Functional Interactions in Complex Networks: A Three-Step Methodology for the Implementation of the Relevance Index (RI)

Riccardo Righi(✉) and Sofia Samoili

European Commission, Joint Research Centre (JRC),
Unit B6 - Digital Economy, Seville, Spain
{riccardo.righi,sofia.samoili}@ec.europa.eu

Abstract. In order to enable the management of the large presence of similar groups of agents, namely masks, resulting from the implementation of the Relevance Index (RI) algorithm, the 'PoSH-CADDy' three-step methodology is here proposed. The developed procedure is based on (i) several rounds of analysis to be performed over reducing sets of agents (with a Progressive Skimming procedure), (ii) the consideration of the overlaps among masks emerging from the output of each round (by means of a Hierachical Cluster Analysis), (iii) a final analysis of the masks remaining from the previous steps (by considering those with a minimum Degree of Dissimilarity). The methodology is implemented in a real socio-economic complex network. Insights from a first explorative analysis are provided.

Keywords: Functional interactions · Physical order
Relevance Index · Progressive skimming · Hierarchical clustering

1 Introduction

Since the widespread use of network analysis in mid 90's [1], social sciences have mostly focused on the comprehension of the structure of durable relationships (e.g. friendship) and its evolution. However, in some contexts the concept of connections is required to represent something that is more similar to a series of flickering and dynamic interactions, than to stable relationships[1]. When dynamic interactions are observed, the presence and the evolution of meso-structures[2]

The views expressed are purely those of the author and may not in any circumstances be regarded as stating an official position of the European Commission.

[1] There are cases in which, even if any new relationship is established, flickering interactions occur: people daily exchange messages with long-time friends, and enterprises repeatedly collaborate with partners they already know.

[2] In the present work, the concepts of (i) masks, (ii) groups, (iii) communities, or (iv) meso-structures, are all treated indistinctly since they all refer to subset of agents belonging to the same system.

M. Pelillo et al. (Eds.): WIVACE 2017, CCIS 830, pp. 212–228, 2018.
https://doi.org/10.1007/978-3-319-78658-2_16

can be scarcely investigated through the use of methods that are suitable for a process of stepwise creation/dissolution of connections [2]. The continuous activation/inactivation of links between agents demands the use of methodologies that, instead of considering the statistical significance of the formation/modification of the relational architectures, focus on the physical order contained in the occurred phenomena [3]. One example of a methodology that allows this is the Relevance Index algorithm, henceforth RI [4–6]. The RI, in order to investigate emergent temporal patterns in dynamic complex systems, uses a statistical approach to evaluate the significance of the integration in terms of entropy of agents' joint behaviors.

Although in complex networks analyses the detection of groups of agents is typically performed by focusing on agents' similarity or through the analysis of the network structure [7,8], the creation and the implementation of the RI algorithm provides a new approach for community detection analysis. With the RI algorithm researchers can detect groups of agents characterized by high levels of behavioral integration. These behaviors, being significantly far from randomness, are expected to reveal the presence of a common function jointly pursued by all the involved members. Since low levels of entropy are determined by the repetition of specific combinations of joint individual statuses over time, the emergence of a non-random temporal pattern unveils the alignment of the actions of these individuals towards a common function. Nevertheless, to implement the RI algorithm in dynamic complex networks that have at least some thousands of agents, and that are observed in a number of instants that is sensibly lower than the number of agents involved, additional methodological steps have to be developed. In particular, in the present work the three-step 'PoSH-CADDy' methodology is developed so as to provide a possible solution to refine the presence of redundancy in the results provided by the RI algorithm when implemented on temporal networks having the aforementioned characteristics.

In Sect. 2 in which an overview of the proposed methodology is presented. In Sect. 3 the principles of the RI algorithm are introduced. Then, in Sect. 4, the first step of the methodology is described, regarding the run of the RI algorithm several times over sets of agents progressively reducing. Then, in Sect. 4 the second step of the methodology is described, regarding the implementation of a hierarchical agglomerative cluster analysis over masks, i.e. subsets of agents belonging to the analyzed system, detected at the previous step. Section 5 follows with the third and last step of the methodology, regarding a final treatment for redundancy of masks detected in all round. Because of this last step, a final set of masks, i.e. a partition of the system, is detected. Finally, in Sect. 6 the implementation of the methodology in a case study is presented. After selecting combinations of the introduced parameters, explorative considerations are made on partitions selected according to (i) a principle of maximization of the overall percentage of agents involved in the partition[3] and (ii) a principle of minimization of the percentage of agents that belong to more than one mask.

[3] Not necessarily all agents belong to at least one masks/subset.

2 Overview of the Methodology

Acknowledging other ongoing researches with similar objectives [9,10], the present work addresses the issue of redundancy that arises by implementing RI methodology in systems with a small ratio between the number of instants in time over which agents can be observed, and the number of agents involved. More specifically, the methodology aims to identify a limited number of masks of agents, i.e. subsets of agents detected by the RI algorithm, so as to allow a final simple representation of the functional meso-structures that are present in the considered complex network. The proposed methodology is based on the following three parameters:

(1) R, i.e. the number of RI rounds of analysis that are performed,
(2) v_{OV}, i.e. a threshold used as reference to limit the presence of overlapping agents among the subset of masks finally considered from each round of RI analysis,
(3) v_{SM}, i.e. a second threshold used as reference to reduce redundancy among the masks remaining after all the previous steps.

Each parameter has a strong connection with one of the steps of the methodology: R parameter defines the length of a process of Progressive Skimming, based on the reiteration of the RI algorithm in rounds of analysis in which the best mask obtained in the previous round is dropped; v_{OV} parameter defines the development of a Hierarchical Cluster Analysis of the masks detected in each round; v_{SM} parameter defines the final process of refinements in which the remaining masks after all rounds are analyzed in terms of their Degree of Dissimilarity. The methodology is named 'PoSH-CADDy' and is summarized in Pseudo-Code 1. The refinement of the output of the RI analysis is performed attempting (i) a spread and wide exploration of the meso-structures of the system under analysis through progressive skimming, and moving towards the detection of masks that (ii) are the most significant (in terms of integration of the behaviors of the agents belonging to them), and that (iii) produce a limited degree of overlaps among them, so as to favor simplicity in the analysis of complex network's dynamics. The 'PoSH-CADDy' procedure (independently from the RI algorithm) is implemented with the R language with a CPU Intel Core i5 2.6 GHz processor and 8 GB RAM. The computational time (with R = 24 and where v_{OV}, v_{SM} is tested with 21 different values each) is approximatively of 5 h. This time period is essentially required for the computation of the distance matrices that are needed to implement the cluster analysis of each group of 15.000 masks that are detected by the RI algorithm in each round. The other steps require a computational time of some minutes. The work does not take into consideration the computational performance of the RI algorithm, as what developed applies to a procedure of refinements of its results.

Pseudo-code 1: 'PoSH-CADDy' methodology for implementation of the RI algorithm over a system $A = \{a_1, a_2, \ldots, a_n, \ldots, a_N\}$, where a_n is the n-th agent

```
function PoSH-CADDy ($R \in \mathbb{N}^+, v_{OV} \in \mathbb{R}_{\geq 0 \wedge \leq 1}, v_{SM} \in \mathbb{R}_{\geq 0 \wedge \leq 1}$)
    for each $r$ round of analysis, where $r \in \mathbb{N}^+$ and $r \leq R$, do
        Skimming of the best mask detected in the previous round
        Definition of $A_r \subseteq A$ including the agents considered in the new round
        Detection of the set of masks $\mathbb{O}(A_r)$ by means of RI analysis over $A_r$
        for each possible number of clusters, i.e. $\kappa \in \mathbb{N}^+$, in which to split $\mathbb{O}(A_r)$, do
            Hierarchical agglomerative cluster analysis for binary data
            (with simple matching coefficient and complete linkage method)
            for each obtained $k$-th cluster, i.e. $\mathbb{C}_{k,\kappa}(A_r)$, do
                Selection of the mask with the highest $t_{CI}$
            end for
            Measurement of the resulting overlaps by means of $s_{OV}(r, \kappa)$, i.e. the ratio between
            the number of agents included in at least two of the remaining masks, and the number
            of agents included in at least one of the remaining masks
        end for
        Definition of the set $\mathcal{K}_{r,v_{OV}}$, including those $\kappa$ such that $s_{OV}(r, \kappa) \leq v_{OV}$
        Selection of $\tilde{\kappa}_{r,v_{OV}}$, i.e. the highest value of $\kappa \in \mathcal{K}_{r,v_{OV}}$
        Definition of set $\mathbb{P}_{r,v_{OV}}$, by finally considering the results of the introduced cluster analysis
        with a number of clusters equal to $\tilde{\kappa}_{r,v_{OV}}$ and considering for each cluster only the mask
        with the highest $t_{CI}$
    end for
    Definition of the unique set of masks $\mathcal{P}_{R,v_{OV}}$, including all the $\mathbb{P}_{r,v_{OV}}$
    Definition of $\mathcal{P}^+_{R,v_{OV}}$, by sorting the masks in $\mathcal{P}_{R,v_{OV}}$ in decreasing order of $t_{CI}$
    Computation of dissimilarity between all masks in $\mathcal{P}^+_{R,v_{OV}}$
    for each couple of masks having a Jaccard index $> v_{SM}$ do
        Drop of the mask with the lower $t_{CI}$
    end for
    Definition of the final set $\mathcal{F}_{R,v_{OV},v_{SM}}$, including the remaining masks
end function
```

3 Principles of the RI Algorithm

The Relevance Index algorithm takes its origin from the neurological studies of Giulio Tononi in the 90's. Tononi introduced the notion of functional cluster, defining it as *a set of elements that are much more strongly interactive among themselves than with the rest of the system, whether or not the underlying anatomical connectivity is continuous* [11]. The hypothesis was confirmed as neurons with similar functions are found to demonstrate high level of coordination in their behaviors over time, independently from being (or not) situated in the same brain region [12,13]. The Cluster Index (henceforth, CI), i.e. the statistics developed and tested by Tononi in his work [12], is based on two information theory concepts derived from the Shannon entropy: Integration (I) and Mutual Information (MI). Formally, given the set $A = \{a_1, a_2, \ldots, a_n, \ldots, a_N\}$ made of N agent and a mask of agents B^m such that $B^m \subset A$, the CI of B^m is written as follows

$$CI(B^m) = I(B^m)/MI(B^m, A \setminus B^m) \tag{1}$$

where $2 \leq |B^m| < |A|$ and $0 < m \leq \xi$, with $m \in \mathbb{N}^+$ and $\xi \approx 2^{|A|}$.

Since integration and mutual information values depend on the size of the subsystem that is under analysis, a homogeneous system made of variables having the same probabilities of the variables of the original system, but that do not have correlation[4] is used [4,5,12]. Finally, the level of significance of the normalized CI, namely t_{CI}, is the value according to which the final ranking of the subsets is produced:

$$CI'(B^m) = \frac{I(B^m)}{\langle I_h \rangle} \Big/ \frac{MI(B^m, A \setminus B^m)}{\langle MI_h \rangle} \tag{2}$$

$$t_{CI} = \frac{CI'(B^m) - \langle CI'_h \rangle}{\sigma(CI'_h)} \tag{3}$$

where $\langle I_h \rangle$ and $\langle M_h \rangle$ indicate respectively the average integration of subsets of dimension $|B^m|$ belonging to the homogeneous system, and the average mutual information between these subsets and the remaining part of the homogeneous system. $\langle CI'_h \rangle$ and $\sigma(CI'_h)$, respectively the mean and the standard deviation of normalized cluster indices of subsets that have the same size of B^m and that belong to the homogeneous system, are used to compute the statistical index t_{CI}.

The concept of CI and t_{CI} was introduced in the research areas of artificial network models, of catalytic reaction networks and of biological gene regulatory systems, contributing to the identification of emergent meso-level structures [4]. Since an exhaustive computation of the t_{CI} statistic is possible only in small artificially designed networks, as those that were initially used to test the efficacy of the method [4–6], a genetic algorithm aimed to investigate the relevant subsets was implemented [6] in the RI algorithm. When implemented in large systems that can be observed in a relatively small number of instants in time, the RI algorithm produces a large number of possible B^m, which may differ among them just for the presence/absence of a single agent. As many similar masks are detected, redundancy emerges.

4 Step 1: Progressive Skimming of the Best Mask Detected

In order to address the large presence of similar masks detected in the considered system $A = \{a_1, a_2, \ldots, a_n, \ldots, a_N\}$ made of N agent, the first step that is proposed is the run of several rounds of the RI algorithm. Each round $r \in \mathbb{N}^+$, with $r \leq R$, and where $R \in \mathbb{N}^+$ indicates the number of rounds finally performed, is set to produce the detection of a same number of masks[5]. At the

[4] A homogeneous system is a system having the same number of agents of the system to which it is referred; each agent has a random generated behavior in accordance with the probability of the states it assumes in the reference system.

[5] The fact that the number of masks detected does not change, is just a choice of the researcher. This parameter could change but, since this work is not aimed at considering the increasing of the value of M, which has been fixed equal to 15.000 in each round of analysis, M is taken for given. Because of that, M will not be indexed with the number of the round r.

same time, in each r round a different set of agents is considered, namely $A_r = \{a_1, a_2, \ldots, a_{n_r}, \ldots, a_{N_r}\}$ where $A_r \subset A$ and with $|A_r| = N_r$. In order to formally describe the output of any round r of the analysis, the sets of masks detected by the RI algorithm, namely $\mathbb{O}(A_r)$, is defined according to the corresponding round of analysis. Formally,

$$\mathbb{O}(A_r) = \{B_r^1, B_r^2, \ldots, B_r^m, \ldots, B_r^M\} \tag{4}$$

where

i. $B_r^m = \{a_{n_r} \in A_r : b_{m,n_r} = 1\}$ is the m-th mask detected

ii. $b_{m,n_r} = \begin{cases} 1, & \text{if the agent } a_{n_r} \text{ is detected in the } m\text{-th mask} \\ 0, & \text{otherwise.} \end{cases}$

iii. $t_{CI}(B_r^m) \;\geq\; t_{CI}(B_r^{m+1})$

For the definition of each different set of agent A_r, a cascade process is used. Before each round, the agents belonging to best mask detected in the previous round, i.e. B_{r-1}^1, are dropped from the analysis, such that the cardinality of the set of agents considered, i.e. A_r, decreases after each round. Formally, each A_r can be described as

$$A_r = A \setminus \bigcup_{q=0}^{r-1} B_q^1 \tag{5}$$

where $q \in \mathbb{N}$ indicates one of the rounds preceding the r-th round[6], and where $0 \leq q \leq (R-1)$. Therefore, $A_r \subset A_{r-1}\ \forall\ r$. As in each round, the set that is analyzed with the RI algorithm does not include any of the best masks detected in the previous rounds, this procedure is called ‘progressive skimming’.

The RI algorithm produces a list of ordered binary masks that may differ among them just for the presence/absence of one single agent. Therefore, for the best detected mask of agents, i.e. B_r^1, also many similar masks are detected (as they are likely to perform well also from a point of view of entropy) in $\mathbb{O}(A_r)$. Because of this redundancy[7] the progressive skimming of masks is implemented, so as to perform an extended exploration of the system. This procedure, even if it deals with a loss of information and a reduction (and so also change) of the considered system when continuing the analysis round after round, allows the researcher to analyze how the rest of the system works independently from what in the previous rounds has been detected the group of agents with the most integrated behaviors. Interactions between the best mask detected in round r and masks detected in following rounds are limited, since the agents belonging to B_r^1

[6] Since the initial round that is performed is $r = 1$, if $r = 1 \rightarrow q = 0$. As there is no round 0, if $q = 0 \rightarrow B_0^1 = \varnothing$. Therefore, from Eq. 5, when $r = 1$, we have that $A_1 = A \setminus B_0^1 = A \setminus \varnothing = A$.

[7] Furthermore, the problem of redundancy in $\mathbb{O}(A_r)$ does not affect only the best mask B_r^1. It is important to remark that it is also present for masks different from the best one. Therefore, it can be said that when the system is large, in each $\mathbb{O}(A_r)$ a lack of variety comes up.

are removed from the sets of agents that is going to be analyzed in the rounds following the r-th. However, because of the implementation of a hierarchical agglomerative cluster analysis (Sect. 5), in each round r also all the other masks different from B_r^1 are taken into account. Therefore, the progressive skimming does not imply that the best mask B_r^1 stands in a condition of isolation. If mask B_r^1 has significant intersections/interactions with other masks B_r^m detected in the same round, evidences should appear in the cluster analysis of the whole $\mathbb{O}(A_r)$. In contrary, if masks substantially different from B_r^1 do not emerge from the cluster analysis, some clues of a functional detachment between the agents in B_r^1 and the agents that belong to the rest of the system are detected.

5 Step 2: Clusters of Masks Within Each r-th Round

5.1 The Cluster Analysis of Masks in $\mathbb{O}(A_r)$

With the Simple Matching Coefficient (SMC) distance is measured between couples of masks, and with the Complete Linkage (CL) criterion for the progressive merging of clusters, a hierarchical agglomerative cluster analysis is then implemented. This analysis is here represented by the function θ_κ assigning each mask B_r^m to one (and only one) cluster. Formally,

$$\theta_\kappa^{SMC,CL}(B_r^m) = k \tag{6}$$

where $k \le \kappa$, with $k \in \mathbb{N}^+$ indicating the specific cluster to which each mask $B_r^m \in \mathbb{O}(A_r)$ is assigned through the hierarchical cluster analysis (with SMC and CL) in which the masks of $\mathbb{O}(A_r)$ are allocated in a number of clusters equal to $\kappa \in \mathbb{N}^+$. Since the number of clusters is not established a-priori, at this stage the definition of each cluster, namely $\mathbb{C}_{k,\kappa}(A_r)$, has to take into account the fact that κ can vary. Therefore, each cluster $\mathbb{C}_{k,\kappa}(A_r)$ is formally defined as

$$\mathbb{C}_{k,\kappa}(A_r) = \{B_r^m \in \mathbb{O}(A_r) : \theta_\kappa(B_r^m) = k\} \tag{7}$$

where $\mathbb{C}_{k,\kappa}$ is the k-th cluster, obtained by dividing in κ clusters the masks contained in $\mathbb{O}(A_r)$.

5.2 The Selection of a Representative Masks for Each Cluster

For any cluster obtained, only the mask with the highest t_{CI} is considered, as representative of the cluster itself. Formally, this mask, henceforth indicated as $\tilde{B}_{r,k,\kappa}$, has the following properties:

$$\tilde{B}_{r,k,\kappa} \in \mathbb{C}_{k,\kappa}(A_r) \quad \text{and} \quad t_{CI}(\tilde{B}_{r,k,\kappa}) = max\ t_{CI}(\mathbb{C}_{k,\kappa}(A_r)). \tag{8}$$

Therefore, each cluster is represented by the mask that, belonging to it, is also the one whose agents present a joint behavior that is the significantly farthest from randomness. By adopting this criterion, the principles underpinning the RI algorithm are respected. Even if several different combination may be present, the analysis of the similarity reveals groups of masks that have to be intended just as possible modification of the one of reference, i.e. the most relevant one.

5.3 Overlaps and the s_{OV} Statistic

The cluster analysis of $\mathbb{O}_r(A_r)$ and the selection of the mask with the highest t_{CI} for each cluster, can produce the affiliation of agents to more than one masks[8]. In order to set the value of κ, i.e. to determine the number of clusters, a criterion concerning the limitation of the progressive emergence of overlaps in the observed structure of masks is adopted. In order to understand which degree of overlap is associated with the values of κ, starting from 1 and continuing in increasing order, the statistic $s_{OV}(r, \kappa)$, where the subscript 'OV' stands for OVerlaps, is computed as

$$s_{OV}(r, \kappa) = \frac{|\bigcup\limits_{k_\alpha, k_\beta=1}^{\kappa} (\tilde{B}_{r,k_\alpha,\kappa} \cap \tilde{B}_{r,k_\beta,\kappa})|}{|\bigcup\limits_{k=1}^{\kappa} \tilde{B}_{r,k,\kappa}|} \qquad \forall\ k_\alpha \neq k_\beta \tag{9}$$

where $k_\alpha, k_\beta \in \{1, \ldots, k, \ldots, \kappa\}$ are the indices of two distinct clusters $\mathbb{C}_{k,\kappa}(A_r)$, obtained by implementing the function θ_κ over the set of masks $\mathbb{O}(A_r)$. The statistic $s_{OV}(r, \kappa)$ calculates, for each possible value of r and of κ, the ratio between the number of agents that belong to at least two masks (numerator) and the number of agents that belong to at least one mask (denominator). The introduced statistic aims to evaluate the degree of simplicity associated to each possible number value of κ, i.e. the number of clusters in which to group the masks included in $\mathbb{O}(A_r)$. The simplicity lies on the fact that masks have to be recognizable and distinct from each other. If the structure of the detected masks is characterized by a high degree of overlap, the masks are so intertwined that they cannot be assumed as unitarity entities and the representation of the whole system, that they are suppose to provide, is finally unreadable.

5.4 Selection of the Number of Clusters by Means of v_{OV} Parameter

In order to define the value of κ, i.e. the number of clusters in which to split each set of masks $\mathbb{O}_r(A_r)$, the criterion adopted lies in the comparison between the statistic $s_{OV}(r, \kappa)$, defined by Eq. 9, and a percentage threshold used as reference, namely $v_{OV} \in \mathbb{R}_{\geq 0}$, with $0 \leq v_{OV} \leq 1$. Given a specific value of v_{OV}, the value κ is chosen in order to have the highest number of clusters among those to which corresponds a $s_{OV}(r, \kappa)$ lower than, or equal to, the percentage threshold v_{OV}. For each r-th round, a set of possible value of κ is so selected. These sets, namely $\mathcal{K}_{r,v_{OV}}$, are formally described as follows.

$$\mathcal{K}_{r,v_{OV}} = \{\kappa \in \mathbb{N}^+ :\ s_{OV}(r, \kappa) \leq v_{OV}\} \tag{10}$$

For each round r, depending on the threshold v_{OV}, all the values of κ that produce a partition for which the percentage of agents that belong to more

[8] The allocation in one exclusive cluster does not concern agents. The same agent can be detected in two masks that are not included in the same cluster.

than one group (up to the number of agents overall included) is less or equal to the considered threshold v_{OV}, are considered admissible. Then, among all the elements contained in $\mathcal{K}_{r,v_{OV}}$, the value $\tilde{\kappa}_{r,v_{OV}}$, i.e. the final value in which finally to split the resulting masks contained in $\mathbb{O}_r(A_r)$ given the specific threshold v_{OV}, is defined as

$$\tilde{\kappa}_{r,v_{OV}} = \max\ \mathcal{K}_{r,v_{OV}} \tag{11}$$

By identifying $\tilde{\kappa}_{r,v_{OV}}$, the highest number of cluster, given the threshold v_{OV}, is selected. Therefore, the soft partition[9] obtained in any of the r rounds, namely $\mathbb{P}_{r,v_{OV}}$, and can be formally defined as

$$\mathbb{P}_{r,v_{OV}} = \{\tilde{B}_{r,k,\kappa} \in \mathbb{O}(A_r)\ :\ \kappa = \tilde{\kappa}_{r,v_{OV}}\} \tag{12}$$

6 Step 3: Final Treatment of Redundancies

6.1 The Set of Masks Resulting from All the Rounds: $\mathcal{P}_{R,v_{OV}}$

At the end of an entire process of analysis always[10] with the same value of the parameter v_{OV}, R sets of masks are obtained, and each of them is identified by the corresponding $\mathbb{P}_{r,v_{OV}}$. Therefore, since the analysis is developed with a specific value of R and a specific value of v_{OV}, it is possible to assemble all the masks in a unique set, namely $\mathcal{P}_{R,v_{OV}}$, that can be formally defined as

$$\mathcal{P}_{R,v_{OV}} = \{\tilde{B}_{r,k,\kappa} \in \bigcup_{r=1}^{R} \mathbb{O}(A_r)\ :\ \kappa = \tilde{\kappa}_{r,v_{OV}}\} \tag{13}$$

where $\tilde{\kappa}_{r,v_{OV}}$ is the number of clusters in which the specific $\mathbb{O}(A_r)$ is divided, as a result of the process described in Eqs. (8–11), and where, as explained in Eq. (8), the tilde (˜) over the mask $B_{r,k,\kappa}$ indicates that in the cluster of masks to which it belongs, i.e. $\mathbb{C}_{k,\kappa}(A_r)$, the mask $\tilde{B}_{r,k,\kappa}$ presents the highest t_{CI}. Once the set $\mathcal{P}_{R,v_{OV}}$ is defined, the last issue addresses the consequence of having implemented a reiterated procedure of analysis, i.e. multiple rounds of the RI algorithm. As at the beginning of each round r exclusively the agents belonging to B^1_{r-1} are dropped, the presence of similar masks (among all those detected in an entire process of analysis) is not prevented[11]. The following, and last, steps aim to manage this redundancy.

[9] A soft partition is intended to be a set of masks of agents that do not necessarily belong to exclusively one masks. Therefore, as explained above, an agent can belong to more than one mask.

[10] From the first round $r = 1$, to the last round $r = R$.

[11] The set $\mathcal{P}_{R,v_{OV}}$ can present redundancies since, even if the rest of the system that at each new round r is analyzed does not include the best masks detected in round $(r-1)$, it can include the agents that belong to the second/third/etc. masks detected in the round $(r-1)$. Therefore, it could happen that those masks that were detected as second/third/etc. masks in $(r-1)$, are detected also in the round r.

6.2 Sorting the Masks of $\mathcal{P}_{R,v_{OV}}$ in Decreasing Order of t_{CI}

All the masks belonging to $\mathcal{P}_{R,v_{OV}}$ are sorted in decreasing order, according to the value of their t_{CI}. In this way, from the set of masks $\mathcal{P}_{R,v_{OV}}$, the sorted set of masks $\mathcal{P}^{+}_{R,v_{OV}}$ is generated. Formally,

$$\mathcal{P}^{+}_{R,v_{OV}} = \{\tilde{B}^{(1)}_{R,v_{OV}}, \tilde{B}^{(2)}_{R,v_{OV}}, \dots, \tilde{B}^{(j)}_{R,v_{OV}}, \dots, \tilde{B}^{(J)}_{R,v_{OV}}\} \tag{14}$$

where $|\mathcal{P}_{R,v_{OV}}| = J$, and $\tilde{B}^{(j)}_{R,v_{OV}}$ is one of the masks belonging to $\mathcal{P}_{R,v_{OV}}$ and that were previously indicated as $\tilde{B}_{r,k,\tilde{\kappa}_{r,v_{OV}}}$. Moreover, the index in the superscript, i.e. $j \in \mathbb{N}^{+}$ where $j \le J$, refers to the ordinality of the masks of $\mathcal{P}^{+}_{R,v_{OV}}$, so that is true the condition $t_{CI}(\tilde{B}^{(j)}_{R,v_{OV}}) \; > \; t_{CI}(\tilde{B}^{(j+1)}_{R,v_{OV}})$.

6.3 Final Drop of Similar Masks According to the Paramater v_{SM}

Once the masks are ordered according to their t_{CI}, a final analysis of their similarity is performed. Starting from the best mask $\tilde{B}^{(1)}_{R,v_{OV}}$, all the masks that are too similar to it are dropped. Then, the same procedure is repeated in cascade process. The second best mask of those remaining is then compared with those having a lower t_{CI}, and so forth with the third best mask remaining, the fourth, etc. This procedure continues until there are no more masks that can be used as a reference. In this way, only masks that have a minimum degree of dissimilarity are kept.

The similarities between masks are calculated in terms of JaCcard Index[12] (henceforth, JC), i.e. the percentage of the number of agents in the intersection of the two considered masks (up to the number of agents in the union set of the same two masks). Then, the set of masks $\mathcal{P}^{+}_{R,v_{OV}}$ is filtered using a threshold regarding SiMilarity, namely $v_{SM} \in \mathbb{R}_{\ge 0}$, with $0 \le v_{SM} \le 1$. The resulting (and final) set of masks, namely $\mathcal{F}_{R,v_{OV},v_{SM}}$, can be formally defined as

$$\mathcal{F}_{R,v_{OV},v_{SM}} = \{\tilde{B}^{(j)}_{R,v_{OV}} \in \mathcal{P}^{+}_{R,v_{OV}} \;:\; JC(\tilde{B}^{(i)}_{R,v_{OV}}, \tilde{B}^{(j)}_{R,v_{OV}}) \; < \; v_{SM}\} \tag{15}$$

where

i. $\tilde{B}^{(i)}_{R,v_{OV}} \in \mathcal{P}^{+}_{R,v_{OV}}$,
ii. i and j, where $i \in \mathbb{N}$ and $j \in \mathbb{N}^{+}$ and $0 \le i < j$, are used to indicate the mask of $\mathcal{P}^{+}_{R,v_{OV}}$ by making reference to their ordinality, as described in Eq. 14,
iii. $\tilde{B}^{(0)}_{R,v_{OV}} = \varnothing$, so that $JC(\tilde{B}^{(0)}_{R,v_{OV}}, \tilde{B}^{(1)}_{R,v_{OV}}) = 0$.

[12] While in Step 2 of the proposed methodology the SMC is used to evaluate similarity (see Sect. 5), in this Step the JC is considered as more appropriate. JC focuses its attention on the intersection of two masks (with regard of the union set), while SMC considers as a condition of similarity also the simultaneous absence of a same element. While in Step 2 was important to consider also the co-absence of agents as an element of similarity, so as to evaluate where the algorithm had moved (in terms of agents considered and not considered), here only the presence of overlapping agents, i.e. the intersection, is relevant.

7 Case Study - Region Tuscany Innovation Policies

In this Section, an example of an implementation of the RI+PoSH-CADDy methodology in an empirical analysis, is presented. The considered case study addresses a regional programme implemented by Tuscany Region (Italy) in the period 2000–2006, aiming to support innovation projects. The considered network policy programme sustained the development of innovation processes by fostering interactions between local agents (enterprises, universities, public research centers, local government institutions, service centers, etc.) [14–16]. Starting in 2002 (and ending in 2008), the programme of public policies was consisted of nine waves not uniformly distributed over time: they had different durations and they overlapped, producing periods in which no wave was active, and periods in which three waves were simultaneously active. The degree of formation and of dissolution of connections was so high that resulted in a situation of intense discontinuity over time. Therefore, a new appropriate tool that does not investigate the flourishing of communities looking at the stepwise creation of network frameworks, was deemed necessary [2]. Moreover, by using the RI algorithm the analysis could take into account the presence of functional meso-structures. Finally, because of the objective of policies taken into consideration, i.e. fostering of innovative processes, the focus on interactive dynamics, more than on network's relational architectures, is even more meaningful[13] [17].

7.1 Available Data and Pre-processing

The most important aspect regarding the implementation of RI analysis in the present case study regards the definition of the informational basis describing agents' statuses of activity. Since the available data contains information on the starting and the ending dates of agents' participations in the projects, it is possible to define a set of 59 instants in time[14] to observe the system. With these dates, a complete behavioral profile for each of the agents involved in the policy programme is structured. In each instant, the number of projects in which each agent was active is considered. A series of 58 variables is generated taking into account how the levels of activity vary from one instant to the following

[13] In this case study, the agents' activities coincide with interactions. Agents are considered to be active when they are participating in a project. And since in each project partnerships have to be established (no single-participant projects are allowed), it follows that to be active implies to be interacting.

[14] Considering all the dates of starting and the ending of the projects, 59 different dates were identified.

one[15]. Regarding the size of the system, the agents participating in the described policies of Region Tuscany are 1121, and the majority of them participated just in one project. The scarcely active agents are removed from the analysis. The focus is set on those with a minimum degree of activity. Therefore, only agents that at least participated in 2 projects are considered. Finally, 352 agents remain. These agents constitute the initial set of the analysis, namely A.

7.2 Setting the Parameters

The RI analysis with the PoSH-CADDy methodology is implemented over the set A, consisted of 352 agents, observed in 58 instants over time. The number of rounds to be performed, i.e. the parameter R, is set equal to 24. With the progressive skimming, as described in Sect. 4, A_{24} is consisted of 204 agents. Therefore, A is extensively explored, as the procedure is stopped after having removed the 45.17% of the agents initially involved. Regarding the threshold v_{OV}, since no specific theoretical reasons suggest the a-priori identification of a specific value, a discrete set V_{OV} of percentage thresholds is used and each $v_{OV} \in V_{OV}$ is considered to implement a process of analysis. The set V_{OV} is defined as follows:

$$V_{OV} = \{v_{OV} \in \mathbb{R}_{>0} : \; v_{OV} = \frac{1}{40}\, x\} \quad \forall\, 0 \leq x \leq 20,\; x \in \mathbb{N} \tag{16}$$

Regarding the setting of the threshold v_{SM}, the same set of conditions are applied also for v_{OV}. A discrete set V_{SM} is created in the same way of V_{OV} and each $v_{SM} \in V_{SM}$ is considered to implement the analysis. For both, values larger than 0.5 are not taken into consideration as, in principle, they go in the opposite direction of the general objective of the present work, that is to reduce redundancy[16].

As one value for R, and 21 values for v_{OV}, and 21 values for v_{SM} are considered, 441 different $\mathcal{F}_{R,v_{OV},v_{SM}}$ are finally computed. Each of these final sets of RI masks constitute a soft partition[17] of the system A. In Fig. 1a, the 441

[15] These variables assume four different values that correspond to one of the following four situations: inactivity, decreasing activity, stable activity or increasing activity. The ‘activity’ status is defined by considering the number of projects in which the agent is participating in the corresponding instant, with regard to the number of projects in which it was participating in the previous instant. With these series of variables, a second order Markov condition in taken into account, since agents' activity is not described just for what is in each instant, but for what it is in the present conditioned to what it was in its nearest past. As a variation in time is considered, the number of variables finally computed equals the number of variables initially present minus 1.

[16] To have more than the 50% of agents producing an overlaps among the masks of a generic $\mathbb{P}_{r,v_{OV}}$, or to allow in $\mathcal{F}_{R,v_{OV},v_{SM}}$ couples of masks generating an intersection that is the 50% or more of the corresponding union set, has been considered as not pertinent for the objective of this work.

[17] Overlaps among groups (determined by the fact that each agent can belong in more than one group) are allowed and are present.

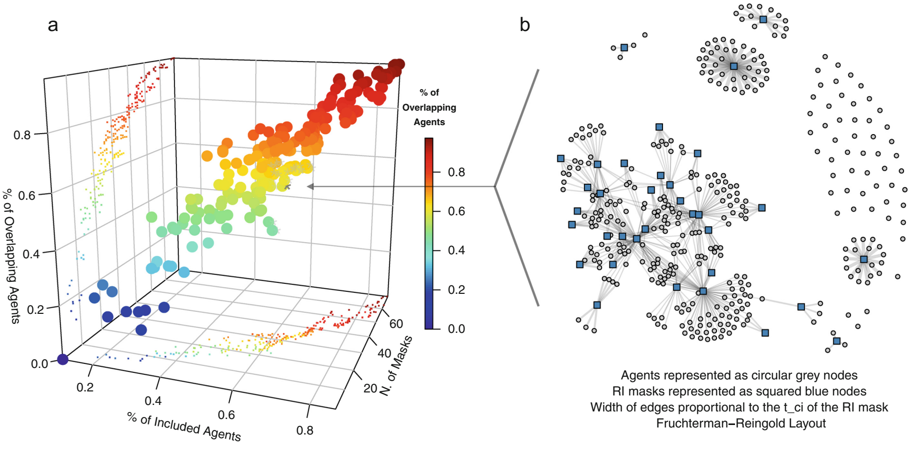

Fig. 1: (a): Colored dots represent the final partitions obtained, i.e. all the $\mathcal{F}_{R,v_{OV},v_{SM}}$ resulting from the possible combinations of the three parameters $R = 24, v_{OV} \in V_{OV}$ and $v_{SM} \in V_{Sm}$, as described in Eq. (16). The y-axis describes the percentage of agents that, in the corresponding $\mathcal{F}_{R,v_{OV},v_{SM}}$, belong to more than one mask (up to the number of agents that belong to at least one mask). The x-axis describes the percentage of agents that belong at least to one mask (up to the total number of agents included in the initial considered set A). The z-axis describes the number of masks that are present in the corresponding partition. The color of the 441 dots is in accordance with the % of overlapping agents. Small grey asterisks indicate the 47 partitions that include at least the 60% of agents of A, and that have less than the 70% of overlapping agents. Big colored dots are projected on the lateral and on the bottom faces of the cube delimiting the three-dimensional space. (b): Bipartite graph representing affiliations of agents (of set A) in the specific final set of RI masks $\mathcal{F}_{24,0.325,0.225}$ (indicated in the 3D representation on the left, with a darker grey asterisk). Grey circular nodes represent agents, and blue squared nodes represent RI mask. The width of edges is proportional to the t_{CI} of the mask. (Color figure online)

obtained $\mathcal{F}_{R,v_{OV},v_{SM}}$ are illustrated in a three-dimensional space describing the number of agents included (up to the total number of agents included in A), the percentage of agents belonging to more than one mask (up to the number of agents overall included in the partition), and the number of masks included in the corresponding partition.

7.3 Exploration of the Results

As represented in Fig. 1a, the considered combinations of the three parameters lead to different $\mathcal{F}_{R,v_{OV},v_{SM}}$. Even though currently evaluation on the parameters' space is not effectuated, the choice of the value to be considered is attributed with an a-posteriori unbiased procedure. In the present work, only those partitions including (i) at least 60% of the agents of the initial set A, and (ii) having less than 70% of agents belonging to more than one community, are considered. The parameters' space is narrowed in order to address two objectives, which both concern the readability of the final representations of the system. These objectives are: (i) to consider partitions in which a large part of the initial system is analyzed, and (ii) to avoid the selection of partitions in which extreme overlapping of the detected subsets prevents a simple interpretation of the system. Statistics regarding the feature of the single masks are not taken into account, and a-priori biased considerations on the values of v_{OV} and v_{SM} are not made. Currently, the parameters' space is not explored with a standardized method. However, the parameters are not selected based on the properties of the single masks, so as to avoid bias.

Based on the aforementioned conditions, 47 partitions (up to 441) are identified. These partitions are indicated with grey asterisks in Fig. 1a. In order to proceed with the exploration of the first results provided by the methodology, the presence of similar features within all the groups of 47 partitions is suggested. Currently, only one is heuristically selected, namely the partition with $v_{OV} = 0.325$ and $v_{SM} = 0.225$, which is indicated with a black asterisk in Fig. 1a. The corresponding set of masks, i.e. the masks included in $\mathcal{F}_{24,0.3,0.075}$, is intended as a weighted bipartite graph, as represented in Fig. 1b. The agents involved, represented by grey circular nodes, are connected to the RI masks, represented by blue squared nodes, in which they are included, and the weight of their connection is based on the value of the t_{CI} of the masks[18]. This partition is composed by 34 masks that overall include 298 agents of the initial set A. The network is consisted of 6 components, and 54 agents are not included in any mask. The 5 masks with the highest t_{CI} (the ones with the widest edges in Fig. 1b) include agents which participated in few projects, with behavioral profiles characterized by few changes over time. The reason is that these masks are identified as highly integrated as the activity of the agents involved is almost constant. Although low levels of entropy are generated, given that the activity of the involved agents is close to minimum, they can cannot be considered as the most relevant subsets. As these 5 not conducive masks generate independent components, the ongoing analyses are focused on the remaining 29 masks, which determine the largest component of 222 agents.

After the computation of the weighted betweenness centrality, the first results suggest a modification in the rank of centrality of the nodes. Although in the real-observed network, where agents are connected together if they co-participated in projects, the centrality of agents is related to the number of projects in which they participated, in the resulting network of RI masks this does not apply. More specifically, in the largest component of the one-mode projection of the weighted bipartite graph determined by the final set of masks $\mathcal{F}_{24,0.325,0.225}$, the following elements are emerging: (i) nodes with the largest number of participations in projects appear to be close to each other in one periphery of the network; (ii)

[18] In case of agents belonging together to more than one community, the corresponding t_{CI} have been summed.

nodes with the smallest number of participations in projects appear to be close to each other in the opposite periphery of the network (with respect to the nodes with large number of projects); (iii) nodes with an average number of participations in projects appear to be very central; (iv) nodes with a high number of participations in projects and nodes with few participations in projects present few direct connections between them; (v) the shortest paths between very active nodes and scarcely active nodes (in terms of participations in projects) pass through agents with average activity.

The centrality ranking that the can be inferred after these initial results reveals an entire change with respect to the observation in the original network of participation in projects. As the RI methodology allows the investigation of the joint integration of agents' dynamics, these first insights suggest that the agents with average number of activities, that now are the most central, harmonize the very intense activity of the nodes with many participations, namely the most central in the network of projects, with the scarce activity of those agents that participated in few projects. While the structure of the observed network of participations indicates that one of the most important and recognized laws of real complex network is respected, i.e. preferential attachment, the analysis of the functionality reveals insights that suggest new interpretations. These insights will be addressed in future research. Currently, because of the tests on the 447 considered partitions, observation do not suggest contradictory indications.

8 Conclusions

As physical order is addressed as a key dimension to the comprehension of the operation and the evolution of socio-economic complex systems [3], the main aim of this research is to contribute in the development the analysis of the entropy of joint behavioral time dynamics characterized by discontinuity, e.g. interactions. The objective of this work is to facilitate the implementation of a methodology that detects functional meso-structures with information theories [11–13]. In addition, the present work attempts to facilitate the implementation of entropy-related methods in the field of social sciences, and in particular in the analyses of socio-economic dynamic complex networks. The RI algorithm is extended with the PoSH-CADDy three-step methodology so as to reduce redundancy issues. The proposed approach is implemented in a real-world dynamic network (economic agents participating to Region Tuscany Network Policies from 2000–2006) consisted of ≈350 agents, where the proportion between the number of agents and the number of instants is ≈6:1. In a complex dynamic network where the number of time instants is considerably lower than the number of involved agents, the proposed procedure accomplished to successfully detect a final set of 34 RI masks representing 34 groups of agents, whose behaviors are considered as integrated, namely not random. For the scope of this study, the focus is set in those partitions with a minimum percentage of agents included in at least one mask, and without too many overlaps among masks. The revealed ranking of the nodes' centrality appears to be substantially different from the one observed in the network of participations in projects.

In the perspective of this research are (i) the development of analytic models to statistically describe agents' characteristics in relation to the topology of the network of RI masks, (ii) the analysis of partitions obtained by combinations of the presented parameters of the methodology, (iii) the implementation of the methodology in other case studies, (iv) and the implementation of the methodology based on the edges' activation over time, instead of agents' statuses, as system variables.

References

1. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
2. Righi, R., Roli, A., Russo, M., Serra, R., Villani, M.: New paths for the application of DCI in social sciences: theoretical issues regarding an empirical analysis. In: Rossi, F., Piotto, S., Concilio, S. (eds.) WIVACE 2016. CCIS, vol. 708, pp. 42–52. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57711-1_4
3. Hidalgo, C.: Why Information Grows: The Evolution of Order, from Atoms to Economies. Basic Books, New York (2015)
4. Villani, M., Filisetti, A., Benedettini, S., Roli, A., Lane, D., Serra, R.: The detection of intermediate-level emergent structures and patterns. In: ECAL, pp. 372–378 (2013)
5. Villani, M., Benedettini, S., Roli, A., Lane, D., Poli, I., Serra, R.: Identifying emergent dynamical structures in network models. In: Bassis, S., Esposito, A., Morabito, F.C. (eds.) Recent Advances of Neural Network Models and Applications. SIST, vol. 26, pp. 3–13. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-04129-2_1
6. Filisetti, A., Villani, M., Roli, A., Fiorucci, M., Serra, R.: Exploring the organisation of complex systems through the dynamical interactions among their relevant subsets. In: Proceedings of the European Conference on Artificial Life 2015 (ECAL 2015), vol. 13, pp. 286–293 (2016)
7. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3), 75–174 (2010)
8. Fortunato, S., Hric, D.: Community detection in networks: a user guide. Phys. Rep. **659**, 1–44 (2016)
9. Sani, L., et al.: Efficient search of relevant structures in complex systems. In: Adorni, G., Cagnoni, S., Gori, M., Maratea, M. (eds.) AI*IA 2016. LNCS (LNAI), vol. 10037, pp. 35–48. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49130-1_4
10. Vicari, E., et al.: GPU-based parallel search of relevant variable sets in complex systems. In: Rossi, F., Piotto, S., Concilio, S. (eds.) WIVACE 2016. CCIS, vol. 708, pp. 14–25. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57711-1_2
11. Tononi, G., McIntosh, A.R., Russell, D.P., Edelman, G.M.: Functional clustering: identifying strongly interactive brain regions in neuroimaging data. NeuroImage **7**(2), 133–149 (1998). https://doi.org/10.1006/nimg.1997.0313
12. Tononi, G., Sporns, O., Edelman, G.M.: A measure for brain complexity: relating functional segregation and integration in the nervous system. Proc. Natl. Acad. Sci. U.S.A. **91**(11), 5033–5037 (1994). ISSN: 0027-8424
13. Tononi, G., Sporns, O., Edelman, G.M.: A complexity measure for selective matching of signals by the brain. Proc. Natl. Acad. Sci. U.S.A. **93**(8), 3422–3427 (1996). ISSN: 0027-8424

14. Russo, M., Rossi, F.: Cooperation networks and innovation: a complex system perspective to the analysis and evaluation of a EU regional innovation policy programme. Evaluation **15**, 75–100 (2009). https://doi.org/10.1177/1356389008097872
15. Caloffi, A., Rossi, F., Russo, M.: The emergence of intermediary organizations: a network-based approach to the design of innovation policies. In: Handbook on Complexity and Public Policy, pp. 314–331 (2015). ISBN: 978-1-78254-951-2
16. Rossi, F., Caloffi, A., Russo, M.: Networked by design: can policy requirements influence organisations' networking behaviour? Technol. Forecast. Soc. Chang. **105**, 203–214 (2016). https://doi.org/10.1016/j.techfore.2016.01.004
17. Lane, D.A.: Complexity and innovation dynamics. In: Handbook on the Economic Complexity of Technological Change. Edward Elgar Publishing (2011). ISBN: 978-0-85793-037-8