British Journal of Mathematics & Computer Science
4(16): 2344-2351, 2014

SCIENCEDOMAIN *international*
www.sciencedomain.org

Multi-clustering Gives Robustness to Modules in Networks

Alain Guénoche[1*]

[1] *Institute of Mathématics of Marseille, CNRS - Aix-Marseille Université, France.*

Original Research Article

Received: 29 March 2014
Accepted: 27 May 2014
Published: 21 June 2014

Abstract

Aims/ objectives: A protein-protein interaction network is considered as a simple indirected graph, weighted or non weighted. A partition of the vertex set, into connected, eventually overlapping, clusters having an edge density larger than the whole graph, is searched. Such a cluster is denoted as a *module*. The cellular functionality of proteins is predicted from this network decomposition. To improve the prediction quality, we need to evaluate the robustness of these modules.

Methodology: We propose a new method which consists in :

- selecting a non deterministic algorithm for graph partitioning into separate clusters (optimizing a modularity criterion);
- applying this algorithm several times to generate a set of close partitions;
- calculating a consensus partition from this set.

Results: This set of partitions permits to evaluate the robustness of any class as the average percentage of partitions joining any protein pair in this class. This robustness function can be applied to compare the consensus partition resulting of this procedure to the usually single partition computed from the graph.

Then, we develop a simulation protocol selecting random graphs having a more or less strong community structure. We show that the multi-clustering method provides modules closer to the communities which are more robust than those of a single partition.

Finally, we present a simple procedure to extend a strict partition into an overlapping class system, making multi assignment for proteins that could be placed equally into several modules, because their contributions to modularity are similar.

Keywords: Graph Partitioning; Consensus of Partitions; Robustness

2010 Mathematics Subject Classification: 53C25; 83C05; 57N16

**Corresponding author: E-mail: alain.guenoche@univ-amu.fr*

1 Introduction

Assessing the quality of partition's clusters, or the quality of a whole partition, appeared with the beginning of the clustering methodology and still remains an open subject. It has often been reduced to the seek for a partition that optimizes some criterion. However, these criteria are very diverse, and none of them can be retain to compare partitions computed according to different principles especially for graphs. Furthermore, they do not indicate the reliability of the computed classes and/or partition.

For undirected simple graphs, weighted or not, in which "natural classes", called communities exist, one can measure the cluster quality by edge density or by the percentage of intra/inter edges. More recently a *modularity criterion* has been largely adopted [1]. It is defined for any partition and it is based on modularity values for any pair of vertices. The maximum modularity value of a partition is high when the community structure is strong. It is an additive function of the cluster values and permits to detect those which have a large contribution and so a good quality.

To quantify the class robustness, we adopt the following strategy. Given a weighted graph $G = (V, E, A)$, using a non deterministic algorithm, we build a series of q partitions $(P_i)_{i=1,...q}$ of the vertex set V, making a *profile*. The first one, usually the single computed partition, is denoted P_{ini} in the following. From this profile, we calculate a consensus partition P_{cons}. Thus, one can measure for any pair of joined vertices the percentage of partitions in the profile joining them. Some robustness coefficients for classes and partitions can be deduced easily.

This *multi-clustering* approach, is not new. It follows a least two similar methods denoted *Bootstrap clustering* [2] and *Consensus clustering* [3]. The common part consists in generating several partitions of V. But this contribution differs from the first one, which needs to establish q graphs similar to G to realize the partition set, and from the second one which uses, as we do, other stochastic partitioning algorithms but never made the connection to robustness.

Our method needs to

- Select a fast partitioning algorithm, because it must be applied q times to graphs having several thousands of vertices. We have defined the TFR method, similar to TFit (Gambette & Guénoche, 2011), which determines the number of classes, not necessarily the same for each partition ;
- Use again our consensus of partitions method (Guénoche, 2010) which provides a *median* partition for the profile. In fact it the same algorithm as before, with a stochastic optimization final step.

A simulation protocol, based on random graphs having a graduated module structure, corresponding to a seed partition P_{seed}, permits to assess that the consensus partition P_{cons} is much closer to P_{seed} than P_{ini}. More, the average robustness of the P_{cons} classes is much higher than the P_{ini}'s one. These results permit to quantify the efficiency of the Multi-Clustering method. More, it allows to distinguish outliers that are vertices clustered irregularly in the q partitions, because they become singletons in the consensus partition.

The paper is organized as follows. In section 2, we recall the modularity formalism for unweighted and weighted graphs and we tackle the optimization problem introducing our non deterministic algorithm to establish a partition profile. In section 3, we come back to the consensus partition problem and define the robustness of clusters and partitions. A simulation protocol, to prove the efficiency of the multi-clustering procedure, is described in section 4. We add, in section 5, a new and simple algorithm to extend a strict partition in separate clusters into an overlapping class system, allowing to control the number of multi-assigned vertices.

2 Graph Partitioning

We don't want to examine here the graph partitioning problem with its huge diversity. For recent developments about the modularity optimization, we refer to Lancichinetti & Fortunato, 2012.

2.1 Modularity optimization

Let $G = (V, E)$ be a connected simple graph without loop, having $|V| = n$ vertices and $|E| = m$ edges. We want to detect modules in G and so build a partition $P = \{V_1, V_2, ..V_q\}$ with a high modularity value. For a partition, this criterion quantifies the difference between the proportion of internal edges in classes and this same quantity if there was no community and so edges were selected at random with the same degree distribution. It is the gap between what is observed in a given partition and what is expected by chance according to this null model. More formally, we refer to the Newman formula :

$$W(P) = \frac{1}{2m} \sum_{x=1}^{n} \sum_{y=1}^{n} (A_{xy} - \frac{d_x d_y}{2m}) \delta_{P(x)P(y)}, \tag{2.1}$$

where (A_{xy}) is the adjacency matrix of G, d_x is the degree of vertex x and δ is the usual Kronecker symbol ; $\delta_{P(x)P(y)}$ is the square matrix of order n such that

$$\delta_{P(x)P(y)} = \begin{cases} 1 & \text{if vertices } x \text{ and } y \text{ belong to the same class in } P, \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

The $M(P)$ modularity of partition P is proportional to the sum of the $w(x, y) = A_{xy} - \frac{d_x d_y}{2m}$ values on joined pairs in P. These are negative if $(x, y) \notin E$ otherwise positive when $d_x d_y < 2m$. The modularity function can be rewritten :

$$M(P) = \sum_{k=1}^{q} \sum_{x,y \in V_k} w(x, y). \tag{2.3}$$

The modularity of partition P is high when there are many edges within classes making a large density. To maximize the modularity function is a *Clique partitioning* problem for complete graph on X weighted by the positive or negative values w.

This formulation can be extended to graphs weighted by $A : E \to \mathbb{R}$. The adjacency matrix is now the matrix with $A(x, y) = 0$ iff $(x, y) \notin E$. In equation (2.1) degrees are replaced by the row or column sums of A ($s_x = \sum_{y|(x,y) \in E} A(x, y)$) and function

$$w(x, y) = A(x, y) - \frac{s_x s_y}{2m}$$

always defines the positive or negative weights of a complete graph.

2.2 Optimization problem

To maximize $M(P)$ over the set $\mathcal{P}_V$ of all the partitions on V is to build a set of separate cliques in $(\mathbf{K}_n, w)$ having a maximum sum of weight. It is the Zahn problem [4] for weighted graphs, well known to be NP-hard, and so no polynomial algorithm is known to give an optimal solution. Many authors adopt this formulation : Given a partition P, they pose $\alpha_{xy} = \delta_{P(x)P(y)}$, and M becomes

$$M(P) = \sum_{x<y} \alpha_{xy} w(x, y) \tag{2.4}$$

assuming P is an equivalence relation on V. The optimization problem is to find a matrix α maximizing M under constraints :

$$\begin{cases} \forall (x < y), \alpha_{xy} \in \{0, 1\} \\ \forall (x < y < z), \alpha_{xy} + \alpha_{yz} - \alpha_{xz} \leq 1. \end{cases}$$

It is a discrete linear programming NP-hard problem with $n(n-1)/2$ variables and $O(n^3)$ constraints. Optimal resolution methods exist establishing α, and so partition π, realizing the global maximum of function M over $\mathcal{P}_V$. But they cannot be applied to large problems. Recently [5], using column generation technics, have proved modularity optimality for graphs with 512 vertices. For much larger problems having several thousands vertices, heuristics must be used. We introduce the *Randomized Transfert-Fusion* (RTF) algorithm derived from our TFit method validated for Bootstrap Clustering.

2.3 Randomized Transfert-Fusion method (RTF)

It starts from the *atomic partition* of V, each vertex being a singleton. RTF is a method which iteratively applies the two following procedures :

- For the first one, called *Atomic Transfer*, the weight of the assignment of any vertex x to any class k is first computed. Let $K(x,k) = \sum_{y \in V_k} w(x,y)$ be this weight. If $x \in V_k$, $K(x,k)$ is the modularity contribution of x to its own class, and so to M. For any other class $V_{k'}$ it corresponds to the possible assignment of x to $V_{k'}$. The difference $K(x,k') - K(x,k)$ is the criterion variation after a transfer of x from class V_k to class $V_{k'}$.

  At each step, the transfer of a random vertex x is tested. If there is a gain (a positive variation) x is assigned to the class for which this gain is maximum. So, x is placed either in another class or creates a new supplementary class if its contribution to any class is negative. It that case, x becomes a singleton and provide a contribution equal to 0, increasing M. This procedure stops when after n consecutive unproductive trials, that is when no transfer has been made and M does not increase during n steps.

- The second one, called *Fusion* transforms G into its quotient graph according to the final partition at the end of the Atomic Transfer procedure. The new vertices are the classes, and the new weight for any pair of classes is equal to the sum of the $w(i,j)$ values of all the interclass element pairs. At this time begins a transfer procedure of the classes, following the same principle as before. Two random classes linked by a positive weight are merged, until all the interclass pairs have negative weights. This lead to partition $\pi = (V_1, \ldots, V_q)$ such that any partition π_{ij} joining classes V_i and V_j has a lower modularity score : $M(\pi_{ij}) < M(\pi)$.

This partition π is then proposed to the Atomic Transfer procedure. If there is no feasible transfer TFR stops ; otherwise, the modified classes are proposed to the Fusion Procedure.

TFit is the non randomized identical algorithm, because vertices are examined in the label order. It is very close to the *Méthode de Louvain* [6], except the Atomic Transfer procedure which is tested at each level. This latter takes time and so RTF is less efficient but it gives better partitions on classical benchmark graphs, after a few runs.

3 Consensus Partition

Former works on consensus partitions were motivated by clustering items described by nominal variables. In his pioneer paper, Régnier [7] introduced the notion of *partition centrale*, defined as the partition with minimum sum of distances to those in the profile. In other words, it is a *median* partition. Indeed it has been empirically assessed that other consensus definitions, more strict or formal, do not lead to satisfying practical results.

Given a profile $\Pi = (P_1, \ldots, P_q)$ of partitions over V, the *consensus partition problem* consists in finding $\pi \in \mathcal{P}$ minimizing the sum of symmetric difference distances to Π. Let T_{xy} be the number of partitions joining x and y in the same class. The score of a partition P relatively to profile Π is :

$$\begin{aligned} S_\Pi(P) &= \sum_{x<y} \Big(\alpha_{xy} T_{xy} + (1-\alpha_{xy})(q - T_{xy}) \Big) \\ &= 2 \sum_{x<y} \alpha_{xy} T_{xy} + \sum_{x<y} q - \sum_{x<y} \alpha_{xy} q - \sum_{x<y} T_{xy} \end{aligned}$$

Quantities $\sum_{x<y} q$ and $\sum_{x<y} T_{xy}$ only depend on the profile Π and not on P. Thus, maximizing $S_\Pi(P)$ is equivalent to maximize :

$$\sum_{x<y} \alpha_{xy} T_{xy} - \frac{1}{2} \sum_{x<y} \alpha_{xy} q.$$

Let $J(P)$be the set of joined pairs in P. An equivalent criterion to $S_\Pi(P)$ is :

$$W_\Pi(P) = \sum_{(x<y)\in J(P)} \Big(T_{xy} - \frac{q}{2}\Big). \tag{3.1}$$

Criterion W_Π can be intuitively understood as follows : for a partition P, a joined pair in $J(P)$ has a positive (resp. negative) contribution when both elements are joined in more (resp. less) than half the partition number in Π.

Let $\mathbf{K}_n$ be the complete graph on V, in which the pairs are weighted by $w : V \times V \to \mathbb{R}$, with $w(x,y) = T_{xy} - q/2$ and let P be a partition into p classes $P = (V_1, \dots, V_p)$. The quantity $W(V_k) = \sum_{(x,y)\in V_k} w(x,y)$ is the weight of all the pairs (a clique) in V_k. We have,

$$W_\Pi(P) = \sum_{k=1,..p} W(V_k) = \sum_{k=1,..p} \sum_{(x,y)\in V_k} \Big(T_{xy} - \frac{q}{2}\Big). \tag{3.2}$$

Thus the consensus partition problem, as it is detailed in (8), is also a *Clique partitioning* problem on the complete graph on V, weighted now by $w(x,y) = T_{xy} - \frac{q}{2}$. The weights $w(x,y)$ are positive or negative, according to the number of times x and y are joined. We will use the same algorithm (RTF) as before.

3.1 Robustness of Classes and Partitions

The score of a partition $W_\Pi(\pi)$ is defined as the sum of joined pair weights. So the score of a class is high when its pairs are frequently joined in the profile. One can evaluate the robustness of a class by the percentage of partitions in the profile joining its elements. As $T_{xy} = |\{P \in \Pi = \{P_1, \dots, P_q\}$ such that $P(x) = P(y)\}|$, we set :

$$Rob(V_k) = \frac{2\sum_{x,y\in V_k} T_{xy}}{q \times |V_k| \times (|V_k| - 1)}. \tag{3.3}$$

This quantity (between 0 and 1) is the average ratio of partitions joining pairs of elements in class V_k over its maximum number. So, one can compare classes using $Rob(V_k)$, the best ones containing only pairs often joined in the profile.

This definition can be extended to partitions. Their robustness is the average, over joined pairs (x,y), of the percentage of partitions joining them. Let us recall that $J(P)$ is the set of joined pairs in P. We obtain

$$Rob(P) = \frac{1}{q \times |J(P)|} \sum_{(x,y)\in J(P)} T_{xy}. \tag{3.4}$$

4 Simulation Protocol

We have developed a simulation protocol with unweighted random graphs made of 200 vertices distributed in 5 connected balanced communities defining a *seed partition*, P_{seed}. Each graph is generated by an Erdös-Reyni procedure with two parameters, the internal density (intra-class edges) d_i and the external density (inter-class edges) d_e. There are three families of graphs corresponding to densities $(d_i = .30, d_e = .10)$, $(d_i = .20, d_e = .05)$ and $(d_i = .10, d_e = .01)$. They generate more and more difficult problems, not because graph communities vanish, but the consensus classes do not fit the seed partition when the average degree decreases.

The RTF algorithm is applied to obtain an initial partition P_{ini} the first one and, to get a profile Π containing $q = 30$ partitions (a larger value has been tested and does not provide any improvements). Its consensus partition P_{cons}, is computed using RTF again, without applying any stochastic procedure. The two partitions, P_{ini} and P_{cons}, are compared to the seed partition P_{seed} by the way of the corrected Rand index (9) and also their robustness values, as previously defined. The results corresponding to 100 trials, that are 100 seed graphs with the same density values, are printed in Table 1.

		Rand		Robustness	
d_i	d_e	P_{ini}	P_{cons}	P_{ini}	P_{cons}
.30	.10	.825	.883	.883	.938
.20	.05	.689	.811	.745	.857
.10	.01	.615	.676	.715	.838

Table 1 - Corrected Rand index and robustness of initial and consensus partitions

According to the Rand index corrected by chance, problems are more and more difficult ; the initial partitions becomes far from seed partitions. But the consensus partitions are much closer to the seed ones. The robustness of the initial partition depends on the computed profile. Finally, P_{cons} always has a robustness value larger than P_{ini}, which can be expected by consensus definition.

What about the modularity value of the consensus partition and its number of classes ? Concerning modularity, we observe small variations, around 1%. Concerning the number of classes, counting clusters with at least 3 elements, the average number of such classes does not much vary and consensus partition isolates unstable vertices that are differently clustered along the profile. In P_{cons} they make singletons or very small classes.

5 From Strict Partitions to Overlapping Class System

In many practical problems of graph partitioning, a strict partition is not satisfying, because some vertices can belong to several classes. It is clear for social networks, as co-author groups in bibliographical lists or countries exchanging goods. It is the same for protein networks, an edge corresponding to a contact, revealing a functional biological complex. But proteins can be expressed in several tissues to make different complexes and so can belong to several classes.

5.1 Contribution of a vertex to its class and to the others

Given a partition $P = \{V_1, V_2, \ldots, V_q\}$ of the vertices of a graph $G = (V, E, A)$ in q classes, the contribution of vertex x to its class V_k has been denoted :

$$K(x,k) = \sum_{y \in V_k} w(x,y). \tag{5.1}$$

It is the sum of weights of the pairs containing x that are counted in the modularity values of class V_k, and also in $M(P)$. For any other class $V_{k'}$, this quantity corresponds to the possible assignment of x to $V_{k'}$. If partition P is computed optimizing modularity, each vertex is assigned to class V_k for which $K(x,k)$ is maximum.

5.2 Multiple assignments

Differences between classes can be small or null and vertex x could be assigned to class $V_{k'}$ if $K(x,k')$ is close to $K(x,k)$. Let κ be the class index not equal to k such that $K(x,k) - K(x,\kappa)$ is minimum. Vertex x can be assigned to class V_κ if

$$\tau(x) = \frac{K(x,k) - K(x,\kappa)}{K(x,k)}$$

is small. This is the relative gap to the second best class for x.

To fix a threshold σ for $\tau(x)$ is difficult and can generate a large number of multi-assigned vertices. Consequently, we choose to fix the rate of multi-assigned vertices, τ_m and to calculate the corresponding threshold σ. If there are 1000 vertices and if τ_m is fixed to 10% the threshold is equal to the 100-th largest value of $\tau(x)$ which defines σ.

Consequently, vertex x will be assigned to any class $V_{k'}$ such that

$$\frac{K(x,k) - K(x,k')}{K(x,k)} \leq \sigma. \tag{5.2}$$

6 Conclusions

a It is clear that the Multi-Clustering method improves the quality of the computed partitions, especially for graphs with a low rate of edges. Compared to the previous Bootstrap-Clustering method, results are similar. But Multi-Clustering allows to avoid the "graph like" definition and to fix parameter values, as the *elongation rate* or the *added edge rate*. In the average, the consensus partition is closer to the seed partition than the initial one, with a very close modularity value. In any case, classes have a better robustness value and outliers are isolated into singletons.

b The multi assignment procedure makes it possible to transform a strict partition of graph vertices into an overlapping class system. It is very efficient since the K contribution table of each vertex to each class is computed in $O(n^2)$. One of the major advantages of this procedure is the possibility to make the number of multi-assigned vertices vary, which is not possible with the OCG algorithm (10) and other methods to build overlapping classes in graphs.

.

Acknowledgment

This work has been made with a CNRS PEPS-BIM financial support.

Competing Interests

The author declares that no competing interests exist.

References

[1] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys. Rev. E. 2004;69:026133.

[2] Gambette Ph., Guénoche A. Bootstrap Clustering for graph partitioning, RAIRO. 2011;45(4):339-352.

[3] Lancichinetti A, Fortunato S. Consensus clustering in complex network. Scientific Reports. 2012;2:336. DOI:10.1038/srep00336.

[4] Zahn CT. Approximating symmetric relations by equivalence relations. SIAM J. on Appl. Math. 1964;12:840-847.

[5] Aloise D, Cafieri S, Caporossi G, Hansen P, Perron S, Liberti L. Column generation algorithms for exact modularity maximization in networks, Physical Review E. 2010;82:046112.

[6] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of community hierarchies in large networks, Journal of Statistical Mechanics: Theory and Experiment. 2008;P10008.

[7] Régnier S. Sur quelques aspects mathématiques des problèmes de classification automatique, Mathématiques et Sciences humaines. 1983;, **82**:13-29, reprint of I.C.C. bulletin. 1965;4:175-191. French.

[8] Guénoche A. Consensus of partitions : a constructive approach. Advances in Data Analysis and Classification. 2010;5(3):215-229.

[9] Hubert L, Arabie P. Comparing partitions. J. of Classification. 1985;2:193-218.

[10] Becker E, Robisson R, Chapple CE, Guénoche A, Brun C. Multifunctional Proteins Revealed by Overlapping Clustering in Protein Interaction Network. BioInformatics. 2012;28:1:84-90. DOI :10.1093.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
www.sciencedomain.org/review-history.php?iid=567&id=6&aid=5043