline{1^{st}E}(GA, BCLC)$ | $0.74 \pm 0.23$ | $0.72 \pm 0.23$ |
| | $0.85 \pm 0.16$ | $0.83 \pm 0.19$ |
| $\overline{Sum_{80}}(GA, BCLC)$ | $1.96 \pm 1.14$ | $1.88 \pm 0.96$ |
| | $1.34 \pm 0.51$ | $1.61 \pm 0.83$ |