Chapter 12
Beyond Networks: Search for Relevant Subsets in Complex Systems

Andrea Roli, Marco Villani, Alessandro Filisetti, and Roberto Serra

1 Introduction

Most natural and artificial systems show some form of organization that can be modelled as a network of nodes. A prominent example of the success of this approach is the field known as complex networks science, which has achieved outstanding results in the analysis, design and control of complex network systems. Research in complex networks has so far mainly focused on the topological properties of the network that affect or explain the behavior observed. Nevertheless, in most interesting cases, the topology of the network might not provide sufficient information on system dynamics. For example, consider network models in which nodes are subject to nonlinear update functions: in this case, the actual relations among variables might not be captured by the topology, but they are rather well represented in terms of coordinated dynamical behavior of groups of nodes. Notable examples are Boolean networks [12], coupled map lattices [3] and functional connectivity graphs in neuroscience [9, 10]. Moreover, in several cases, the interactions

A. Roli (✉)
Department of Engineering and Computer Science (DISI), Alma Mater Studiorum University of Bologna, Via Venezia 52, I-47521 Cesena, Italy

European Centre for Living Technology, Ca Minich, S. Marco 2940, 30124 Venezia, Italy
e-mail: andrea.roli@unibo.it

M. Villani · R. Serra
Department of Physics, Informatics and Mathematics, University of Modena e Reggio Emilia, Via Campi 213b, 41125 Modena, Italy

European Centre for Living Technology, Ca Minich, S. Marco 2940, 30124 Venezia, Italy
e-mail: marco.villani@unimore.it; rserra@unimore.it

A. Filisetti
Current address: Explora s.r.l, Rome, Italy

European Centre for Living Technology, Ca Minich, S. Marco 2940, 30124 Venezia, Italy
e-mail: alessandro.filisetti@gmail.com

G. Minati et al. (eds.), *Towards a Post-Bertalanffy Systemics*, Contemporary Systems Thinking, DOI 10.1007/978-3-319-24391-7_12

among the interesting variables are largely, or at least partly, unknown; therefore, it is necessary to infer some hypotheses about the organization by observing system's behavior "from the outside". For these reasons, methods are required to go beyond network models, taking into account multiple and dynamical relations among nodes.

In this paper we address the issue of identifying sets of variables that are good candidates as "relevant subsets" (RSs) for describing the organization of a dynamical system. We will assume that some variables can be observed and that they change in time (possibly reaching an attractor state) and we will look for groups of variables that can represent the required RSs for describing the system organization. Note that the very notion of a RS is somewhat ill-defined. However this feature is shared by several other interesting concepts, like e.g. that of a cluster. In general, we stress that a good candidate RS (CRS for short) should provide important indications about some key features of the system organization. The main features of good CRSs can be tentatively identified as follows:

1. The variables that belong to a RS have a strong interaction and integration with the other variables of the same RS
2. They have a weaker integration with the other system variables or RSs

The outcome of the analysis we propose here is essentially a list of subsets, ranked according to the above criteria, that provide clues to understand the system organization. The list cannot be used *tout court*: its application requires some care, but it can lead one to discover very interesting non obvious relationships, above and beyond those provided by topological or structural information. Note also that RSs can overlap, i.e., they don't need to be disjoint sets.

A peculiar feature of the method that will be illustrated in the following section is that it does not require knowledge of the system structure and rules, as it only requires knowledge of the behavior in time of the variables $x_1 \ldots x_N$, without any prior knowledge of the topology or of the dynamical rules. All that is needed is a set of numerical values of the relevant variables in time, therefore the method can be applied to models as well as to real-world data. Of course, if further pieces of information are available (e.g., concerning the topology), they can be profitably used.

The reason why our method does not require prior knowledge on system topology and rules is that it infers information on structure and dynamical relations among variables by exploiting techniques from information theory. While the idea of applying concepts and measures of information theory to the study of complex systems is certainly not new, we found that a measure of this kind, the Cluster Index—introduced by Tononi and Edelman [11] in the study of neural networks close to a stationary state—can be profitably generalized to study dynamical systems. This generalization is called here the Dynamical Cluster Index (DCI) and it has been introduced in [13]. This method has been successfully applied in uncovering the structure of artificial ad hoc systems (e.g., simple leader-followers models), of models of autocatalytic reaction sets, of genetic networks (e.g., the Arabidopsis Thaliana), of protein networks (e.g., the Mapk signalling system[1]) and robot controllers [5].

[1] Results presented in a contribution currently under review.

We present and discuss the DCI-based method in Sect. 2, while in Sect. 3 we illustrate the application of the CRSs search procedure to typical examples of Boolean networks. We conclude the contribution with Sect. 4.

2 The Dynamical Cluster Index

In this section we succinctly introduce the DCI. The interested reader can find more details in [2, 13].

Let us consider a system modelled with a set U of N variables assuming finite and discrete values. The cluster index of a subset S of variables in U, $S \subset U$, as defined by Tononi [11], estimates the ratio between the amount of information integration among the variables in S and the amount of integration between S and U. These quantities are based on the Shannon entropy of both the single elements and sets of elements in U.

According to information theory, the entropy of an element x_i is defined as:

$$H(x_i) = -\sum_{v \in V_i} p(v) \;\; log\, p(v) \tag{12.1}$$

where V_i is the set of the possible values of x_i and $p(v)$ the probability of occurrence of symbol v. The entropy of a pair of elements x_i and x_j is defined by means of their joint probabilities:

$$H(x_i, x_j) = -\sum_{v \in V_i} \sum_{w \in V_j} p(v,w) \;\; log\, p(v,w) \tag{12.2}$$

Equation (12.2) can be extended to sets of k elements considering the probability of occurrence of vectors of k values. In this work, we deal with observational data, therefore probabilities are estimated by means of relative frequencies.

The cluster index $C(S)$ of a set S of k elements is defined as the ratio between the integration $I(S)$ of S and the mutual information between S and the rest of the system $U - S$.

The integration of S is defined as:

$$I(S) = \sum_{x \in S} H(x) - H(S) \tag{12.3}$$

$I(S)$ represents the deviation from statistical independence of the k elements in S. The mutual information $M(S; U - S)$ is defined as:

$$M(S; U - S) \equiv H(S) - H(S|U - S) = H(S) + H(U - S) - H(S, U - S) \tag{12.4}$$

where $H(A|B)$ is the conditional entropy and $H(A,B)$ the joint entropy. Finally, the cluster index $C(S)$ is defined as:

$$C(S) = \frac{I(S)}{M(S;U-S)} \tag{12.5}$$

Since C is defined as a ratio, it is undefined in all those cases where $M(S;U-S)$ vanishes. In this case, the subset S is statistically independent from the rest of the system and it has to be analyzed separately. As $C(S)$ scales with the size of S, cluster index values of systems of different size need to be normalized. To this aim, a reference system is defined, i.e., the homogeneous system U_h, randomly generated according to the probability of each single state measured in the original system U. Then, for each subsystem size of U_h the average integration "I_h" and the average mutual information "M_h" are computed. Finally, the cluster index value of S is normalized by means of the appropriate normalization constant:

$$C'(S) = \frac{I(S)}{\langle I_h \rangle} / \frac{M(S;U-S)}{\langle M_h \rangle} \tag{12.6}$$

Furthermore, to assess the significance of the differences observed in the cluster index values, a statistical index T_c is computed:

$$T_c(S) = \frac{C'(S) - \langle C'_h \rangle}{\sigma(C'_h)} \tag{12.7}$$

where $\langle C'_h \rangle$ and $\sigma(C'_h)$ are the average and the standard deviation of the population of normalized cluster indices with the same size of S from the homogeneous system.

We emphasize that definitions (12.5)–(12.7) are made without any reference to a particular type of system. In their original papers, Edelman and Tononi considered fluctuations around a stationary state of a neural system. In our approach, this measure is applied to time series of data generated by a dynamical model. In general, these data lack the stationary properties of fluctuations around a fixed point. Moreover, depending upon the case at hand, either transients from arbitrary initial states to a final attractor, or collections of attractor states can be considered, as well as responses to perturbations of attractor states. In all these cases we will use Eq. (12.5), that will therefore be called the Dynamical cluster index (DCI).

The search for CRSs of a dynamical system by means of the DCI requires first the collection of observations of the values of the variables at different instants. It is not necessary to know the values of all the important variables, although of course the quality of the results can be negatively affected by lack of information. Moreover, in principle it is not required to have a time series: in fact, since the DCI is computed on the basis of symbol frequencies, a collection of snapshots of the system variables is sufficient.

In order to find CRSs, in principle all the possible subsets of system variables should be considered and their DCI computed. In practice, this procedure is feasible only for small-size subsystems in a reasonable amount of time. Therefore, heuristics are required to address the study of large-size systems. A simple heuristic consists in evaluating samples of subsets of increasing cardinality k: at first, subsets of size k are uniformly sampled and evaluated, then the samples of size $k+1$ are composed of random samples plus all the $(k+1)$-size neighbours of the k-size subset with the highest DCI value. This heuristic has proven to be quite effective, because usually the subsets with highest DCI value are composed of subsets which in turn have a high DCI value, compared to the subsets of the same cardinality. In our experiments, we always relied on the procedure described above. However, other search procedures can be adopted.[2] Once all the samples for size up to $N-1$ are evaluated, the T_c is computed so as to rank the CRSs.

3 Relevant Subsets in Boolean Networks

In this section we illustrate the application of the CRSs search procedure on networks whose nodes are updated according to Boolean functions, namely Boolean networks (BNs). BNs are an important framework frequently used to model genetic regulatory networks [4], also applied to relevant biological data [1, 7, 8] and processes [6, 12]. In the BNs we consider in the following, the nodes update dynamics is synchronous and deterministic, therefore every system state has only one successor state and the system trajectories in the state space can be decomposed into a transient and a cyclic attractor (which can also be a fixed point, i.e., a cyclic attractor of period 1).

The first example consists of a BN, named BN1, made of two independent components (see Fig. 12.1a).[3] The data used in order to compute the relative frequencies are obtained from the states of the various attractors, each one weighted according to its basin of attraction. The rationale for using the attractor states for this analysis is, intuitively, that attractors should be able to capture the important functional relationships in a system. The DCI analysis is able to correctly identify the two separated subnetworks of BN1 (data not shown for lack of space). This case simply provides an example of the concordance of the results provided by the DCI-based method and a topological clustering technique.

Example involving BN2 (Fig. 12.1b) is of great interest: as it can be observed, BN2 is composed of two interacting subnets: one composed of nodes $\{4,5,6\}$ (identical to a BN module in BN1) and nodes $\{1,2\}$, which constitute an oscillator. The two subnets feed node 3, which make a XOR of nodes 1 and 6. If we analyzed BN2 in terms of its topology, we would split it into the three above-mentioned components. This analysis, though, does not capture the dynamical essence of the network, which instead involves all but one node (either node 1 or node 2 is excluded, as they are totally correlated), because they are all equally important to reconstruct the dynamical behavior of BN2. The DCI-based analysis returns one main CRS composed of nodes $\{2,3,4,5,6\}$, so it correctly clusters the relevant nodes for the

[2] As an ongoing work, we are experimenting with a genetic algorithm for searching the CRS with highest DCI value for each size.

[3] The following three examples has already been described in [13], however in this contribution we emphasize different aspects w.r.t. previous work.

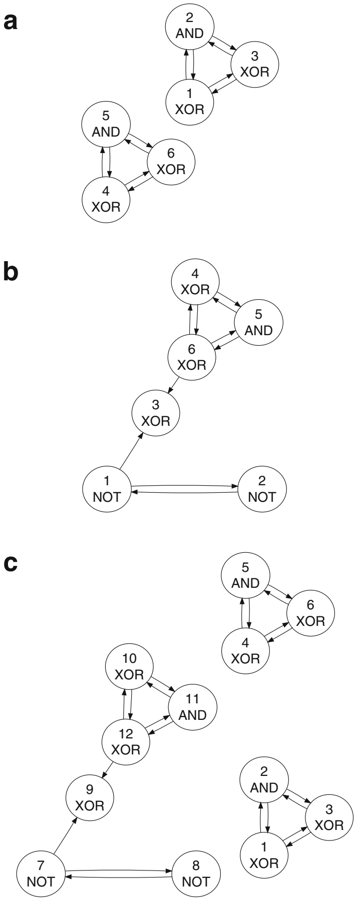
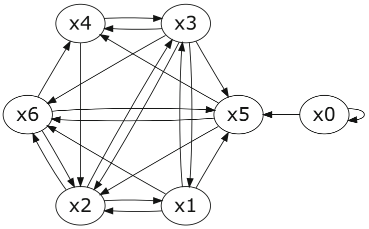


Fig. 12.1: (**a**) Independent Boolean networks (BN1); (**b**) interdependent networks (BN2); (**c**) a system composed of the union of both previous networks (BN3). The boolean function associated to each node is shown inside the node itself

dynamic of the system. Finally, the union of BN1 and BN2, named BN3, is depicted in Fig. 12.1c. An analysis based on the DCI correctly returns in the highest positions of the rank the two subnets of BN1 and the CRS of BN2.

One might argue that these examples are trivial, because they consider very small and simple BNs: on the contrary, while these BNs are simple if topology and rules are known, it is far from trivial to identify the most CRSs by the sole observation of the stationary states of the BNs.

In the following we present further notable cases of networks in which the topological structure is not sufficiently informative to detect the actual RSs. They



```
Node functions:

x0 = !x0
x1 = (x2 & !x3) | (!x2 & x3)
x2 = (x1 & !x3) | (!x5 & x3) |
     (x3 & x4) | (!x5 & !x6)
x3 = (x1 & !x2) | (!x4 & x2)
x4 = (x5 & !x3) | (!x6 & x3)
x5 = (x6 & !x1) | (!x1 & x3) & x0
x6 = (x1 & x3) | (!x2 & x5)
```

Fig. 12.2: BN with a topology composed of a cluster and a peripheral node, which indeed strongly influences the rest of the network

represent the not so rare case in which topology even defeats dynamical community identification. The first BN under exam is depicted in Fig. 12.2, along with node update functions. The network has 5 attractors: two fixed points, 2 cycles of period 2 and one of period 14. A topological analysis would identify two clusters, namely $\{x_0\}$ and $\{x_1,\ldots,x_6\}$. Nevertheless, a dynamical analysis made by means of the DCI returns as first CRS the subset $\{x_0,x_1,x_2,x_4,x_5,x_6\}$, showing that from the dynamical point of view node $x_0$is not disconnected from the rest of the network. This is a prominent example of the situation in which an agent external to a (topological) community can anyway influence it.

An example related to defeating community detection techniques based on pairwise relations is provided by the BN depicted in Fig. 12.3. The BN has 3 attractors of period 4, 5 and 13, respectively. Despite the typical structure displaying two loosely connected clusters, the dynamics of the network is such that nodes x_1 and x_6 are indeed strongly interdependent and the DCI-based analysis shows that the most CRSs involve nodes from both the topological clusters. For example, among the first five suggested CRSs, the largest one is composed of 8 nodes, all except nodes x_2 and x_4.

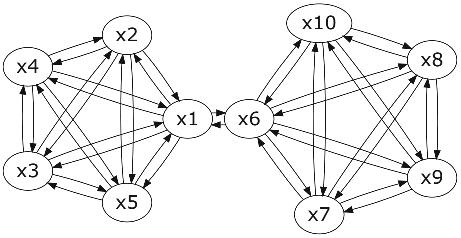


Fig. 12.3: BN whose topology is composed of two cliques (node update functions are omitted for lack of space). The actual dynamics, though, is such that the nodes of the two clusters are strongly integrated

4 Conclusion

In this contribution we have presented a method, based on information theory, which makes it possible to detect CRSs of variables in a system, exploiting the data deriving from the dynamics of the system. Although the advantage of the method is to identify CRSs emerging from the dynamics of the system and it does not require prior knowledge on system's structure, a promising direction for future work consists in combining the topological information with the dynamical one.

Acknowledgements The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n^{o} 284625.

References

1. Balleza, E., Alvarez-Buylla, E., Chaos, A., Kauffman, S., Shmulevich, I., & Aldana, M. (2008). Critical dynamics in genetic regulatory networks: Examples from four kingdoms. *PloS One, 3*(6), e2456.
2. Filisetti, A., Villani, M., Roli, A., Fiorucci, M., Poli, I., & Serra, R. (2014). On some properties of information theoretical measures for the study of complex systems. In: C. Pizzuti and G. Spezzano (Eds.), *Advances in artificial life and evolutionary computation*, CCIS 445, pp. 140–150, Springer.
3. Kaneko, K. (1990). Clustering, coding, switching, hierarchical ordering, and control in a network of chaotic elements. *Physica D, 41*, 137–172.
4. Kauffman, S. (1993). *The origins of order: Self-organization and selection in evolution*. Oxford: Oxford University Press.
5. Roli, A., Villani, M., Serra, R., Garattoni, L., Pinciroli, C., & Birattari, M. (2013). Identification of dynamical structures in artificial brains: An analysis of boolean network controlled robots. In *AI*IA2013: Advances in Artificial Intelligence. Lecture notes in computer science* (Vol. 8249, pp. 324–335). New York: Springer.
6. Serra, R., Villani, M., Barbieri, A., Kauffman, S., & Colacci, A. (2010). On the dynamics of random Boolean networks subject to noise: Attractors, ergodic sets and cell types. *Journal of Theoretical Biology, 265*(2), 185–193.
7. Serra, R., Villani, M., Graudenzi, A., & Kauffman, S. (2007). Why a simple model of genetic regulatory networks describes the distribution of avalanches in gene expression data. *Journal of Theoretical Biology, 246*, 449–460.
8. Serra, R., Villani, M., & Semeria, A. (2004). Genetic network models and statistical properties of gene expression data in knock-out experiments. *Journal of Theoretical Biology, 227*, 149–157.
9. Shalizi, C., Camperi, M., & Klinkner, K. (2006). Discovering functional communities in dynamical networks. In *Proceedings of ICML 2006. Lecture notes in computer science* (Vol. 4503, pp. 140–157). New York: Springer.
10. Sporns, O., Tononi, G., & Edelman, G. (2000). Theoretical neuroanatomy: Relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cerebral Cortex, 10*(2), 127–141.
11. Tononi, G., McIntosh, A., Russel, D., & Edelman, G. (1998). Functional clustering: Identifying strongly interactive brain regions in neuroimaging data. *Neuroimage, 7*, 133–149.
12. Villani, M., Barbieri, A., & Serra, R. (2011). A dynamical model of genetic networks for cell differentiation. *PloS One, 6*(3), e17703.
13. Villani, M., Filisetti, A., Benedettini, S., Roli, A., Lane, D., & Serra, R. (2013). The detection of intermediate-level emergent structures and patterns. In *Advances in Artificial Life, ECAL 2013* (pp. 372–378). New York: The MIT Press.