Entropic Detection of Chromatic Community Structures

Franck Delaplace

Entropic Detection of Chromatic Community Structures

Franck Delaplace

IBISC-lab, Paris-Saclay University, Univ. Evry
franck.delaplace@univ-evy.fr

Abstract

The detection of community structure is probably one of the central trends in complex network emphasizing the complex internal organization of people, molecules or processes behind social, biological or computer networks... The issue is to provide a network partition representative of this organization so that each community presumably gathers nodes sharing a common mission, purpose or property. Usually the identification is based on the difference between the connectivity density of the interior and the boundary of a community. Indeed, nodes sharing a common purpose or property are expected to interact closely. Although this rule appears mostly relevant, some fundamental scientific problems like disease module detection highlight the inability to determine significantly the communities under this connectivity rule. The main reason is that the connectivity density is not correlated to a shared property or purpose. Therefore, another paradigm is required for properly formalize this issue in order to meaningfully detect these communities. In this article we study the community formation from this new principle. Considering colors formally figures the shared properties, the issue is thus to maximize group of nodes with the same color within communities. We study this novel community framework by introducing new measurement called *chromatic entropy* assessing the quality of the community structure regarding this constraint. Next we propose an algorithm solving the community structure detection based on this new community formation paradigm.

Keywords: Community structure, Detection algorithm, Complex Network

1 Introduction

Complex networks model component interactions in diverse real-world domains as in sociology with social or friendships networks, computer science with WEB, and biology with regulatory, metabolic or neural networks. Nodes of these networks are often arranged in closely tight groups called communities. These

communities delineate the organizational supports of function, property, purpose or categories. They thus highlight a structure of the network providing an organizational understanding behind the topology. Formally, the goal is to identify a node partition of the network. A *community structure* is a partition of the vertices of a graph defined according rules structuring the vertex distribution. Although there is no firm answer concerning these rules [17], it is commonly admitted that the definition of a community relates to a difference in connection density between its interior and its boundary. The density of connection between nodes inside a community must be higher than the density of connection across communities. Such community obtained by this method is called the *topological community* [13]. Community detection algorithms capture this difference of connection density for detecting communities in a network [8, 18]. The quality of a community structure is evaluated by a measure assessing this partitioning rule. A recognized standard is the *modularity* introduced by Newmman [5]. The modularity is based on the comparison of the network with a random one having the same topological characteristics than the original one (*i. e.,* same number of nodes, same node degree). Therefore a good measure must be greater than a community structure having the same characteristics but obtained by chance because this reveals an organizational bias. Finding a community structure maximizing the modularity is NP-hard [4] and different heuristics have been proposed for detecting the best community structure [3, 7, 8, 9].

While the concept of community is central in network science, the connection density rule fails to significantly identify the meaningful community structure of a network for some issues, thus restricting the applicability of community detection algorithms. It is notably the case for *disease module* discovery. A disease module groups genes which are mechanistically linked to the same pathophenotype.The study of the modularity of human disease would provide a causal understanding of the pathogenesis strengthening the etiological explanation and rationally determine clues for drug target discovery.

In [15], the authors carefully demonstrate that disease module are not topological module/community. By using three representative, methodologically distinct algorithms on community structure detection based on density connection, the authors show that the disease genes gathered in a community by connection density method are drastically under-represented, thus prohibiting the ability to assign communities to diseases. Moreover, they also show that this lack of representativeness is not due to an insufficiency of knowledge about genetic diseases, but rather to the inadequacy of the density connection method to properly address the disease module. This empirical analysis is explained by the authors by the fact that the disease proteins do not form particularly dense subgraphs. This conclusion is also confirmed by other works on the disease module domain [16, 20, 19] which propose alternative clustering methods based on other rules than those governing topological community detection.

Because of its overarching importance in health, the identification of disease modules clearly states the need to extend this framework for detecting community structures by including other categories of problems. Therefore, based

on the disease module, our objective is to generalize its principles from the proposed method in order to characterize an alternative community detection paradigm.

DIAMOND [15], GLADIATOR [20], and SCA [21] are three computational methods solving the disease module detection based on different approaches. However, they share some common features allowing us to state the fundamental rules for finding disease module.

The genes implicated in a disease are retrieved from databases analysis as OMIM[2, 14] for Mendelian diseases or ORPHANET [19] for orphan diseases. They constitute the landmarks of the disease at molecular level and reciprocally a fundamental property assigned to these genes from which the disease module can be detected. Hence this property is central and monitor the community structure detection.

A backbone of network biology lies on the "local hypothesis" stating that genes or proteins involved in the same disease have a tendency to interact with each other [11] and to cluster in the same neighborhood [6, 10]. Hence, all disease related genes in a module are necessary connected together over a short distance. Connectivity analysis depends on algorithmic methods, and two disease-related genes may or may not be considered neighbors. DIAMOND examines the neighborhood of gene by identifying a typical connection pattern that must differ to random/null model connection. They are thus looking for a characteristic connectivity pattern between disease genes. The connection rules of GLADIATOR are based on the reproduction of connection obtained by phenotypic similarity analysis, while SCA reconnects the disease seeds by few extra hidden nodes qualified as seed connectors while complying with a short connectivity distance between seeds.

All these algorithms aims at finding the largest modules encompassing the greatest number of genes related to a disease, and stop when no improvements are possible. Therefore the definition of a module relates here to largest number of connected nodes which mostly share the same property.

Disease module detection exemplifies an important problem for community structure inference where the condition underpinning the node partition is related to alternative criteria than connection density difference between the interior and the boundary of a community. Therefore, it seems greatly beneficial for extending the scientific questioning on network community that the resolution of this problem is achieved in a broader context than disease modules, impelling to generalize the statement of this problem.

The common property which is responsible for the formation of the community must be understood in a broad sense including a wide variety of situations such as involvement in the same process or function, membership of a social or ethnic group, identical characteristics, sharing a common topic of interest, common purpose or mission etc., more generally any trait that can be shared by a community and qualifying its members. This property will be formally assimilated to a "color" leading to assigning the same color to the nodes having the same property. Accordingly, the issue of *chromatic community structure* detection is to find communities of connected nodes that maximize the density

of the major color within each.

Such problem statement explains why the connection density based algorithms may fail to detect such communities because the nodes with the same colors can be sparsely connected since only the connectedness prevails and potentially separated by nodes differently colored. As there is no a priori relationships between colors and connections, nodes with the same color can be located through communities obtained by connection density rule.

In this article, we study the chromatic community structure detection problem and propose an algorithm for finding partition of communities. In Section 2 we mathematically formalize the problem. We then define in Section 3 the *chromaritic entropy* which is a measure assessing the significance of a chromatic community structure. We detail in Section 4 an algorithm finding a chromatic community structure. The algorithm is then evaluated in Section 5 before concluding (Section 6).

2 Formalizing the coloring

In this section we address the basic notions related graph coloring. Let $G = \langle V, E\rangle$ be a graph where V is a set of vertices and $E \subseteq V \times V$ a set of edges, a *community* p is a subset of V (*i. e.,* $p \subseteq V$) and a *community structure* P is a partition of V, namely:

$$\bigcup_{p_i \in P} p_i = V \wedge \forall p_i, p_j \in P : p_i \cap p_j \neq \emptyset \implies p_i = p_j.$$

A community structure based on color selection criteria is called a *chromatic community structure.*

Coloring profile. Coloring assigns a color to each vertex of a graph which is described by a *coloring profile* corresponding to an application from vertex to color $c : V \to C$ where C denotes the set of colors. The set of colors C will be represented by an integral interval $[1, r]$ where integers define colors. For example $c = \{1 \mapsto 1, 2 \mapsto 1, 3 \mapsto 2, 4 \mapsto 3, 5 \mapsto 3\}$ assigns color 1 to nodes $1, 2$, color 2 to node 3 and color 3 to nodes $4, 5$. The restriction of the coloring to community denoted c_p for community $p \subseteq V$ is defined as: $c_p = \{v \mapsto c(v) \mid v \in p\}$.

If the vertices correspond to an integral interval $V = [1, n]$ then the coloring profile can be described by a vector such that the index stands for a vertex label and its corresponding value for a color (*i. e.,* $c(i) = k \iff i \mapsto k \in c$). For example $c = \{1 \mapsto 1, 2 \mapsto 1, 3 \mapsto 2, 4 \mapsto 3, 5 \mapsto 3\}$ is described by the vector $(1, 1, 2, 3, 3)$.

Colored Graph. A *colored graph* is a $3-$uple $\langle V, E, c\rangle$. The colored graph in Figure 1 uses 3 colors $C = [1, 3]$ where: green$= 1$, red$= 2$ and yellow$= 3$. From

its coloring profile:

$$c = \{1 \mapsto 1, 2 \mapsto 3, 3 \mapsto 1, 4 \mapsto 2, 5 \mapsto 1, 6 \mapsto 3\}$$

the vector representation is $(1, 3, 1, 2, 1, 3)$. Given the following chromatic community structure:

$$P = \{p_1 = \{1, 3, 4, 5\}, p_2 = \{2, 6\}\},$$

we deduce the following coloring profiles restricted to p_1, p_2:

$$c_{p_1} = \{1 \mapsto 1, 2 \mapsto 3, 4 \mapsto 2, 5 \mapsto 1\}, c_{p_2} = \{2 \mapsto 3, 6 \mapsto 3\}.$$

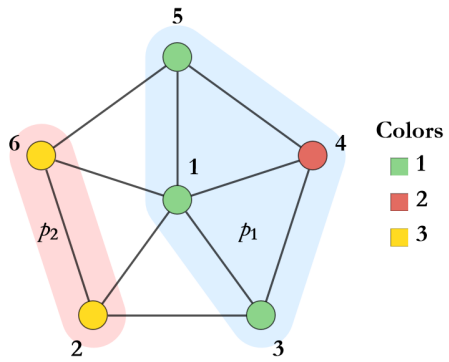


Figure 1: Community structure of a colored graph.

Transparency. The absence of properties of a vertex is represented by the *transparency* (denoted 0) since a color is assumed to qualify a property or an attribute of a vertex. The transparency is not a color *i. e.,* $0 \notin C$. Transparent vertices are therefore never involved by color, but the transparent vertices still exist as vertices.

Chromatic function. A *chromatic function* $\chi : (V \to C) \to C \to \mathbb{N}$ counts the number of occurrences of each color in a coloring profile. The formal definition of the chromatic function is based on the *counting operator* (**Count**) which is a function counting the positions/nodes of each element corresponding to values of a vector or a function. $\mathbf{Count}(X, y)$ specifically counts the number of occurrences of element y in vector/function X:

$$\mathbf{Count}(X, y) = \left\{y \mapsto \left|\{i \mid X(i) = y\}\right|\right\}.$$
$$\mathbf{Count}(X) = \bigcup_{i=1}^{|X|} \mathbf{Count}(X, X(i)).$$

The chromatic function is thus defined from a coloring profile c as:

$$\chi_c = \bigcup_{k \in C} \mathbf{Count}(c, k) \tag{1}$$

The chromatic function of c of the example in Figure 1 is:

$$\chi_c = \{1 \mapsto 3, 2 \mapsto 1, 3 \mapsto 2\}.$$

For the following coloring profile $c = (0, 0, 1, 2, 1, 0, 1, 3, 0, 3)$ with transparent color, we also have the same chromatic function because the transparency is not accounted as color by definition (1) since $0 \notin C$.

Finally, the density also includes the transparency since it corresponds to the ratio of the number d of vertices with the same color by the number n of vertices in a community ($\frac{d}{n}$). As example, from the previous coloring profile with transparency, the density of color 1 ($d = 3$) is $\frac{3}{10} = 0.3$.

Dominant color. A coloring profile with d vertices of the same color, will be called a *$d-$coloring profile*. This notion is also applied to community from their local coloring profile. A *$d-$colorful community* p implies that:

$$\exists k \in C : \chi_{c_p}(k) = d. \tag{2}$$

Notice that these coloring profiles may also have several subsets of vertices with the same color of cardinality greater or equal to d. The graph in Figure 1 is a $3-$coloring profile for color 1, but also a $2-$coloring profile for color 3, and $1-$coloring profile for color 2.

Among the $d-$coloring profiles we specifically focus on the class of profiles where d is the cardinality of the color occurring the most. These profiles are said *$d-$dominant* by this main color. Hence a coloring profile is $d-$dominant if and only if:

$$\exists k \in C : \chi_c(k) = d \wedge \forall k' \in C : \chi_c(k') \leq d. \tag{3}$$

In this case, color $k \in \arg\max \chi_{c_p}$ is said *dominant*. In Figure 1 the dominant color is 1 and the coloring profile is thus $3-$dominant. By extension, a community is said $d-$dominant if the restriction of the coloring profile to this community is $d-$dominant. In Figure 1, p_1 is $3-$dominant for color 1 and p_2 is $2-$dominant for color 3. Notice that several dominant colors may exist in a coloring profile.

3 Chromatic Entropy

The meaningfulness of a chromatic community structure will be deduced from a measure. Although, the significance of the colorful communities closely depends on the application fields for interpreting the colors, the issue is to define a generic measure assessing the significance of a chromatic community structure.

Basically this measure is related to the dominant color in each community. Intuitively more a color dominates more significant a community is.

However, this characteristic is not enough for relevantly qualifying the significance of community structure. Indeed, as an extreme illustrative example, let us consider a community structure where each community is reduced to a single node. Such structure leads to optimal coloring of the communities since the single node owns the dominant color in its community because it covers it totally. However such community structure clearly tells us nothing of value about community organization since all nodes remain isolated.

A relevant measure should assess the intentionality behind the design of a community. By considering that the human design driven by intention is opposed to chance, a significant community should thus lead to gather more nodes of the dominant color than would be expected by chance. Indeed, the situation that cannot be delivered by chance underpins a mechanistic organization representing an human intention. As a result, we can safely conclude that the structure of the chromatic community excluding the chance would provide a meaningful structure underpinning an intentional organization. Such perspective raises two major issues: 1) defining a measure characterizing the intention in community design, 2) formally characterizing the probability to generate a d-colorful community by chance.

3.1 Chromatic entropy definition

In complex system analysis, the entropy is a concept commonly used to quantify disorder, randomness, chaos, or uncertainty in various fields [1, 12]. By incorporating entropy into the community detection process, the goal is to find partitions that maximize the quality of community structure while minimizing the randomness within communities considered as the sign of community disorder and disorganization. This approach would reveal meaningful communities in complex networks, leading to a better understanding of the underlying community structure and their organization law. This framework thus appears suitable for assessing how much a community is the proceed of an intentional construction.

In our context, the *chromatic entropy* H quantifies the intentionalness of the community design. The chromatic entropy will relate to the coloring of a community obtained by chance: the more likely a community is to be colored by chance, greater its entropy. A community structure with a small entropy thus emphasizes a meaningful community structure. Accordingly, the chromatic entropy is based on the quantification of the community organization intentionally designed, called the *intentionalness quantity* and denoted $I : \Delta_1 \to \mathbb{R}$. Intuitively, this quantity defines how much a community is intentionally organized. This measure is semantically equivalent to the measure of information introduced by Shannon. It is expected that the higher the probability of random community generation, the lower the Intentionalness quantity.

Let $\Delta_m = \{(p_1, \cdots, p_m) \mid 0 \leq p_i \leq 1 \wedge \sum_{i=1}^{m} p_i \leq 1\}$ be the sets of (m-ary),

possibly incomplete, probability distributions on m communities. The entropy H is a continuous function defined as $H : \Delta^* \to \mathbb{R}$ with $\Delta^* = \bigcup_{m \geq 1} \Delta_m$. Δ^* as domain is used for mathematical convenience to accommodate any community structure cardinality and $\Delta_1 \subseteq [0, 1]$ thus stands for a subset of the unit interval. For characterizing the chromatic entropy, we focus on the axiomatic properties framing the definition of this function (Table 1). Notice that the maximality is a property specific to our context that does not necessarily apply to the other notions of entropy as are the other properties. It is worth noticing that the

Definition	**Property**
Non negative : The entropy cannot be negative since it is a metric.	$H(p) \geq 0$
Expansibility: adding a community with probability zero does not change the entropy of the structure.	$H(p, 0) = H(p)$
Symmetry: The entropy is insensitive to a permutation on probability distribution.	$H(p_1, p_2) = H(p_2, p_1)$
Sub Additivity: The entropy of a community structure is less than or equal to the sum of the entropies of the communities composing it.	$H(p_1 p_2) \leq H(p_1) + H(p_2)$.
Minimality: The community structure is assumed to be totally meaningful with a minimal entropy when the probability is null.	$H(0) = 0$
Maximality: The entropy is maximal when the probability is 1 because the community is assumed to be fully random.	$H(1) = \infty$

For the sake of simplicity, we define the properties with the minimal number of parameters requested for their definition.

Table 1: Properties of the entropy

definition of the Shannon entropy [1] cannot be straightforwardly used due to the maximality property, since $-p \log_2 p = 0$ with $p = 1$ and not ∞.

By setting the intentionalness quantity as $I(p) = \log_2(1-p)$ which fulfills the expected requirements the chromatic entropy can be finally defined as follows (Definition 1):

Definition 1 (Chromatic Entropy)**.**

$$H(p) = -p\, I(p) = -p \log_2(1 - p)$$

The extension to a distribution of probabilities Δ_m follows the usual gener-

alized form of entropy-function:

$$H(p_1, \cdots, p_m) = \sum_{i=1}^{m} H(p_i).$$

3.2 Probability of random coloring

The probability to randomly generate a d-colorful community with a particular color chosen among $|C| = r$ colors by chance is defined by the ratio of the favorable cases to the possible cases. The number of the whole possible colored communities is r^n corresponding to the cardinal of the complete enumeration of the possible combinations of vertex coloring among r colors. The definition of the favorable cases necessitates to combinatorically enumerate them which is harder to characterize than the possible cases community structure is assumed to be totally unorganized. Two issues are addressed:

1. the enumeration of the d-colorful communities of size n considering r colors;

2. the enumeration of the $d-$dominant colorful communities of size n considering r colors.

The first issue does not impose the domination but just the cardinality of a subset of vertices with the same color while the second refers exactly to the definition of a $d-$dominant coloring profile. The separation of the enumeration problem in two issues is motivated by the computational complexity of the resulting combinatorial formulas explained in Section 5. We thus need to enumerate the favorable colorful communities for each issue. Subsection 3.2.1 defines the combinatorial formula enumerating the favorable colorful communities for issue 1, while Subsection 3.2.2 determines it for issue 2.

3.2.1 Enumeration of $d-$colorful communities

Different coloring of d vertices are obtained using any color. Let D_k be the set of colorful communities having d vertices of color k, the count of all communities containing a $d-$color profile obviously corresponds to the cardinality of the union of these sets, namely: $|\bigcup_{k=1}^{r} D_k|$. Some communities may have a d-color profile for different colors, meaning that these sets intersect. The enumeration formula of $|\bigcup_{k=1}^{r} D_k|$ is based on the *Poincaré sieve* (inclusion-exclusion principle), for the cardinal of the union:

$$\left|\bigcup_{k=1}^{r} D_k\right| = \sum_{k=1}^{r}(-1^k) \sum_{1 \leq i_1 \leq \cdots \leq i_j \leq \cdots \leq i_k \leq r} |D_{i_1} \cap \cdots \cap D_{i_j} \cap \cdots \cap D_{i_k}|$$

.

For example let us considering 3 sets, from the *Poincaré sieve* the cardinal is then (see Figure 2)

$$\begin{aligned}|D_1 \cup D_2 \cup D_3| = |D_1| + |D_2| + |D_3| \\ - (|D_1 \cap D_2| + |D_1 \cap D_3| + |D_2 \cap D_3|) \\ + |D_1 \cap D_2 \cap D_3|.\end{aligned}$$

To obtain the formula enumerating the $d-$colorful communities, we thus need to define a combinatoric formula for each set and each intersection of sets.

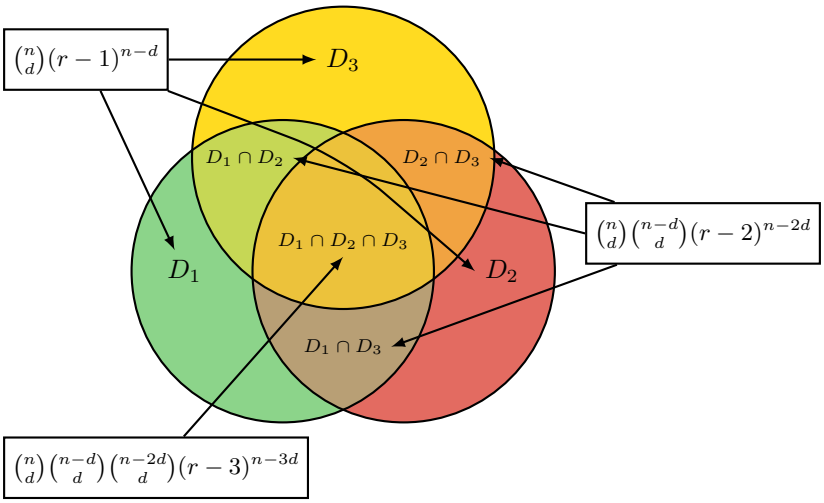


Figure 2: Ven diagram of the union of 3 sets of $d-$colorful communities.

Figure 2 shows the combinatorial formulas for all intersection cases of 3 sets (see the Appendix for a detailed explanation of each formula). For 3 sets the formula is thus:

$$3\binom{n}{d}(r-1)^{n-d} - 3\binom{n}{d}\binom{n-d}{d}(r-2)^{n-2d} + \binom{n}{d}\binom{n-d}{d}\binom{n-2d}{d}(r-3)^{n-3d},$$

which can be simplified by setting $r = 3$ into:

$$\binom{n}{d}\left(\left(0^{n-3d}\binom{n-2d}{d} - 3\right)\binom{n-d}{d} + 3\ 2^{n-d}\right),$$

considering that $0^0 = 1$.

Theorem 1 provides the general enumeration formula deduced from the Poincaré sieve once each intersection is combinatorically defined.

Theorem 1. *The count of $d-$colorful communities of size n with r colors is given by κ function:*

$$\kappa(r,n,d) = \sum_{k=1}^{\min\left(r,\lfloor \frac{n}{d} \rfloor\right)} \frac{(-1)^{k-1}\binom{r}{k} n!(r-k)^{n-kd}}{(n-kd)!(d!)^k}$$

The proof is in the Appendix □

3.2.2 Enumeration of the $d-$dominant colorful communities

The domination implies to include the dominance constraint in comparison to the $d-$colorful communities enumeration, leading to specify the different equivalence classes of communities complying with the dominations conditions 3. Each class addresses the number of nodes for each color while fulfilling the dominance condition. Since the conditions of domination are only based on the number of vertices of the same color regardless the color, if two chromatic functions of two communities p, q are equal up to a permutation on colors $\pi : C \to C$, $\chi_{c_p} = \pi \circ \chi_{c_q}$ then these communities share the same domination property. Thus they belong to the same equivalence class related to the color distribution.

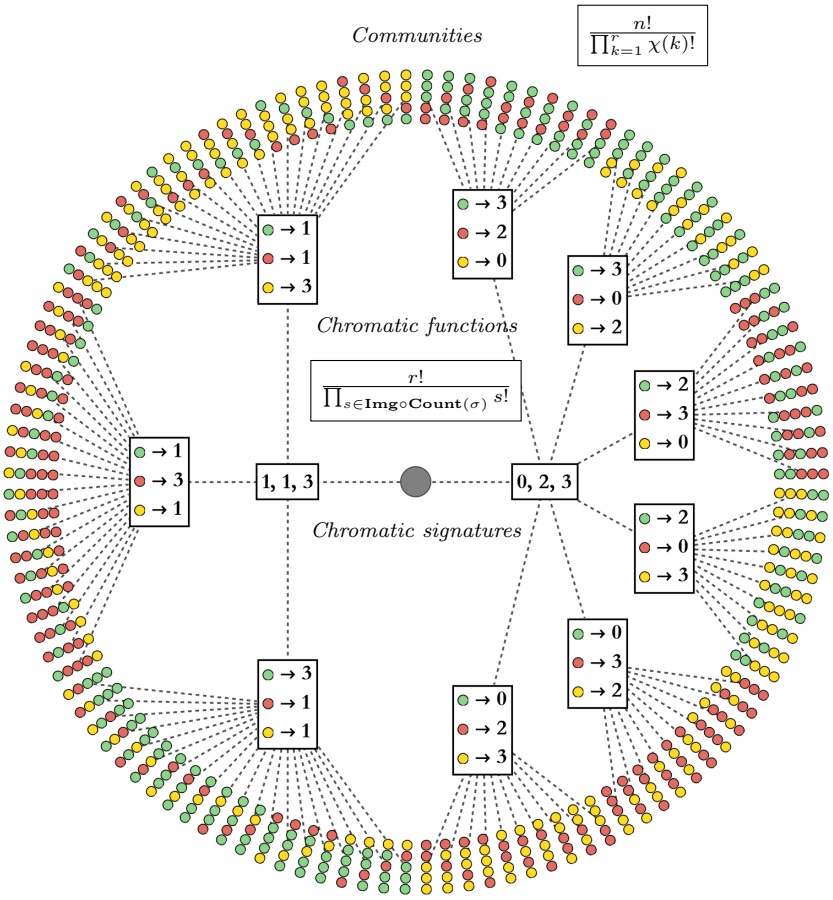
We introduce the notion of *chromatic signature* σ to capture this equivalence on chromatic functions. A signature of a chromatic function is a vector of color count corresponding to its ordered image (Definition 4)

$$\sigma_p = \mathbf{Sort} \circ \mathbf{Img}\ \chi_{c_p} \tag{4}$$

Several chromatic functions may have the same signature. For example the two chromatic functions: $\{1 \mapsto 0, 2 \mapsto 3, 3 \mapsto 2\}$ and $\{1 \mapsto 3, 2 \mapsto 0, 3 \mapsto 2\}$ have the same chromatic signature which is: $(0, 2, 3)$. The signatures are at he heart of the combinatorial formula enumerating the $d-$dominant coloring profiles by abstracting the chromatic functions. We can deduce that a signature of a $d-$dominant color profile complies with the following conditions:

$$\begin{cases} \sigma(r) = d \wedge \\ \sum_{i=1}^{r} \sigma(i) = n \wedge \\ \forall 1 \leq i \leq r : \sigma(i) \leq d \wedge \\ \forall 1 \leq i, j \leq r : i \leq j \implies \sigma(i) \leq \sigma(j). \end{cases} \tag{5}$$

A chromatic signature properly defines an equivalence class on communities with regard to the domination property. Indeed, two communities with an equal chromatic signature share the same domination property (*i. e.,* $\sigma_p = \sigma_q \iff p \sim q$). Thus, each equivalence class specializing the property of domination according to the count of each color leads to a specific signature (Figure 3). Let $\mathcal{S}_{r,n,d}$ be the set of all possible dominant signatures (DSS) with respect to parameters r, n, d, this set is explicitly generated by collecting all the



PARAMETERS: $r = 3, n = 5, d = 3$. Two dominant signatures are deducted $(1, 1, 3)$ and $(0, 2, 3)$ which respectively correspond to 3 and 6 chromatic function groups. 20 communities are associated with each chromatic function of the first group and 10 for the second. A total of 120 communities are 3−dominant. The framed formulas correspond respectively to the number of chromatic functions of a signature (under "Chromatic functions") and to the number of communities dominant for a chromatic function (near "Communities").

Figure 3: Enumeration of 3−dominant communities of size 5 with 3 colors.

signatures following Definition 5. As example the DSS for $r = 4, n = 9, d = 4$ is $\mathcal{S}_{4,9,4} = \{(0,1,4,4),(0,2,3,4),(1,1,3,4),(1,2,2,4)\}$. It represents the core of the combinatorial formula enumerating the $d-$dominant colorful communities. The algorithm computing this set is given in the Appendix.

Figure 3 shows the distribution of the dominant colored communities into two equivalence classes distinguished by their color count. The communities are first grouped according to the chromatic function equality and next according to their signature equality by gathering the chromatic functions with the same signature. Counting all the $d-$dominant colorful communities intuitively follows this hierarchical division. From signatures, we first count the chromatic functions corresponding to them and then for each chromatic function we count the possible coloring profiles leading to this chromatic function. The final count of the dominant colorful communities is the product of these two steps. Theorem 2 defines the count of the $d-$dominant communities.

Theorem 2. *The count of all possible $d-$dominant communities of size n with r colors is given by γ function:*

$$\gamma(r,n,d) = n!r! \sum_{\sigma \in \mathcal{S}_{r,n,d}} \frac{1}{\prod_{s \in \mathbf{Img} \circ \mathbf{Count}(\sigma)} s! \ \prod_{i=1}^{r} \sigma(i)!}.$$

The proof is in the Appendix. □

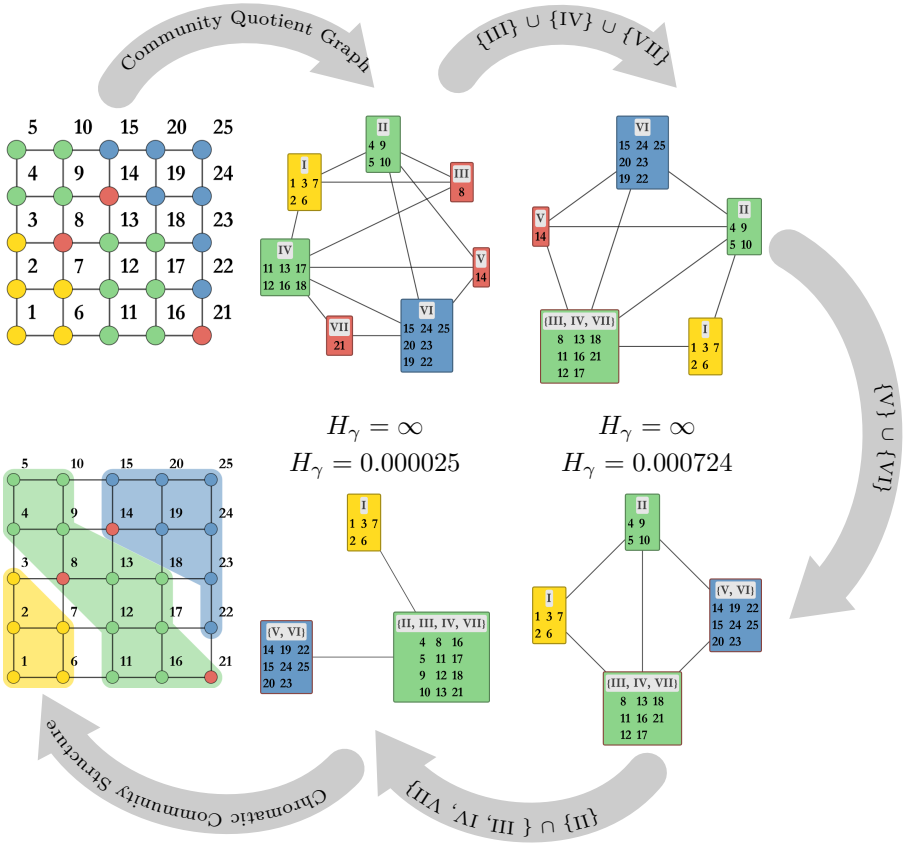
3.2.3 Probability of random coloring

From the enumeration of the colorful communities, we can formally define the probability of random coloring for normal or dominant coloring. This corresponds to the ratio of the favorable cases to the possible cases where the favorable cases is given by κ or γ while the number of all possible colorful communities is r^n. Therefore, for a community p such that $n = |p|$ with a coloring profile c_p distributing r colors into vertices of p, and considering the largest number of vertices of the same color $d = \max \chi_{c_p}$. These probabilities are respectively:

$$p_\kappa = \frac{\kappa(r,n,d)}{r^n},\ p_\gamma = \frac{\gamma(r,n,d)}{r^n}$$

4 Chromatic community structure detection

The chromatic community detection algorithm (CHROCODE) finds a partition of a colored graph minimizing the chromatic entropy H. The algorithm is divided in two phases: first a partition grouping connected nodes of the same color is built, forming as partition of monochrome communities, and next these communities are iteratively merged to decrease the chromatic entropy until no merges can improve the solution. The input parameters of the algorithm are the colored graph G, c, a neighborhood distance δ, and a probability law p_κ or p_γ. The algorithm was originally inspired by the Louvain algorithm [7] although



The labels of the cluster of nodes that are vertices of the quotient graph are in Roman while the nodes of the original graph are labeled in Arabic.
PARAMETERS: $n = 25, \delta = 2, r = 4$.

Figure 4: CHROCODE algorithm steps.

the specificity of the chromatic community structure framework leads to a significantly different program. CHROCODE is freely distributed in two open-source implementations: in Python [23] and in Mathematica-Wolfram [22]. The algorithm is completely detailed in the Appendix. In more detail, the tasks carried out during these two stages are:

1) Connected monochrome community structure. From a colored graph $\langle V, E, c\rangle$, a community is designed from a vertex seed by first integrating neighboring vertices of the same color, then extending it by integrating their respective neighborhood having the same color and so on. Once no supplementary vertices can be added, the current community is closed and stored. Another vertex is then chosen as seed until no vertices are available. The resulting community structure P is composed of monochrome communities.

2) Fusion of monochrome communities. From the monochrome communities P previously obtained, we define a quotient graph Q where each community becomes a node of this graph ($Q = \langle P, E_P\rangle$). There exists a link between two community-nodes if there already exists a link between some nodes composing the respective communities ($E_P = \{(p_i, p_j) \mid \exists (v_i, v_j) \in E : v_i \in p_i \wedge v_j \in p_j\}$).

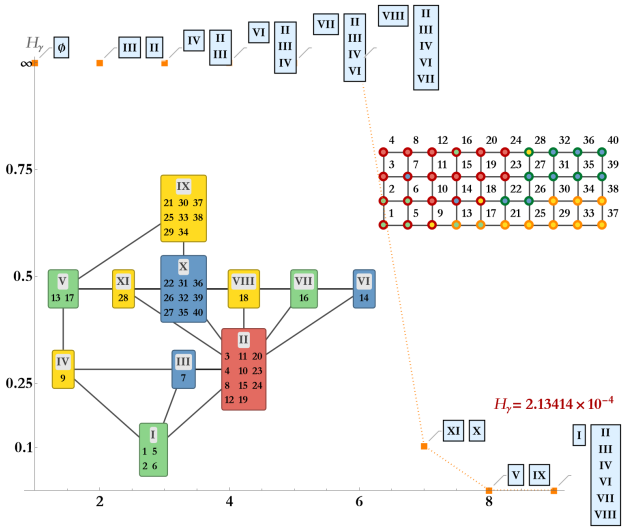
Next the communities are merged to decrease the entropy. Iteratively, the community p with the largest chromatic entropy is selected from P, and its neighborhood N of distance δ is computed. The algorithm evaluates whether merging p to a neighbor node will minimize the chromatic entropy. p is finally merged with neighbor q that minimizes the chromatic entropy the most. The node-communities located in the shortest path from p to q are also merged in order to fulfill the connectedness property within the new resulting community. Once the assembly of nodes is achieved they will now form a new community-node corresponding to their union.

The quotient graph is then updated by replacing the merged nodes by this new node-community and by updating the quotient graph. The process ends when no merges can decrease the entropy.

Let $\langle G, E, c\rangle$ be a colored graph, the complexity of the first phase is in $\mathcal{O}(|E|)$ since all nodes are visited from neighborhood to neighborhood to merge them into monochrome communities. Now considering the worst case for monochrome community reduced to a set of node singletons because the colors of all nodes are different, and assuming that at each step the new community merges only two communities, we deduce that the complexity is in $\mathcal{O}(|V|^2(|E| + |V|\log(|V|)))$.

Figure 4 shows the evolution steps of the algorithm. First the monochrome community quotient graph is defined. Let us remark that two connected node-communities have necessary a different color. Next the algorithm starts by merging the reduced single-node communities into larger communities because their chromartic entropy is ∞ constituting the greatest possible value. After, the communities are grouped together for forming larger communities decreasing the chromatic entropy until finally reaching 0.000025.

Figure 5 shows the computation steps on a larger example where the curve describes the chromatic entropy progression by indicating the community assemblies at each step. The grid graph was chosen because it provides a clean presentation of the final communities on the graph but the algorithm can be applied to any graph.



Bottom left: initial quotient graph of monochrome communities; upper right: the final colored graph where each final community contains the vertices with the same color border; at left of each step point of the curve: the communities to be merged, a column of roman numbers indicates a previously merged community.
PARAMETERS: $n = 40, \delta = 1, r = 4, p_\gamma$.

Figure 5: Chromatic community structure computation.

5 CHROCODE evaluation

CHROCODE will be analyzed with regard to three network topologies: small world, scale free, and Erdös Reny. the exploitation of these different network topologies allows us the assessment of their respective influence on the performance of the algorithm.

5.1 Probability Law Analysis

The choice of the probability law p_κ or p_γ seemingly alters the community structure obtained by CHROCODE algorithm. How significant can this difference be? This issue is crucial because the computational time between p_κ or p_γ could drastically differ. The complexity of p_κ is in $\mathcal{O}(rn)$ while the complexity of p_γ

depends on the cardinality of the DSS $\mathcal{S}$, in $\mathcal{O}(|\mathcal{S}|rn)$. Figure 6 shows the evolution of the cardinality of the DSS by choosing optimally the parameters r, n, d to maximize its growth. Notice that the optimal r is $r = n - d + 1$.

The size of the DSS grows exponentially when n increases by selecting the optimal parameters r, d (see Figure 6.1). However it is also worth noting that this growth is limited if r remains small ($\lesssim 10$) (Figure 6.3) which is often the case in practice.

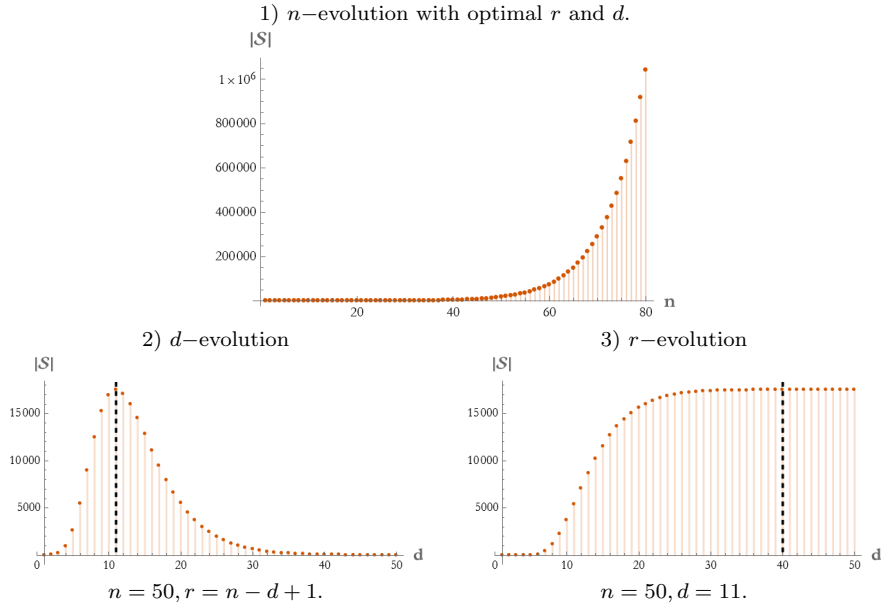


Figure 6: Evolution of the cardinality of the DSS $\mathcal{S}$.

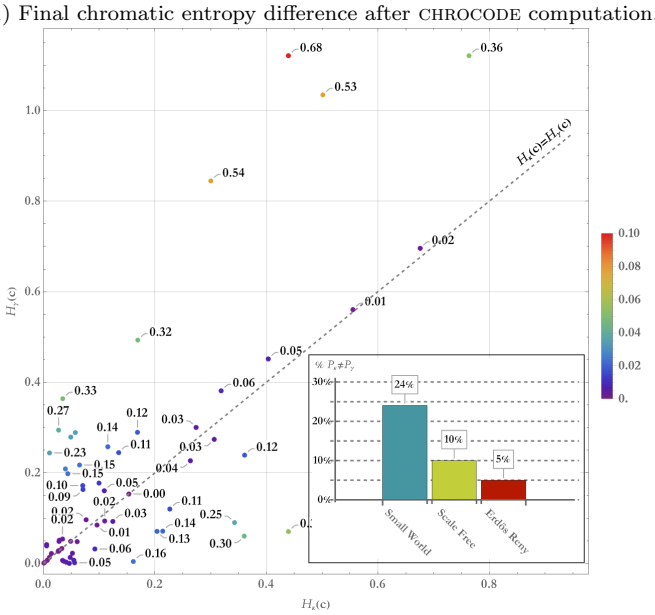
5.2 Impact of the probability laws on chrocode

When the computation of the DSS becomes intractable due to its size, we do wonder know if we can validly use p_κ instead of p_γ. To answer to this question we compare the entropy of the community structure computed by CHROCODE using respectively the probability laws p_κ and p_γ as input ((Figure 7).

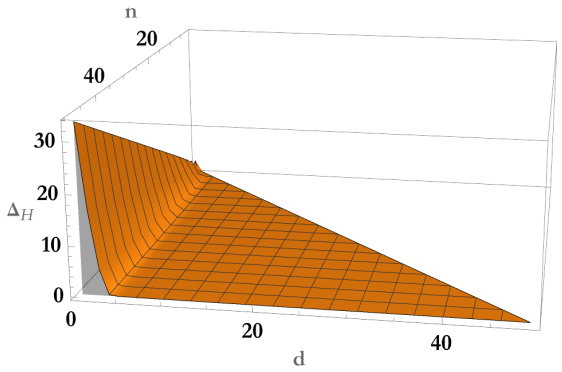
For each topology we generate 10 networks by increasing their size from 10 to 100 by 10. Therefore for each topology 100 networks are produced (300 networks in total). For the benchmark, we use 4 colors ($r = 4$) and a neighborhood of distance 2 ($\delta = 2$). Based on these networks, we isolate the cases where the resulting community structure computed by CHROCODE differs using p_κ or p_γ possibly leading to two distinct community structures: $P_\kappa = \text{CHROCODE}(G, c, \delta, p_\kappa)$ and $P_\gamma = \text{CHROCODE}(G, c, \delta, p_\gamma)$ with $P_\kappa \neq P_\gamma$.

The percent of networks where the community structures differ closely depends on the topology. The small world topology induces more differences than

1) Final chromatic entropy difference after CHROCODE computation.



2) Brute chromatic entropy difference



1) The analysis is achieved on 49 different cases ($r = 4, \delta = 2$).
2) 1275 differences are computed since d never exceeds n. For the mean difference computation (at right), the zero values are removed unless only this value exists. The error bar represents the standard error.

Figure 7: Difference between p_κ and p_γ

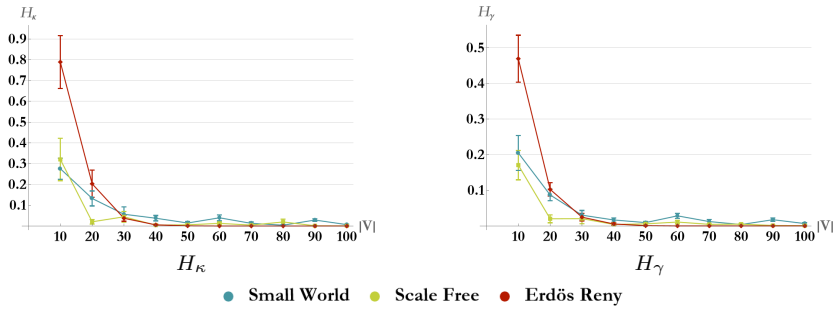
the other topologies. We also determine the entropy distance on the resulting community structures with the same initial graph: $\delta_H = |H(P_\kappa) - H(P_\gamma)|$. Even a community structure is computed from one probability law, but we also calculate the entropy with both probability laws. Figure 7.1 describes the observed distances. This distance never exceeds 1. with the tested networks whatever the topologies. The difference between the chosen probability laws thus depends on network topology and seemingly remains moderate on the tested cases.

Moreover, we also have evaluated the chromatic entropy distance $\Delta_K(c) = |H_\kappa(c) - H_\gamma(c)|$ by making n, d varying with the optimal parameter for $r = n - d + 1$ (Figure 7.2). This evaluation is focused on the difference between the entropies using the two probability laws ($\Delta_H(c) = |H_\gamma(c) - H_\kappa(c)|$), providing a complementary approach to the previous one. This evaluation shows that the difference is significant when d is small. If $d > 3$ this difference is in the order of 10^{-1} and if $d > 5$ the difference is negligible in the order of 10^{-6}. The decrease of Δ_H is exponential and lower than 1 when $d > 1$ whatever n.

In conclusion, from these two evaluations, the quality of the community structures appears almost equivalent whatever the probability law used. Hence, they can be somehow considered practically similar although their definition differ.

5.3 Network size sensitivity

The efficiency of the algorithm is sensitive to the network topology and its size. Figure 8 clearly shows that the entropy decreases when the size increases. The topology does not seem to affect the result significantly when the number of nodes exceeds 30 for the tested networks. The curves for both probability laws are similar because a real difference between them occur only when CHROCODE provides different community structures according to the used probability law. We have previously shown that only few cases induce a difference that remains low. (Figure 7).



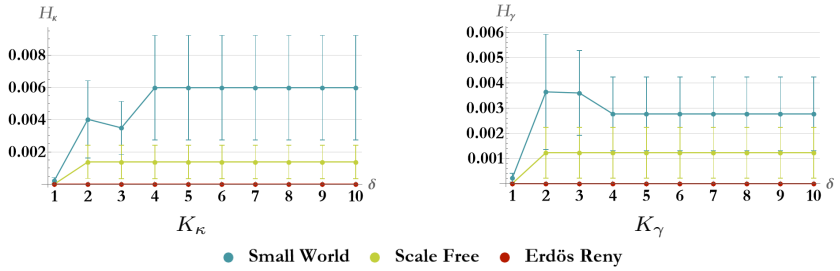
The curves correspond to the mean of the entropy on 10 trials and the error-bars describe the standard error. ($r = 4, \delta = 2$).

Figure 8: Evolution of the chromatic entropy w.r.t. the vertex size.

5.4 Neighborhood distance

The variation of the distance of the neighborhood (δ) also affects the final result. Figure 9 shows the consequence of the variation of the neighborhood distance from 1 to 10 on the chromarities with 10 networks of size 100 for each value of δ on different topologies. for each trial CHROCODE was computed twice using both chromarities and the community structure with the smallest chromaty is kept when the result differs. We can observe that the optimal distance δ differs on the topologies. The optimal δ is 1 whatever the topologies, and the variation of δ does not significantly affect the result for Erdös Reny topology. It is also worth noticing that the variation is always stabilized after $\delta \geq 4$ for all network trials.

Therefore, a possible improvement of the algorithm is to perform 4 tests by varying δ from 1 to 4 and to keep the structure with the lowest entropy among these tests.



The diagram reports the mean of the entropy on 10 trials of networks of size 100 and the error-bars describe the standard error. A different network is generated for each trial. ($r = 4, |V| = 100$).

Figure 9: Impact of δ on CHROCODE result.

6 Conclusion

We propose a new approach to detecting communities that relies on new criteria to identify them. Instead of a difference in connection density between its interior and its border, defining a community will minimize the chromatic entropy which is the entropy measure adapted to the problem of gathering nodes with the same colors.

This new paradigm provides an alternative approach to connectivity rule. It takes on its full meaning in challenges where the connection of nodes sharing the same property remains loose and therefore cannot be captured by an examination of the connection density as has been shown for disease modules.

This clustering criterion finds its application in problems where the community organization is essentially based on the aggregation of nodes sharing the same property without apparent correlation with the law of connectivity.It is based on the idea that a relevant organization is opposed to a design by chance.

We characterize two probability laws for defining the probability to generate a community by chance: the difference between them lies on the fact that we consider whether the color is dominant.

The significance of a community is assessed by its entropy. Low entropy means that the community structure cannot have been the result of chance confirming its organizational relevance. We have proposed a CHROCODE heuristic solving this problem in polynomial time. The tests analyzing the performance of this algorithm highlight the proximity of the two probability laws as well as a very good performance of the algorithm.

A first perspective would be to improve the algorithm by refining the heuristic criteria for better aggregating communities notably based on the topology of the graph. Another perspective would be to study how the grouping of nodes according to the major color rule could also integrate connectivity between nodes of the same color. Indeed, sharing the same property, these nodes could develop a particular connectivity structure characterizing a connection pattern that can be specific to the shared property. Such a perspective would allow recognition of a property-dependent community through an hybrid model, combining the identical property recognition with connectivity rules for detecting communities.

Appendix

Proofs

Proof of Theorem 1. To combinatorially specify the cardinality of $|\bigcup_{k=1}^{r} D_k|$, we need to determine the cardinality of all the intersection sets $|\bigcap_{k=l}^{m} D_k|, 1 \leq m \leq r$. We illustrate the characterization of the formulas on 3 sets before generalizing it to ease the explanation. The main issue is to formulate the cardinality of any intersection of set by a combinatorial formula.

Basically $D_k = \{p \mid \exists k \in c : \chi_{c_p}(k) = d\}$ is a set of communities with d vertices of color k. The different possible selections of d vertices among n is given by $\binom{n}{d}$. The count of the rest of the profile once the color k is assigned to vertices equals $(r-1)^{n-d}$, leading to the following combinatorial characterization of $|D_k|$:

$$|D_k| = \binom{n}{d}(r-1)^{n-d}.$$

Let us remark that this formula can be applied for all the colors and the number of possible used color is $\binom{r}{1}$ (which is 3 for $r = 3$). We deduce that $|D_1| + |D_2| + |D_3| = \binom{r}{1}\binom{n}{d}(r-1)^{n-d} = 3\binom{n}{d}2^{n-d}$.

By extension, for the intersection of two sets $D_k \cap D_j = \{p \mid \chi_{c_p}(k) = d \wedge \chi_{c_p}(j) = d\}$ the cardinal is defined by first considering the selection of a subset of size d for color k and next the selection of a size d vertices of color j in the remaining $n - d$ vertices. The following combinatorial formula formalizes these two steps of vertices selection.

$$|D_k \cap D_j| = \binom{n}{d}\binom{n-d}{d}(r-2)^{n-2d}.$$

Similarly, the number of possible color pairs is given by $\binom{r}{2}$ (which is 3 for $r = 3$). Then, we conclude that:

$$D_1 \cap D_2| + |D_1 \cap D_3| + |D_2 \cap D_3| = \binom{r}{2}\binom{n}{d}\binom{n-d}{d}(r-2)^{n-2d} = 3\binom{n}{d}\binom{n-d}{d}.$$

The same reasoning can be applied for the cardinal of the intersection of the three sets $|D_1 \cap D_2 \cap D_3|$ and more generally for any intersection.

$$|D_1 \cap D_2 \cap D_3| = \binom{r}{3}\binom{n}{d}\binom{n-2d}{d}\binom{n-d}{d}(r-3)^{n-3d}$$

The formula holds under considering that $0^0 = 1$ since $(r-3)^{n-3d} = 0^{n-3d}$ which must not be null or undefined when $n = 3d$ *e. g.,* for $r = 3, n = 6, d = 2$ we have $|D_1 \cap D_2 \cap D_3| = 90$. The formula defining the cardinal of the union of the 3 sets is finally:

$$|D_1 \cup D_2 \cup D_3| = \binom{r}{1}\binom{n}{d}(r-1)^{n-d} - \binom{r}{2}\binom{n}{d}\binom{n-d}{d}(r-2)^{n-2d} + \binom{r}{3}\binom{n}{d}\binom{n-d}{d}\binom{n-2d}{d}(r-3)^{n-3d}.$$

The generalization to any number of colors based on the Poincaré sieve finally leads to:

$$\left|\bigcup_{k=1}^{r} D_k\right| = \sum_{k=1}^{r}(-1)^{k-1}\binom{r}{k}\left(\prod_{i=0}^{k-1}\binom{n-id}{d}\right)(r-k)^{n-kd}.$$

By simplification of the product and by considering that the product is null if $rd > n$ the number of $d-$ colorful communities of size n with r colors is finally given by κ function.

$$\kappa(r,n,d) = \sum_{k=1}^{\min\left(r,\lfloor\frac{n}{d}\rfloor\right)} \frac{(-1)^{k-1}\binom{r}{k}n!(r-k)^{n-kd}}{(n-kd)!(d!)^k} \quad \square$$

□

Proof of Theorem 2. The enumeration of $d-$dominant communities is based on DSS by applying the formula counting the number of permutations with repetition. Considering m distributed on $n > m$ positions having each $k_i, 1 \leq i \leq m$ repetitions, the n elements having each k_i repetitions such that $\sum_{i=1}^{m} k_i = n$, let us recall that the number of permutation with repetition is:

$$\frac{n!}{\prod_{i=1}^{m} k_i!}$$

Indeed, first we count the number of color profiles of size n for a specific chromatic function. Since the vertices of the same color cannot be distinguished in a community, the number of communities having the same chromatic function corresponds to the number of permutations where the vertices of the same color are repeated, that is:

$$\frac{n!}{\prod_{k=1}^{r} \chi(k)!}$$

Notice that we can similarly define it using the signature σ by $\frac{n!}{\prod_{k=1}^{r} \sigma(k)!}$. This formula using the signature σ will be used in the sequel.

Next we need to enumerate all the chromatic function related to a signature. Let us remark that the number of chromatic function with the same signature

is also obtained by the application of a permutation with repetition. The repetition comes from the possible equality of the number of vertices for distinct colors, thus decreasing the number of different chromatic functions. For example in Figure 3, the number of chromatic functions for $(1, 1, 3)$ is 3 while it is 6 for $(0, 2, 3)$ because there is a repetition of the number 1 in the first and none in the second. Therefore, the formula counting chromatic functions taking into account the equal number of occurrences for different colors is:

$$\frac{r!}{\prod_{s \in \mathbf{Img} \circ \mathbf{Count}(\sigma)} s!}$$

Finally, the number of communities associated to a signature is then the product of these two formulas, leading to:

$$\frac{n!r!}{\prod_{s \in \mathbf{Img} \circ \mathbf{Count}(\sigma)} s! \prod_{j=1}^{r} \sigma(j)!}$$

This formula counts the number of communities having the same signature. The total number of communities is the sum of this count for all signatures. Let $\mathcal{S}_{r,n,d}$ be the DSS according to parameters r for the number of colors, n for the community size and d for the maximal number of vertices of the same colors, the formula counting the dominant communities is finally given by γ function:

$$\gamma(r, n, d) = n!r! \sum_{\sigma \in \mathcal{S}_{r,n,d}} \frac{1}{\prod_{s \in \mathbf{Img} \circ \mathbf{Count}(\sigma)} s! \prod_{i=1}^{r} \sigma(i)!}. \quad \square$$

$\square$

CHROCODE Algorithm

The main variables and used functions are:

$\textsc{Vget}(V)$	gets a vertex in V (randomly).
$\textsc{Path}(G)$	set of pathes of graph G.
$\textsc{ShortestPath}(E, p, q)$	finds a shortest path between p, q.
c	function giving the color of a node
H_ω	chromatic entropy with a probability law ω as parameter
P	community structure.
N	neighborhood of radius δ.
up	Boolean variable determining whether P must be updated.

function $\textsc{ChroCoDe}(\langle V, E, c\rangle$: colored graph, r: number of colors,δ: radius,ω: probability)
  $W \leftarrow V$;
  while $W \neq \emptyset$ **do** ▷ Generate the monochrome communities.
    $v \leftarrow \textsc{Vget}(W)$; $W \leftarrow W \setminus \{v\}$;
    $p_{\text{new}} = \{v\}$;$p = \emptyset$;
    while $p_{\text{new}} \neq \emptyset$ **do**
      $w \leftarrow \textsc{Vget}(p_{\text{new}})$; $p_{\text{new}} \leftarrow p_{\text{new}} \setminus \{w\}$;
      $p \leftarrow p \cup \{w\}$;
      $N \leftarrow \{w' \mid (w, w') \in E \wedge c(w') = c(v) \wedge w' \notin p\}$;
      $p_{\text{new}} \leftarrow p_{\text{new}} \cup N$;
    end while
    $W \leftarrow W \setminus p$;
    $P \leftarrow P \cup \{p\}$;
  end while ▷ $\langle P, E_P\rangle$ is the quotient graph.
  $E_P \leftarrow \{(p, p') \mid \exists v \in p, \exists v' \in p' : (v, v') \in E, p, p' \in P\}$;
  $W \leftarrow P$;
  while $W \neq \emptyset$ **do** ▷ Assemble the communities.
    $p \leftarrow \arg\max \{H_\omega(q, r) \mid q \in W\}$;
    $W \leftarrow W \setminus \{p\}$;
    $N \leftarrow \{q \mid q \in E_P \wedge 1 < \textsc{ShortestPath}(E_P, p, q) \leq \delta\}$;
    $h_{\min} \leftarrow H_\omega(P, c, r)$; $up \leftarrow False$;
    for $q \in N$ **do** ▷ Find the merging of communities minimizing H_ω in N.
      $SP \leftarrow \textsc{ShortestPath}(E_P, p, q)$;
      $h \leftarrow H_\omega((P \setminus SP) \cup \left\{\bigcup_{p_i \in SP} p_i\right\}, c, r)$;
      if $h_{\min} \geq h$ **then**
        $up \leftarrow True$;
        $h_{\min} \leftarrow h$; $SP_{\min} \leftarrow SP$;
      end if
    end for
    if up **then**
      $p = \bigcup_{p_i \in SP_{\min}} p_i$; ▷ Merge the communities of the path.
      $P \leftarrow (P \setminus SP_{\min}) \cup \{p\}$; ▷ Update P
      $E_P \leftarrow \{(p, p') \mid \exists v \in p, \exists v' \in p' : (v, v') \in E, p, p' \in P\}$; ▷ quotient graph rebuilt
      $W \leftarrow P$;
    end if
  end while
  return P;
end function

Algorithm computing the Dominant set of signatures (DSS)

We define . is the concatenation operator between two vectors, and a vector is written $(x_1, \cdots, x_m)$.

function SUBSIG(r, n, d, σ)
  var : s
  if $n = 0$ **then**
    $s \leftarrow \{\overbrace{(0, \cdots, 0)}^{r}.\sigma\}$
  else if $d = 0$ **then**
    $s \leftarrow \emptyset$
  else if $r = 1 \wedge n \leq d$ **then**
    $s \leftarrow \{(n).\sigma\}$
  else if $r = 0$ **then**
    if $n = 0$ **then**
      $s \leftarrow \{\sigma\}$
    else
      $s \leftarrow \emptyset$
    end if
  else
    $s \leftarrow \emptyset$
    for $d' \leftarrow \lceil \frac{n}{r} \rceil$ **to** $\min(d, r)$ **do**
      $s \leftarrow s \cup$ SUBSIG$(r - 1, n - d', d', (d').\sigma)$
    end for
  end if
  return s
end function

function FINDALLSIGS(r, n, d)
  var : S
  if $r = 0$ **then**
    if $n = 0$ **then**
      $S \leftarrow \{\emptyset\}$ ▷ a solution exists but empty
    else
      $S \leftarrow \emptyset$ ▷ No solutions
    end if
  else
    $S \leftarrow$ SUBSIG$(r - 1, n - d, d, (d))$
  end if
  return S
end function

References

[1] C. E. Shannon. "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27.3 (July 1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[2] Ada Hamosh et al. "Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders". In: *Nucleic Acids Research* 33 (suppl_1 Jan. 1, 2005), pp. D514–D517. DOI: 10.1093/nar/gki033.

[3] Pascal Pons and Matthieu Latapy. "Computing Communities in Large Networks Using Random Walks (Long Version)". Dec. 12, 2005.

[4] U. Brandes et al. *Maximizing Modularity Is Hard*. Aug. 30, 2006. DOI: 10.48550/arXiv.physics/0608255. URL: http://arxiv.org/abs/physics/0608255 (visited on 01/16/2023). preprint.

[5] M. E. J. Newman. "Modularity and Community Structure in Networks". In: *Proceedings of the National Academy of Sciences* 103.23 (June 6, 2006), pp. 8577–8582. DOI: 10.1073/pnas.0601602103.

[6] M. Oti et al. "Predicting Disease Genes Using Protein-Protein Interactions". In: *Journal of Medical Genetics* 43.8 (Aug. 2006), pp. 691–698. DOI: 10.1136/jmg.2006.041376.

[7] Vincent D. Blondel et al. "Fast Unfolding of Communities in Large Networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 9, 2008), P10008. DOI: 10.1088/1742-5468/2008/10/P10008.

[8] Santo Fortunato. "Community Detection in Graphs". In: *Physics Reports* 486.3-5 (Feb. 2010), pp. 75–174. DOI: 10.1016/j.physrep.2009.11.002.

[9] Tsuyoshi Murata. "Detecting Communities in Social Networks". In: *Handbook of Social Network Technologies and Applications*. Ed. by Borko Furht. Boston, MA: Springer US, 2010, pp. 269–280. DOI: 10.1007/978-1-4419-7142-5_12.

[10] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. "Network Medicine: A Network-Based Approach to Human Disease". In: *Nature reviews genetics* 12.1 (2011), pp. 56–68.

[11] Laura I. Furlong. "Human Diseases through the Lens of Network Biology". In: *Trends in Genetics* 29.3 (Mar. 2013), pp. 150–159. DOI: 10.1016/j.tig.2012.11.004.

[12] Annick Lesne. "Shannon Entropy: A Rigorous Notion at the Crossroads between Probability, Information Theory, Dynamical Systems and Statistical Physics". In: *Mathematical Structures in Computer Science* 24.3 (June 2014), e240311. DOI: 10.1017/S0960129512000783.

[13] Wei Liu, Matteo Pellegrini, and Xiaofan Wang. "Detecting Communities Based on Network Topology". In: *Scientific Reports* 4.1 (1 July 18, 2014), p. 5739. DOI: 10.1038/srep05739.

[14] Joanna S. Amberger et al. "OMIM.Org: Online Mendelian Inheritance in Man (OMIM®), an Online Catalog of Human Genes and Genetic Disorders". In: *Nucleic Acids Research* 43 (Database issue Jan. 2015), pp. D789–798. DOI: 10.1093/nar/gku1205.

[15] Susan Dina Ghiassian, Jörg Menche, and Albert-László Barabási. "A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome". In: *PLOS Computational Biology* 11.4 (Apr. 8, 2015), e1004120. DOI: 10.1371/journal.pcbi.1004120.

[16] Amitabh Sharma et al. "A Disease Module in the Interactome Explains Disease Heterogeneity, Drug Response and Captures Novel Pathways and Genes in Asthma". In: *Human Molecular Genetics* 24.11 (June 1, 2015), pp. 3005–3020. DOI: 10.1093/hmg/ddv001.

[17] Santo Fortunato and Darko Hric. "Community Detection in Networks: A User Guide". In: *Physics Reports*. Community Detection in Networks: A User Guide 659 (Nov. 11, 2016), pp. 1–44. DOI: 10.1016/j.physrep.2016.09.002.

[18] Bisma S. Khan and Muaz A. Niazi. *Network Community Detection: A Review and Visual Survey*. Aug. 2, 2017. DOI: 10.48550/arXiv.1708.00977. URL: http://arxiv.org/abs/1708.00977 (visited on 11/08/2022). preprint.

[19] Sonia Pavan et al. "Clinical Practice Guidelines for Rare Diseases: The Orphanet Database". In: *PLOS ONE* 12.1 (Jan. 18, 2017), e0170365. DOI: 10.1371/journal.pone.0170365.

[20] Yael Silberberg, Martin Kupiec, and Roded Sharan. "GLADIATOR: A Global Approach for Elucidating Disease Modules". In: *Genome Medicine* 9.1 (Dec. 2017), p. 48. DOI: 10.1186/s13073-017-0435-z.

[21] Rui-Sheng Wang and Joseph Loscalzo. "Network-Based Disease Module Discovery by a Novel Seed Connector Algorithm with Pathobiological Implications". In: *Journal of molecular biology* 430 (18 Pt A Sept. 14, 2018), pp. 2939–2950. DOI: 10.1016/j.jmb.2018.05.016.

[22] Franck Delaplace. *ChroCoS Library - Mathematica*. Version 2.01. Mar. 2023. DOI: 10.5281/zenodo.7767174. URL: https://doi.org/10.5281/zenodo.7767174.

[23] Franck Delaplace. *ChroCoS Module - Python*. Version 2.0. Mar. 2023. DOI: 10.5281/zenodo.7767111. URL: https://doi.org/10.5281/zenodo.7767111.