# An Improved Relevance Index Method to Search Important Structures in Complex Systems

Laura Sani[1,0000−0002−3288−5945], Alberto Bononi[1,0000−0002−9524−3659], Riccardo Pecori[1,4,0000−0002−5948−5845], Michele Amoretti[1,0000−0002−6046−1904], Monica Mordonini[1,0000−0002−5916−9770], Andrea Roli[2,0000−0001−9891−5441], Marco Villani[3,0000−0002−5991−5470], Stefano Cagnoni[1,0000−0003−4669−512X], and Roberto Serra[3,0000−0002−1417−5106]

[1]Dip. di Ingegneria e Architettura, Università di Parma, Italia
[2]Dip. di Informatica - Scienza e Ingegneria
Università di Bologna - Sede di Cesena, Italia
[3]Dip. di Scienze Fisiche, Informatiche e Matematiche
Università degli Studi di Modena e Reggio Emilia, Italia
[4]SMARTEST Research Centre, Università eCAMPUS, Novedrate (CO), Italia

**Abstract.** We present an improvement of a method that aims at detecting important dynamical structures in complex systems, by identifying subsets of elements that show tight and coordinated interactions among themselves, while interplaying much more loosely with the rest of the system. Such subsets are estimated by means of a Relevance Index (RI), which is normalized with respect to a homogeneous system, usually described by independent Gaussian variables, as a reference. The strategy presented herein improves the way the homogeneous system is conceived from a theoretical viewpoint. Firstly, we consider the system components as dependent and with equal pairwise correlations, which implies a non-diagonal correlation matrix of the homogeneous system. Then, we generate the components of the homogeneous system according to a multivariate Bernoulli distribution, by exploiting the NORTA method, which is able to create samples of a desired random vector, given its marginal distributions and its correlation matrix. The proposed improvement on the RI method has been applied to three different case studies, obtaining better results compared with the traditional method based on the homogeneous system with independent Gaussian variables.

**Keywords:** Complex Systems Analysis, Information Theory, Relevance Index, NORTA

## 1 Introduction

The identification of functional structures in dynamical systems composed of many interacting parts is a major challenge in science. In particular, the formation of intermediate-level structures is of particular interest for what concerns biological as well as artificial systems. These structures come from the dynamics of small-scale processes, but possess peculiar characteristics and are able to deeply influence the system they belong to.

Several measures have been proposed to describe the organization of these dynamical complex systems, many of which are based on information theory [10, 19]. Some of the most relevant results of the application of such metrics can be found in the domain of neuroscience [25, 27].

Starting from these results, Villani et al. [31] introduced a method to identify relevant structures in dynamical complex systems, based on a dataset including samples of the system status at different times. In particular, the Relevance Index (RI) quantifies how much the behavior of these relevant structures deviates from the behavior of a reference (homogeneous) system, in which the variables have, individually, the same marginal distributions as in the dataset, and all have the same pairwise correlation. In particular, a system characterized by independent Gaussian variables, i.e., with zero pairwise correlation, was originally taken as a reference [31].

In previous works, we improved the aforementioned RI method by applying some metaheuristics, in order to deal with the curse of dimensionality in computing the index [22, 26], and a GPU-based parallelization scheme, in order to speed up the overall computation [30].

In this paper we propose a further improvement to the RI method, by imposing that the variables in the homogeneous system all have the same nonzero pairwise correlation, matching the average pairwise correlation estimated from the system under analysis. This is achieved by using the NORTA method [6]. The introduction of this pairwise correlation value has allowed us to identify particularly interesting groups of variables, undetected in previous experiments.

The rest of the paper is structured as follows: in Section 2 we summarize some previous applications of the relevance index and of the NORTA method; in Section 3 we describe the most significant theoretical steps underlying the RI computation and the NORTA method; in Section 4 we assess the improvements obtained by applying the proposed modification to some relevant use cases; finally, in Section 5, we draw some conclusions.

## 2 Background

In this section we summarize previous works that take advantage either of the RI method or of the NORTA method, which we are going to combine in the proposed technique.

### 2.1 The Relevance Index Method

Much research has been already focused on the search for particularly informative groups in dynamical complex systems. Because of the emphasis on the (nonlinear) relationships among their constituents, many efforts have been based on the analysis of their representation through either networks [14] (for example community detection [4, 16]), multigraphs [1, 13] or hypergraphs [11]. Nevertheless, often the interactions across these informative groups are not known; in addition, the interaction topology could turn out not to be sufficient by itself

to determine the behavior of the whole system, since it is often necessary to consider also the dynamical movements of the constituents. Besides this, it has been shown that relevant information about emergent structures in dynamical systems can be extracted by observing the system behavior "from the outside", by means of information-theoretical and statistical techniques [2, 3, 18], sometimes combined with dynamical systems analyses [8]. Some previous works have documented the use of these information-theoretical measures for studying complexity [10, 19] and criticality [5, 20, 34, 37]. However, none of the existing methods has all the following desirable properties:

- ability to identify groups of variables that change in a coordinated fashion;
- ability to identify critical states;
- direct applicability to data, without any need to resort to models;
- robustness with respect to sampling effort and system size.

The RI method, which is based on Shannon's entropy, appears to be a step towards obtaining all the aforementioned requirements. Indeed, the RI is a method based on the Cluster Index (CI), introduced by Edelman and Tononi in 1994 and 1998 [28, 29], which detects functional groups of brain regions, assuming system fluctuations around a steady state. The RI method extends the applicability of the CI to a broad range of non-stationary dynamical systems, such as abstract models of gene regulatory networks and simulated chemical [31], biological [32], as well as social [9, 23] systems. Moreover, the experimental analysis concerning two prominent models that exhibit two different kinds of criticality, namely the Ising model for phase transition and the Random Boolean Network (RBN) for dynamical criticality, demonstrated that the RI can be effectively used to identify critical states [21].

### 2.2 The NORTA Method

NORTA ("NORmal To Anything") is a method devised to generate specifically correlated random vectors [6]. This is a mathematical procedure that solves the issue of creating random vectors of correlated samples, given the set of their marginal distributions (marginals) and a measure of the dependence among them.

This is a good choice in our scenario, since, usually, the majority of complex systems components experience a certain degree of mutual dependence [35]. Some recent examples, where NORTA has been successfully employed in different fields, include wind power generation in renewable power supply systems [17], and the modeling of probabilistic load flows, based on Latin Hypercube Sampling [36]. Indeed, NORTA presents some degrees of uncertainty in the estimation of the marginal distributions and of the correlation matrix [35], since it is not always guaranteed that its samples have exactly the desired correlation matrix. However, we found it useful to overcome some issues encountered in applying the original RI method to some simple systems described by a moderate number of variables.

## 3 Theoretical Approach

The RI can be used to study data from a wide range of dynamical system classes, with the purpose of identifying sets of variables that behave in a somehow coordinated way, i.e., the variables belonging to the set are integrated with each other much more than with the other variables not pertaining to the set itself. These subsets can be used to describe the whole system organization, thus they are named Relevant Subsets (RSs).
The computation of the RI, which is an information-theoretical measure based on Shannon's Entropy (H in the following) [7], is usually based on observational data, and probabilities are estimated as the relative frequencies of the values observed for each variable. The theoretical definition of the RI is summarized in the following.

Let us consider a system $U$ composed of $n$ random variables $X_1, X_2, ..., X_n$ (e.g., agents, chemicals, genes, artificial entities) and suppose that $S_k$ is a subset composed of $k$ elements, with $k < n$. The $RI$ of $S_k$ is defined as:

$$RI(S_k) \ = \frac{I(S_k)}{MI(S_k; U \backslash S_k)} \tag{1}$$

where $I(S_k)$ is the integration, which measures the mutual dependence among the $k$ elements in $S_k$, and $MI(S_k; U \backslash S_k)$ is the mutual information, which quantifies the mutual dependence between subset $S_k$ and the remaining part of the system $U \backslash S_k$.

The integration, in turn, is defined as:

$$I(S_k) \ = \sum_{s \in S_k} H(s) - H(S_k) \tag{2}$$

while the mutual information is formalized as follows:

$$MI(S_k; U \backslash S_k) \ = H(S_k) + H(U \backslash S_k) - H(S_k, U \backslash S_k) \tag{3}$$

The integration can be shown to be the Kullback-Leibler Distance [7] between the joint distribution of the system variables and the product distribution of their marginals. Hence the integration is zero whenever the system variables are independent.

Trivially, the RI is undefined if $MI(S_k; U \backslash S_k) = 0$. However, a vanishing $MI$ is a sign of independence (i.e., physical separation) of the subset under exam from the rest of the system, and therefore the subset has to be studied separately.

Since the RI increases with the subset size, a normalization method is required to compare RI values of subsets of different sizes. Moreover, the statistical significance of RI differences should be assessed by means of an appropriate test. For these reasons, a statistical significance index was introduced as [28]:

$$T_c(S_k) \ = \frac{\nu RI(S_k) - \nu \left\langle RI_h \right\rangle}{\nu \sigma(RI_h)} \ = \frac{RI(S_k) - \left\langle RI_h \right\rangle}{\sigma(RI_h)} \tag{4}$$

where $\langle RI_h \rangle$ and $\sigma(RI_h)$ are, respectively, the average and the standard deviation of the RI of a sample of subsets of size $k$ extracted from a reference homogeneous system $U_h$, and $\nu = \langle MI_h \rangle / \langle I_h \rangle$ is its normalization constant.

A post-processing sieving algorithm [33] is used to select the most relevant sets, reducing the list of Candidate Relevant Sets (CRSs) to the most representative ones, i.e., those having the highest $T_c$ values. The sieving algorithm is based on the criterion by which, if a CRS is a proper subset of another CRS and ranks higher than this, then it should be considered more relevant than this. Therefore, the algorithm keeps only those CRSs that are not included in or do not include any other CRS with higher $T_c$: this "sieving" action stops when no more eliminations are possible and the remaining groups of variables are the elementary RSs.

The generation of the homogeneous system is critical, as stated also in [31], and often, in the past, a simple but general and easy to compute solution was preferred. This solution encompassed the computation of the frequency of occurrence of each variable, given the available observations, and the generation of a new random series of samples, where each variable had a prior probability equal to the frequency of the original observations. The homogeneity required by Tononi was achieved by considering the components of the random vector $U_h$ to be Gaussian and independent. This caused:

1. the correlation matrix of the homogeneous system to be a diagonal matrix, i.e., with pairwise correlations set to zero;
2. the integration $I(S_k)$ to be zero for all subsets of the homogeneous system.

The improved version of the method we propose in this paper consists in taking the pairwise correlation between variables describing the homogeneous system into account, requiring that it be not null, which seems a more realistic assumption. In this way, we remove a hypothesis (the independence of the variables) that is not true in general. Moreover, we maintain the homogeneity required by Tononi, by forcing all off-diagonal elements of the correlation matrix to have the same constant value $\rho$:

$$CORR(U_h) = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \dots & \rho & 1 \end{bmatrix}$$

while we normalize all variances to 1. The value of $\rho$ is computed, in a first approximation, as the average value of all pairwise correlations of the observed variables.

In order to generate a homogeneous system with the aforementioned features, we take advantage of the NORTA method [6]. The measure of dependence we used in NORTA is the usual *product-moment* correlation matrix, based on the linear Pearson correlation coefficient with entries defined according the following

formula:

$$\rho(X_i, X_j) = \frac{COV(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}. \quad (5)$$

The NORTA method creates independent and identically distributed replicas of a random vector $\mathbf{X} = (X_1, X_2, ..., X_n)$, based on its (known) marginal distributions $F_i(x) = P(X_i \leq x), \quad i = 1, \ldots, n$ and the correlation matrix $CORR(\mathbf{X})$.
In summary, the NORTA procedure performs the following steps:

1. generates a normal random vector $\mathbf{Z} = (Z_1, Z_2, ...Z_n)$ with zero mean and covariance matrix $COV(\mathbf{Z})$, with 1s on the main diagonal;
2. obtains the prescribed marginal distributions by computing the replica $\mathbf{X}' = (X'_1, X'_2, ..., X'_n)$ according to the following equation:

$$X'_i = F_i^{-1}(\Phi(Z_i)) \;\; i = 1...n, \quad (6)$$

where $\Phi$ is the distribution function of a standard Gaussian random variable and $F_i^{-1}$ is the inverse of $F_i$,defined as:

$$F_i^{-1}(u) = \inf\{x : F_i(x) \geq u\}. \quad (7)$$

3. chooses $COV(\mathbf{Z})$ in order to induce the requested correlation matrix $CORR(\mathbf{X})$. In this case there is no closed-form solution and the method often relies on an efficient numerical search, by solving a number of one-dimensional root-finding problems. In some cases the procedure does not lead to the exact desired correlation matrix, failing to produce a positive semidefinite matrix, which is a requirement for a valid correlation matrix. However, NORTA can often get very close to the desired correlation matrix, even in very high dimensions.

In this work, NORTA is used to generate the homogeneous system based on the R implementation known as NORTARA[1]. This package generates $n$-dimensional random vectors with given marginal distributions and correlation matrix. The NORTA algorithm, which generates a standard normal random vector and then transforms it into a random vector with specified marginal distributions, is combined with the RA (Retrospective Approximation) algorithm, which is a generic stochastic root-finding algorithm.

## 4 Experimental Evaluation

In order to test the presented methodology, we analyzed three different systems whose dynamics are precisely known. In particular, we studied the consequences of using different homogeneous systems: the one produced with the method proposed in this work (where we "inject" the average correlation which characterizes the system under study - $H_{wiC}$, where $wiC$ means "with correlation") and the

[1] http://cran.r-project.org/web/packages/NORTARA/

one produced with the original method ($H_{noC}$). In the following, we focus on the application of the RI analysis, possibly applying the sieving algorithm in order to simplify the results. The binary nature of the variables of the test systems we used allows one to apply the $H_{wiC}$ approach with simple Bernoulli distributions to all situations.

The three case studies we considered are representative of valuable research fields, that is, (i) the dynamics of Boolean networks, (ii) dynamical simulations of autocatalytic reaction systems happening within a Continuous-flow Stirred-Tank Reactor (CSTR for short) and (iii) simplified models of the dynamics of opinion diffusion.

The **Boolean network framework**, despite its apparent simplicity, has obtained remarkable results in simulating several aspects of real gene regulatory networks [12, 24]. In particular, here we present a collection of 5 different Boolean systems (denoted as RBN1, ..., RBN5) composed of 12 nodes, synchronously updated on the basis of either a Boolean function or a random Boolean value generator.
In each analysis considered in this paper, instead of juxtaposing different states belonging to the different attractors of each system [31], we follow single trajectories, perturbed every 20 steps by temporarily changing a randomly chosen variable from 0 to 1 (or vice versa).

The **CSTR** case study simulates a collection of molecules able to collectively self-replicate [12], a situation frequently studied in researches about the origin of life [15]; very similar assemblies could play an important role also in future bio-technological
In this research, we tested a simple system featuring two distinct reaction pathways, a Linear reactions CHain (LCH) and an AutoCatalytic set of molecular Species (ACS). The reactions occur only in the presence of a specific catalyst, since spontaneous reactions are assumed to occur too slowly to affect the system behavior. Both LCH and ACS pathways occur in an open CSTR with a constant influx of feed molecules and a continuous outgoing flux of all the molecular species proportional to their concentration (see [31] for a more detailed description of the model). The problem we address in this paper is the detection of the groups of chemicals that participate in distinct dynamical organizations, by simply observing their concentration in time.
The asymptotic behavior of this kind of systems is a single fixed point [31], due to the system feedback structure. In order to apply our analysis, we need to observe the feedbacks in action; so, we perturb the concentration of some molecules in order to trigger a response in the concentration of (some) other species. We temporarily set to zero the concentration of some species after the system has reached its stationary state. In order to analyze the system response to perturbations we discretize its trajectory by observing it within equally-sized, non-overlapping time windows and by classifying the behavior of the chemical concentrations within this interval as "chemical concentration changing" ("1" tag) and "no change in chemical concentration" ("0" tag).

Finally, we compare the results of the application of the RI to different homogeneous systems on a simple model, in which the integration among variables in a subsystem under observation and its mutual information with the remaining part of the system can be tuned by acting on few parameters. The model abstracts from specific functional relationships among elements of the system and could resemble a basic **Leader-Followers model** (LF), used in opinion dynamics studies.
The system is composed of a vector of $n$ binary variables $\{X_1, X_2, ..., X_n\}$ representing, for example, the opinion in favor of or against a given proposal. The model generates independent observations of the system state, i.e., each observation is a binary $n$-vector generated independently of the others, based on the following rules:

- Variables are divided into three groups, G1 = $\{L_a, F1_a, F2_a, F3_a\}$, G2 = $\{L_b, F1_b, F2_b\}$, and G3 = $\{L_c, F1_c, ... , F8_c\}$.
- $L_a$, $L_b$, $L_c$ are the leaders of their groups[2], and they have a probability $p_{lcpy}$ to copy the value of another leader, and a probability of 1-$p_{lcpy}$ to independently assume a random value in {0,1} (with probability of obtaining a "1" equal to 0.4, 0.3, and 0.3, respectively).
- The values of the followers of the three groups are set as a copy (or negation) of their leaders with probability $p_{copy}$ and randomly (according to a Bernoulli distribution with probability 0.5) otherwise.
- The three groups are submerged into a "sea" of random variables following a Bernoulli distribution with $P(x = 0) = P(x = 1) = 0.5$.

It is possible to tune the integration among elements within groups and the mutual information between groups by changing $p_{lcpy}$ or $p_{copy}$. In our examples, we fixed for simplicity $p_{lcpy}$=0.0 (non-interacting groups) with $p_{copy}$=1.00 (perfect followers) and $p_{copy}$=0.98 (imperfect followers).

### 4.1 Results

In Figure 1, we report the relevant subsets identified by the RI analysis performed on RBNs using the $H_{noC}$ or $H_{wiC}$ homogeneous systems as a reference for the $T_c$ computation. The RBN systems are relatively simple, and the most interesting relevant subsets are evident also without applying the sieving algorithm (see Table 1).

Figure 1 reports the two groups that rank highest according to the $T_c$ value (the first four ranks for case RBN5).
Both approaches find the same solutions in cases RBN1, RBN2 and RBN3. In particular, the two methods directly identify the two correct solutions of RBN1, the two fundamental groups composing the correct solution of RBN2, and the correct solution of case RBN3. In RBN2, the simple iteration of the RI method after the application of the sieving algorithm is able to identify the

[2] In details, $L_b(t) = L_a(t-1)$ and $L_c(t) = L_b(t-1)$.

**Table 1.** Table showing the relationships among nodes of the considered RBNs.

| Node | Node rule | | | | |
|---|---|---|---|---|---|
| | **RBN 1** | **RBN 2** | **RBN 3** | **RBN 4** | **RBN 5** |
| **A** | RND(0.5) | RND(0.5) | RND(0.5) | RND(0.5) | RND(0.5) |
| **B** | RND(0.5) | RND(0.5) | RND(0.5) | RND(0.5) | RND(0.5) |
| **C** | (D⊕E) | (D⊕E) | L^(D⊕E) | (D⊕E) | (D⊕E) |
| **D** | (C⊕E) | (C⊕E) | (C⊕E) | (C⊕E) | (C⊕E) |
| **E** | (C⊕D) | (C⊕D) | (C⊕D) | (C⊕D) | (C⊕D) |
| **F** | RND(0.5) | RND(0.5) | RND(0.5) | (E⊕H) | (E⊕H) |
| **G** | RND(0.5) | RND(0.5) | RND(0.5) | (G+H+I+L)≥2 | RND(0.5) |
| **H** | (I⊕L) | E^(I⊕L) | E^(I⊕L) | (C⊕L) | (I⊕L) |
| **I** | (H⊕L) | (H⊕L) | (H⊕L) | (D+E+G+H)≥2 | (H⊕L) |
| **L** | (H⊕I) | (H⊕I) | (H⊕I) | F⊕(E⊕I) | (E⊕I) |
| **M** | RND(0.5) | RND(0.5) | RND(0.5) | RND(0.5) | RND(0.5) |
| **N** | RND(0.5) | RND(0.5) | RND(0.5) | RND(0.5) | RND(0.5) |

correct big group (formed of variables C, D, E, H, I and L)[3]. In RBN3, the slightly preeminent position of the first triplet in Fig. 1 is due to the particular set of samples that has been chosen; indeed, by analyzing several sets of samples both triplets are equally represented.

The structure in case RBN4 is highly heterogeneous and comprises loosely integrated parts: the $H_{noC}$ approach (though identifying correct nodes) is not able to spot out most variables composing the groups acting within the system, whereas the $H_{wiC}$ approach identifies almost all the correct nodes. The variables not detected are just nodes G and I, which indeed have a very low coupling with the other variables (see Table 1 for details): so, the $H_{wiC}$ approach seems to be more accurate than the $H_{noC}$ approach. In other words, the $T_c$ rank orders obtained by the two approaches are different, but often the $H_{wiC}$ approach identifies larger groups, which are also the correct ones. In case RBN5, for example, the most relevant group is composed of eight nodes and it is immediately identified by the $H_{wiC}$ approach, whereas the $H_{noC}$ approach identifies in the first positions only the small subsets composing the largest group of variables.

This hypothesis is supported by the analysis of the CSTR case: the $H_{noC}$ approach is able to identify merely small subsets of the ACS system, whereas the $H_{wiC}$ approach directly identifies in its first iteration almost all the members of the ACS. At the same time, this approach identifies also the largest part of the LCH structure. In this case, we repeated the RI analysis several times, by using different $H_{noC}$ and $H_{wiC}$ homogeneous systems: all these analyses consistently confirmed these results. Figure 2, which, for simplicity, shows only the relevant sets selected by the application of the sieving algorithm (two sets using $H_{noC}$ and two sets using $H_{wiC}$) which have a much higher value than the other possible sets, strongly supports these observations.

[3] data not shown

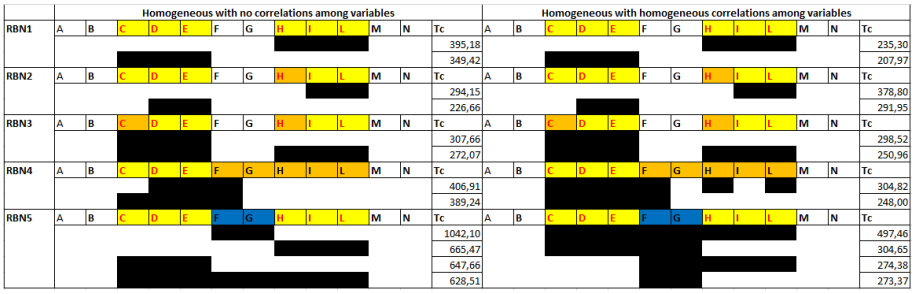


**Fig. 1.** The first two candidate relevant sets for each RBN case (four candidate relevant sets in the RBN5 case) and their $T_c$ values, computed by using the "classical" homogeneous system ($H_{noC}$, left) and by using the homogeneous system built using NORTA ($H_{wiC}$, right). In each row, a black cell indicates that the corresponding variable is selected in the candidate relevant set, whereas white cells denote the variables not belonging to the candidate relevant set. For each RBN case, we also report the correct solution, in which the different colors denote particular nodes or subdivisions of the dynamical structure of the systems. In particular: (i) case RBN1 hosts two dynamically-independent structures, (ii-iii) which in cases RBN2 and RBN3 are linked through the orange nodes; (iv) in case RBN4, a structure, observable also in case RBN1, is providing signals to other 5 nodes (highlighted in orange, and in turn exchanging messages among each other in various ways); (v) in case RBN5, the two structures, present also in case RBN1, are both sending signals to the blue nodes F and G. For a more detailed description see Table 1.

The LF scenario described in the previous part of the section is a particularly difficult case for the RI analysis. Indeed, the addition to a group of size $N_v$ of a variable, which is an almost perfect function of a variable already present within the group itself, leads to a new group of size $N_v + 1$ with a normalized integration very similar to the normalized integration of the initial group. Indeed, the integration of the group of size $N_v$ subtracted from the integration of the group of size $N_v + 1$ is equal to the entropy of the added variable: the same holds for the homogeneous system if the difference between the average integrations of groups of size $N_v + 1$ and of size $N_v$ is taken into consideration.

In the case of $p_{copy}$=1 (perfect followers) the $H_{noC}$ approach ranks in the top 130 positions almost all subsets of group G3 (of sizes 7, 6 and 8, in frequency order)[4], before identifying the correct G3 group, whereas the $H_{wiC}$ approach identifies the correct G3 group (with $T_c$= 522.803) immediately after its 8 subsets composed of 8 variables (with $T_c$ values slightly lower than 524). Figure 3 shows the superposition of these subsets, which highlights the presence of the G3 group, and the corresponding $T_c$ values range: indeed, the $H_{wiC}$ approach is able to discriminate among all the possibilities in a sophisticated way, thereby effectively identifying the correct G3 position.

[4] Notice that, in case of a perfect copy, the action of excluding a particular variable and including another one leads to groups having the same $T_c$ value.


**Fig. 2.** The two candidate relevant sets - obtained by applying the sieving algorithm to the results of the RI analysis - of the CSTR case and their $T_c$ values, computed using the "classical" homogeneous system ($H_{noC}$, first two rows) and using the homogeneous system built with NORTA ($H_{wiC}$, last two rows). The groups of variables remaining after the application of the sieving algorithm have $T_c$ values by far lower than those shown here. The blue and yellow colors indicate the chemical species belonging to LCH and ACS structures, respectively; darker colors indicate the chemical species produced by the reactions happening within the CSTR reactor (for these species the names are also reported). The constant species are not included in the table reported in this figure; the colored nodes without name indicate substrates or intermediate complexes.

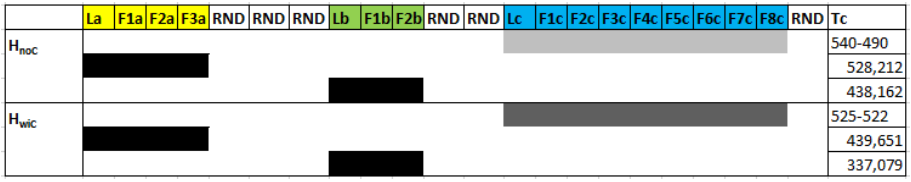


**Fig. 3.** The candidate relevant sets identified by using the "classical" homogeneous system ($H_{noC}$) and by using the homogeneous system built with NORTA ($H_{wiC}$), related with the Leader-Followers case analyzed in this paper (different colors highlight different LF groups). The variable sets reported in the top line (in light grey for $H_{noC}$ and in dark grey for $H_{wiC}$) actually represent the top-ranked 130 ($H_{noC}$) and 8 ($H_{wiC}$) sets, all subsets of the same variables, that were detected; in this case the $T_c$ column shows the range of the subsets' $T_c$ values. Notice that the $T_c$ range identified by the $H\_wiC$ approach is significantly smaller than the range identified by the $H_{noC}$ approach.

Eventually, both systems correctly identify the G1 and G2 groups. Similar results hold for $p_{copy}$=0.98 (data not shown).

## 5 Conclusion

In this paper, we have proposed an improvement to the RI method for identifying relevant subsets in complex systems. In particular, we have introduced a constant nonzero degree of statistical dependence in the variables composing the homogeneous reference system, by imposing that all variable pairs share the same pairwise correlation. The results coming from three relevant case studies demonstrate that this improvement allows one to identify sets of interacting variables of larger size compared with the previous way of generating the homogeneous system. Actually, because the analyzed systems and the $H_{wiC}$ approach feature the same integration values, we suspect that this fast identification of

larger groups might be related to their total amount of integration more than to their size. As future work we plan to verify this hypothesis by analyzing systems where dynamical structures of different size exhibit similar integration levels. Other possible future developments may regard the application of the new method of computing the homogeneous system to complex systems with many more variables and to verify its performance also by applying some metaheuristics or an iterative version of the sieving procedure, in order to identify hierarchical relations among RSs.

## References


1. Balakrishnan, V.: Graph Theory. McGraw Hill (1997)
2. Balduzzi, D., Tononi, G.: Integrated information in discrete dynamical systems: Motivation and theoretical framework. PLOS Computational Biology 4(6), 1–18 (06 2008)
3. Barrett, A.B., Seth, A.K.: Practical measures of integrated information for time-series data. PLOS Computational Biology 7(1), 1–18 (01 2011)
4. Bazzi, M., Porter, M.A., Williams, S., McDonald, M., Fenn, D.J., Howison, S.D.: Community detection in temporal multilayer networks, with an application to correlation networks. Multiscale Model. Simul. 14(1), 1–41 (Jan 2016)
5. Bossomaier, T., Barnett, L., Harré, M.: Information and phase transitions in socio-economic systems. Complex Adaptive Systems Modeling 1(1), 9 (2013)
6. Cario, M.C., Nelson, B.L.: Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Tech. rep. (1997)
7. Cover, T., Thomas, A.: Elements of information theory, 2nd Edition. Wiley-Interscience, New York (2006)
8. Cross, M.C., Hohenberg, P.C.: Pattern formation outside of equilibrium. Rev. Mod. Phys. 65, 851–1112 (Jul 1993)
9. Filisetti, A., Villani, M., Roli, A., Fiorucci, M., Poli, I., Serra, R.: On some properties of information theoretical measures for the study of complex systems. In: Pizzuti, C., Spezzano, G. (eds.) Advances in Artificial Life and Evolutionary Computation: 9th Italian Workshop, WIVACE 2014, Vietri sul Mare, Italy, May 14-15, Revised Selected Papers. pp. 140–150. Springer International Publishing, Cham (2014)
10. Gershenson, C., Fernandez, N.: Complexity and information: Measuring emergence, self-organization, and homeostasis at multiple scales. Complex. 18(2), 29–44 (Nov 2012)
11. Johnson, J.: Hypernetworks in the Science of Complex Systems. Imperial College Press (2013)
12. Kauffman, S.: The origins of order. Oxford University Press (1993)
13. Kivel a, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. Journal of Complex Networks 2(3), 203–271 (2014)
14. Lewis, T.G.: Network Science: Theory and Applications. Wiley Publishing (2009)
15. Mansy, S., Schrum, J., Krishnamurthy, M., Tobe, S., Trecol, D., Szostak, J.: Template-directed synthesis of a genetic polymer in a model protocell. Nature 454 (2008)
16. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113 (Feb 2004)

17. Nuño, E., Cutululis, N.: A heuristic for the synthesis of credible operating states in the presence of renewable energy sources. In: 2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS). pp. 1–7 (Oct 2016)
18. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, NY, USA (2000)
19. Prokopenko, M., Boschetti, F., Ryan, A.J.: An information-theoretic primer on complexity, self-organization, and emergence. Complexity 15(1), 11–28 (2009)
20. Prokopenko, M., Lizier, J.T., Obst, O., Wang, X.R.: Relating fisher information to order parameters. Phys. Rev. E 84, 041116 (Oct 2011)
21. Roli, A., Villani, M., Caprari, R., Serra, R.: Identifying critical states through the relevance index. Entropy 19(2) (2017)
22. Sani, L., Amoretti, M., Vicari, E., Mordonini, M., Pecori, R., Roli, A., Villani, M., Cagnoni, S., Serra, R.: Efficient search of relevant structures in complex systems. In: AI*IA 2016 Advances in Artificial Intelligence: XVth International Conference of the Italian Association for Artificial Intelligence, Genova, Italy, November 29 – December 1, 2016, Proceedings. pp. 35–48. Springer International Publishing, Cham (2016)
23. Sani, L., Lombardo, G., Pecori, R., Fornacciari, P., Mordonini, M., Cagnoni, S.: Social relevance index for studying communities in a facebook group of patients. In: Sim, K., Kaufmann, P. (eds.) Applications of Evolutionary Computation. pp. 125–140. Springer International Publishing, Cham (2018)
24. Serra, R., Villani, M., Semeria, A.: Genetic network models and statistical properties of gene expression data in knock-out experiments. Journal of Theoretical Biology 227(1), 149 – 157 (2004)
25. Shalizi, C., Camperi, M., Klinkner, K.: Discovering functional communities in dynamical networks. In: Airoldi, E. et al. (ed.) Statistical Network Analysis: Models, Issues, and New Directions: ICML 2006 Workshop on Statistical Network Analysis, Pittsburgh, PA, USA, June 29, 2006, Revised Selected Papers. pp. 140–157. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
26. Silvestri, G., Sani, L., Amoretti, M., Pecori, R., Vicari, E., Mordonini, M., Cagnoni, S.: Searching relevant variable subsets in complex systems using k-means pso. In: Pelillo, M., Poli, I., Roli, A., Serra, R., Slanzi, D., Villani, M. (eds.) Artificial Life and Evolutionary Computation. pp. 308–321. Springer International Publishing, Cham (2018)
27. Sporns, O., Tononi, G., Edelman, G.: Theoretical neuroanatomy: Relating anatomical and functional connectivity in graphs and cortical connection matrices. Cerebral Cortex 10(2), 127–141 (2000)
28. Tononi, G., McIntosh, A., Russel, D., Edelman, G.: Functional clustering: Identifying strongly interactive brain regions in neuroimaging data. Neuroimage 7, 133–149 (1998)
29. Tononi, G., Sporns, O., Edelman, G.M.: A measure for brain complexity: relating functional segregation and integration in the nervous system. Proceedings of the National Academy of Sciences 91(11), 5033–5037 (1994)
30. Vicari, E., Amoretti, M., Sani, L., Mordonini, M., Pecori, R., Roli, A., Villani, M., Cagnoni, S., Serra, R.: GPU-Based Parallel Search of Relevant Variable Sets in Complex Systems, pp. 14–25. Springer International Publishing, Cham (2017)
31. Villani, M., Filisetti, A., Benedettini, S., Roli, A., Lane, D., Serra, R.: The detection of intermediate-level emergent structures and patterns. In: Miglino, O. et al. (ed.) Advances in Artificial Life, ECAL 2013. pp. 372–378. The MIT Press (2013), http://mitpress.mit.edu/books/advances-artificial-life-ecal-2013

32. Villani, M., Sani, L., Amoretti, M., Vicari, E., Pecori, R., Mordonini, M., Cagnoni, S., Serra, R.: A relevance index method to infer global properties of biological networks. In: Pelillo, M., Poli, I., Roli, A., Serra, R., Slanzi, D., Villani, M. (eds.) Artificial Life and Evolutionary Computation. pp. 129–141. Springer International Publishing (2018)
33. Villani, M., Sani, L., Pecori, R., Amoretti, M., Roli, A., Mordonini, M., Serra, R., Cagnoni, S.: An iterative information-theoretic approach to the detection of structures in complex systems. Complexity 2018 (2018)
34. Wang, X., Lizier, J., Prokopenko, M.: Fisher information at the edge of chaos in random boolean networks. Artificial Life 17(4), 315–329 (2011)
35. Xie, W., Nelson, B.L., Barton, R.R.: Statistical uncertainty analysis for stochastic simulation with dependent input models. In: Proceedings of the Winter Simulation Conference 2014. pp. 674–685 (Dec 2014)
36. Xu, X., Yan, Z.: Probabilistic load flow evaluation with hybrid latin hypercube sampling and multiple linear regression. In: 2015 IEEE Power Energy Society General Meeting. pp. 1–5 (July 2015)
37. Zubillaga, D., Cruz, G., Aguilar, L.D., Zapotécatl, J., Fernández, N., Aguilar, J., Rosenblueth, D.A., Gershenson, C.: Measuring the complexity of self-organizing traffic lights. Entropy 16(5), 2384–2407 (2014), http://www.mdpi.com/1099-4300/16/5/2384