arXiv:1301.7006v6 [cs.SI] 25 Feb 2014

A Unified Community Detection, Visualization and Analysis method

Michel Crampes

Michel Plantié

Ecole des Mines d'Ales, Parc Georges Besse, 30035 Nîmes Cedex

With the widespread of social networks on the Internet, community detection in social graphs has recently become an important research domain. Interest was initially limited to unipartite graph inputs and partitioned community outputs. More recently bipartite graphs, directed graphs and overlapping communities have all been investigated. Few contributions however have encompassed all three types of graphs simultaneously. In this paper we present a method that unifies community detection for these three types of graphs while at the same time merges partitioned and overlapping communities. Moreover, the results are visualized in a way that allows for analysis and semantic interpretation. For validation purposes this method is experimented on some well-known simple benchmarks and then applied to real data: *photos and tags in Facebook and Human Brain Tractography data. This last application leads* to the possibility of applying community detection methods to other fields such as data analysis with original enhanced performances.

1. introduction

Thanks to the growth of online social networks, community detection has become an important field of research in computer sciences. Many algorithms have been proposed (see several surveys on this topic in [8, 28, 34, 29]). Most of them take unipartite graphs as inputs and produce partitioned communities. In unipartite graphs any node may share an edge with another node. Other contributions have also explored bipartite graphs and directed graphs. In bipartite graphs, nodes are separated in two sets and there are only edges between nodes of different sets. In directed graphs each link has a start node and an end node. These authors generally introduce community detection methods which are specific for each type of graphs, and sometimes for two types of graphs. In this paper we present a method that encompasses all three types of graphs simultaneously in a unique bipartite graph model.

With this respect we consider Newman's modularity [25] and apply it to bipartite graphs. We show that this modularity model can be directly applied to bipartite graphs with the side effect of structurally linking objects of both node sets in the same communities. This structural property is formally demonstrated in Annex 1.

In a second step this model is transformed into a unipartite graph model. As a result any community detection algorithm for unipartite graph may be applied.

2 *Crampes, Plantié*

We chose for experiments the so-called Louvain algorithm [3] which is known for its efficiency in producing partitioned communities from extensive data sets. It is also applicable to weighted and unweighted graphs. Our method extracts communities where both types of nodes are associated. We show that this result is semantically pertinent although it has been criticized by some authors [20, 13, 1] who think that there should not be the same number of communities in both sets. Moreover associating both types of nodes in the same communities opens up new issues. It is possible to merge partitioned and quantified overlapping communities in a unique view and then analyze their structure with different perspectives. Indeed most community detection algorithms such as Louvain use heuristics which lead to local optima. With our approach we can identify and explain the final organization and possibly correct some unwanted node assignments.

In the following we use the term "'semantics"' for qualifying entities which are described by properties or attributes. Community detection is driven by properties that are shared between entities and consequently the resulting communities are semantically described by these properties.

For validation and comparison with other authors the whole method has been experimented on small traditional unipartite and bipartite benchmarks. We have generated interesting insights which extend beyond known results. We can then apply our method on real medium-sized bipartite graphs, in a step that reveals significant properties such as overlapping communities, community compactness and the role of inter-community objects. These results are valuable when observed in data like people-photo data sets targeted by our experiments.

Beyond community detection, our method has also been applied to brain data extracted through 'tractography' by a team of neurologists and psycho-neurologists seeking to extract macro connections between different brain areas. Our results were compared with those they obtained when applying spectral clustering, a traditional data analysis method. Although they were very similar, our method provided new insights in the analysis. In conclusion we observe that after having borrowed algorithms from data analysis methods, community detection may in return offer new tools to these techniques. We also successfully applied our method to most standard unipartite and bipartite graph benchmarks.

The next section will present a state-of-the-art on community detection methods using different types of graphs. Section 3 will follow by focusing on a new method to unify all types of graphs; it uses a definition of modularity for bipartite graphs directly derived from modularity for unipartite graph which is presented in Annex 1 (section 8). Section 4 will then demonstrate how our unifying method is particularly valuable in computing, visualizing and analyzing partitioned and overlapping communities. Section 5 presents several practical results on different types of graph data sets. The conclusion in section 7 discusses the pros and cons of our method in the light of these experimental results.

2. State of the art

As stated above, several state-of-the-art assessments have already addressed the community detection problem: [28, 29, 34, 8]. They are mainly focused on unipartite graph partitioning. The calculation performed is based on maximizing a mathematical criterion, in most cases modularity [25], representing the maximum number of connections within each community and a minimum number of links with external communities. Various methods have been developed to identify the optimum, e.g. greedy algorithms [23, 26], spectral analysis [24], or a search for the most centric edges [25]. One of the most efficient greedy algorithm for extracting partitioned communities from large (and possibly weighted) graphs is Louvain [3]. In a very comprehensive state-of-the-art report [8] other new partitioned community detection methods are described.

The partitioning of communities, despite being mathematically attractive, is not satisfactory to describe reality. Each individual has 'several lives' and usually belongs to several communities based on family, professional, and other activities. As such other methods more recently take into account the possibility for overlapping communities. The so-called k-clique percolation method [27] detects overlapping communities by allowing nodes to belong to multiple k-cliques. A more recent method adapted to bipartite networks, and based on an extension of the k-clique community detection algorithm is presented in [31]. Several methods use local fitness optimization [16][14]. The 'Label Propagation Algorithms' (LPA) are reported to be particularly efficient [12]. [16] uses a greedy clique expansion method to determine overlapping communities via a two-step process: identify separated cliques and expand them for overlapping by means of optimizing a local fitness criteria. [7] derives n order clique graphs from unipartite graphs to produce partitioned and overlapping communities using Louvain algorithm. Some research has provided results in the form of hypergraph communities such as in[6, 5]. Other methods are found in scientific papers, yet most of these are prone to major problems due to computational complexity. More recently Wu [33] proposed a fast overlapping community detection method for large real-world unipartite networks. The method in [7] presents some common features with ours, albeit with a different strategy, since it uses traditional partitioning algorithm to extract overlapping communities.

When considering semantics it becomes necessary to focus on bipartite or "multipartite" graphs i.e. graphs whose nodes are divided into several subsets, and whose edges only link nodes from different subsets. One example of this type of graph is the set of photos from a Facebook account along with their 'tags' [19] or else the tripartite network of epistemic graphs [30] linking researchers, their publications and keywords in these publications. Traditional methods transform the multipartite graph into a unipartite graph by assigning a link between two nodes should they share a common property. In doing so however semantics is lost. Hence many researchers retain the multiparty graph properties by extending the notion of modularity to these types of graphs and then apply algorithms originally designed for

unipartite graphs [32, 22, 1, 20, 7][18].

3. Unifying bipartite, directed and unipartite graphs

3.1. *Bipartite graphs partitioning*

3.1.1. *Turning bipartite graphs into unipartite graphs.*

In formal terms, a bipartite graph $G = (U, V, E)$ is a graph $G' = (N, E)$ where node set N is the union of two independent sets U and V and moreover the edges only connect pairs of vertices (u, v) where u belongs to U and v belongs to V.

$N = U \cup V$, $U \cap V = \emptyset$, $E \subseteq U \times V$.

Let $r = |U|$ and $s = |V|$, then $|N| = n = r + s$

The unweighted biadjacency matrix of a bipartite graph $G = (U, V, E)$ is a $r \times s$ matrix B in which $B_{i,j} = 1$ *iff* $(u_i, v_j) \in E$ and $B_{i,j} = 0$ *iff* $(u_i, v_j) \notin E$.

It must be pointed out that the row margins in B represent the degrees of nodes u_i while the columns' margins represent the degrees of nodes v_j. Conversely, in B^t, the transpose of B, row's margins represent the degrees of nodes v_j and columns' margins represent the degrees of nodes u_i. Let's now define the off-diagonal block square matrix A' :

$A' = \begin{pmatrix} 0_r & B \\ B^t & 0_s \end{pmatrix}$ where 0_r is an all zero square matrix of order r and 0_s is an all zero square matrix of order s.

This symmetric matrix is the adjacency matrix of the unipartite graph G' where nodes' types are not distinguished. It is possible to apply to G' any algorithm for extracting communities from unipartite graphs. A' is also the off-diagonal adjacency matrix of bipartite graph G. Consequently the communities which are detected in G' are also detected in G. The question is to determine the validity of this side effect result: what is the quality of partitioning for G when applying an unipartite graph partitioning algorithm on G'? Barber [1] and Liu/Murata [18] have also introduced the block matrix as a way of detecting communities in bipartite graphs. However we see below that they do not take all consequences of this approach.

3.1.2. *Extending modularity to bipartite graphs*

Modularity is an indicator often used to measure the quality of graph partitions [25]. First defined for unipartite graphs, several modularity variants have been proposed for bipartite graph partitioning and overlapping communities. More recently several authors introduced modularity into bipartite graphs using a probabilistic analogy with the modularity for unipartite graph which will be discussed below. However when applying unipartite graph modularity optimization algorithms to bipartite graphs, it is another expression of probabilistic modularity presented hereafter.

Let $G = (U, V, E)$ be a bipartite graph with its biadjacency matrix B and the unipartite graph G' with the adjacency off-diagonal block matrix A'. Let's consider Newman's modularity [25] for this graph G'. It is a function Q of both matrix A'

and the communities detected in G' :

$$Q = \frac{1}{2m} \sum_{i,j} \left[A'_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{1}$$

where A'_{ij} represents the weight of the edge between i and j, $k_i = \sum_j A'_{ij}$ is the sum of the weights of the edges attached to vertex i, c_i denotes the community to which vertex i is assigned, the Kronecker's function $\delta(u, v)$ equals 1 if $u = v$ and 0 otherwise and $m = 1/2 \sum_{ij} A'_{ij}$. Hereafter we only consider binary graphs and weights are equal to 0 or 1.

After several transformations we show (see Annex 1, Section 8) that this modularity can also be written using the biadjacency matrix B of the bipartite graph $G = (U, V, E)$:

$$Q^B = \frac{1}{m} \sum_{ij} [B_{ij} - \frac{(k_i + k_j)^2}{4m}] \delta(c_i, c_j) \tag{2}$$

where k_i is the margin of row i in B, k_j the margin of column j in B and $m = \sum_{ij} B_{ij} = \frac{1}{2} \times \sum_{ij} A'_{ij} = m$ in (1).

Another interesting formulation to be used is the following (Appendix 1, Section 8):

$$Q^B = \sum_c [\frac{|e_c|}{m} - (\frac{(d_{u|c} + d_{v|c})}{2 \times m})^2] \tag{3}$$

where $|e_c|$ is the number of edges in community c, and $d_{w|c}$ is the degree of node w belonging to c.

This formulation of modularity is the same as Newman's modularity with more detailed information: it explicitly shows that both sets of nodes are structurally associated in the same communities.

Since in the general case B is not symmetric, this definition thus characterizes modularity for bipartite graphs after their extension into unipartite graphs. It then becomes possible to apply any partitioning algorithm for unipartite graphs to matrix A' and obtain a result where both types of nodes are bound in the same communities, except in the case of singletons (i.e. nodes without edges). This definition from unipartite graph modularity given that it is able to bind both types of nodes, is compared in Section 3.2 with other authors' modularity models for bipartite graphs.

3.1.3. *Turning oriented graphs into bipartite graphs.*

A directed graph is of the form $G^d = (N, E^d)$ where N is a set of nodes and E^d is a set of ordered pairs of nodes belonging to N : $E^d \subseteq N \times N$. From the model in (1)

6   *Crampes, Plantié*

Leicht [17] use probabilistic reasoning 'insights' to derive the following modularity for directed network:

$$Q = \frac{1}{m} \sum_{ij} \left[A_{ij} - \frac{k_i^{in} k_j^{out}}{m} \right] \delta(c_i, c_j) \tag{4}$$

where k_i^{in} and k_j^{out} are the in - and out- degrees of vertices i and j, A is the asymmetric adjacency matrix, and $m = \sum_{ij} A_{ij} = \sum_i k_i^{in} = \sum_i k_j^{out}$. Symmetry is then restored and spectral optimization applied to extract non-overlapping communities. This model leads to a node partition that does not distinguish between the in and out roles; the nodes are simply clustered within the various communities.

To compare these authors' method to ours, we transformed directed graphs into bipartite graphs (this transformation was also suggested in Guimera's work [13] when applying their method for bipartite networks to directed graphs, as will be seen below). At this point, let's differentiate the nodes' roles into $N \times N$. Along these lines, we duplicate N and consider two identical sets N^{out} and N^{in}. The original directed graph G^d is transformed into a bipartite graph $G = (N^{out}, N^{in}, E)$ in which nodes appear twice depending on their 'out' or 'in' role and moreover the asymmetric adjacency matrix A plays the role of biadjacency matrix B in bipartite graphs. We can now define modularity for directed graphs as follows:

$$Q^B = \frac{1}{m} \sum_{ij} [A_{ij} - \frac{(k_i^{in} + k_j^{out})^2}{4m}] \delta(c_i, c_j) \tag{5}$$

After applying any algorithm for a unipartite graph on the corresponding adjacency matrix A' we obtain a partition where some nodes may belong to the same community twice or instead may appear in two different communities. Each model has its pros and cons. Leicht's model [17] is preferable when seeking a single partition with no role distinction. Our model is attractive when seeking to distinguish between 'in' and 'out' roles, e.g. between producers and customers where anyone can play either role. The brain data example that follows will demonstrate that our model is particularly well suited for analyzing real data.

3.1.4. *Turning unipartite graphs into bipartite graphs*

In the above presentation, we introduced modularity for bipartite graphs as a formal derivative of unipartite graph modularity. It is dually possible to consider unipartite graphs as bipartite graphs, and extract communities as if unipartite graphs were bipartite graphs. To proceed, we must consider the original symmetric adjacency matrix A as an asymmetric biadjacency matrix B (with the same nodes on both dimensions) and build a new adjacency matrix A' using the original adjacency matrix A twice on the off-diagonal, as if the nodes had been cloned. When applying a unipartite graph partitioning algorithm, we then obtain communities in which all nodes appear twice. This method only works if we add to A the unity matrix

I (with the same dimensions as A) before building A'. The first diagonal in A in fact only contains 0s since no loops are generally present in a unipartite graph adjacency matrix. Semantically adding I to A means that all objects will be linked to their respective clones in A'. This is a necessary step in that when extracting communities, the objects must drag their clones into the same communities in order to maintain connectivity. In practice therefore, for unipartite graphs, we build A' with $A + I$.

It may seem futile to perform such a transformation from a unipartite graph to a bipartite one in order to find communities in unipartite graphs given that for computing bipartite graph partitioning, we have already made the extension into unipartite graphs using their (symmetric) adjacency matrix. This transformation is nonetheless worthwhile for several reasons. First, when appearing twice, nodes should be associated with their clones. If the resulting communities do not display this property, i.e. a node's clone lies in another community, then the original matrix is not symmetric and can be considered as the adjacency matrix of a directed graph. This conclusion has been applied to the human brain tractography data clustering, which will be described in the experimental section below.

Conversely, if we are sure that the original adjacency matrix is symmetric, then a result where all nodes are associated with their clones in the same communities would be a good indicator of the quality of the clustering algorithm and moreover provides the opportunity to compare our bipartite graph approach with other unipartite graph strategies. This is also a method we introduced into our experiment (see the karate and other applications below) for the purpose of verifying the validity of results.

Lastly, the most important benefit consists of building overlapping communities and ownership functions for unipartite graphs using the method explained in Section 4 below. Although transforming unipartite graphs into bipartite graphs requires more computation, it also provides considerable information opening the way to semantic interpretation, which justifies its application in a variety of contexts.

3.2. *Comparison with other modularity models and partitioning algorithms for bipartite graphs*

Most modularity models which have been proposed in the literature for bipartite graphs are inspired by Newman's modularity for unipartite graphs. In some of them the objective is to distinguish the number of communities in each type of nodes [13] [21][32]. However there is a recent consensus on a probability null model introduced by Barber [1] which is very close to the original Newman's modularity null model for unipartite graphs [18]. Although these authors introduce the same block matrix as we do, their modularity model differs from ours.

After small transformations for unifying notation, Barber's model (see [2] equation 19) is the following:

$$Q_b^B = \frac{1}{m}\sum_{i,j}[B_{ij} - \frac{k_i k_j}{m}]\delta(c_i, c_j) \tag{6}$$

This model is slightly different from our model:

$$Q^B = \frac{1}{m}\sum_{ij}[B_{ij} - \frac{(k_i + k_j)^2}{4m}]\delta(c_i, c_j) \tag{7}$$

The formal difference is obvious and deserves some comments. Our modularity expression is formally derived from Newman's unipartite modularity model (see appendix).

As was shown in equation 3, it is equivalent to considering bipartite graphs as unpartite graphs with both types of graphs behaving the same.

We are therefore inclined to directly apply unipartite graph algorithms which are based on this model and expect modularity optimization. Conversely [1][18], although they consider the same block matrix as we do, they specify a different null model which is conceptually sound but not the result of a direct mathematical derivation from the unipartite model. Therefore either these two definitions are equivalent in terms of final optimization, or, if they are not, Barber's model should be used with specific algorithms for bipartite graphs, or with algorithms for unipartite graphs adapted to bipartite graphs.

If their interpretation is different, the effects of using either this formula or the other can be observed according to two perspectives: 1) the number of communities in each set, 2) the node distribution of each type in the communities. According to our definition of modularity, both types of nodes are explicitly bound. Consequently when applying any unipartite graph algorithm for detecting communities, both types of nodes should have the same number of communities and, except for singletons, they should be regrouped into the same communities (a type of node should not be isolated in a community). This side effect is not explicit in Equation 6. However since in this equation $\delta(c_i, c_j)$ specifies that the summation is applied to both types of objects belonging to the same community, the side effect is the same: optimizing the standard bipartite graph modularity should yield a partitioning of both types of nodes in the same communities (this analysis is also found in [21] : "This definition implicitly indicates that the numbers of communities of both types are equal"). Both modularities should then produce the same results in terms of node type distribution.

As far as the number of communities and node ownership are concerned, it is more difficult to compare the results of both these models, in particular if various algorithms are applied depending on the selected model. For instance, in the Southern Women experiment described below, we found 3 communities when applying Louvain, while Murata in [18] found four communities using their original LPAb+ algorithm. These authors however only provided a quantitative evaluation

via comparison with other algorithms on computation performance and modularity optimization; in contrast, we provide hereafter qualitative analysis as well, which allows for semantics justification on the partitioning as will be showed in next section.

4. Detection and analysis of community overlapping

4.1. *Adding semantics to communities*

The fact that both types of nodes are bound in their communities yields several important results. First, in considering one type of nodes, a community can be defined by associating a subset of nodes from the other type. In other words, nodes from one set provides sense and semantics for the grouping of nodes from the other set and moreover may qualitatively explain regroupings, as will be seen below. This semantic perspective has not been considered by any of the other authors, a situation due to the fact that in other contributions, either the number of communities differs for both types of nodes (e.g. [20], or else when both types of nodes contain the same number of communities they are not bound in each community [13, 1].

Binding both types of nodes into the same communities yields other pertinent results. For one thing, it is possible to define belonging functions and consequently obtain quantified overlapping communities. In the following discussion, we will consider three possible belonging functions, which may expose community overlapping in a different light.

4.2. *Probabilistic function*

Let's adopt the Southern Women's benchmark, which will be more thoroughly described in Section 5.3 below. Applying the Louvain community detection algorithm for unipartite graphs yields a partition where Women and Events are regrouped into three exclusive communities. Let's call these communities c_1 , c_2 and c_3. Now, let's suppose the fictitious case in which woman w_1 participated in events e_1, e_2, e_3 and e_4 . furthermore, w_1, e_1 and e_2 are classified in c_1, while e_3 is classified in c_2 and e_4 is classified in c_3. We can then define a probability function as follows:

$$P(u_i \in c) = \frac{1}{k_i} \sum_j B_{ij} \delta(c_j) \tag{8}$$

where c is a community, $k_i = \sum_j B_{ij}$ and $\delta(c_j) = 1$ if $v_j \in c$ or $\delta(c_j) = 0$ if $v_j \notin c$

In $P(u_i \in c)$ the numerator includes all edges linking u_i to properties $v_j \in c$ and the denominator contains all edges linking u_i to all other nodes. With this function in the present example the probability of w_1 being classified in community e_1 equals $\frac{2}{4}$, and her probabilities of being classified in c_2 and in c_3 are $\frac{1}{4}$ each. The probability a node belongs to a given community is the percentage of its links to this community as a proportion of the total number of links to all communities. In

other words, the greater the proportion of links to a given community, the higher
the expectation of belonging to this community.

4.3. *Legitimacy function and overlapping communities*

It is possible to add more meaning in order to decide which community a given node
should join. The legitimacy function serves to measure the node involvement in a
community and other results to show community overlapping. The more strongly a
node is linked to other nodes in a community, the greater its legitimacy to belong
to the particular community. In the Southern Women's example, let's assume that
after partitioning, c_1 contains 7 events, c_2 5 events and c_3 2 events (which is actually
the case in the experiment presented below). Then, w_1 would have a $\frac{2}{7}$ legitimacy
for c_1, $\frac{1}{5}$ for c_2 and $\frac{1}{2}$ for c_3. The legitimacy function can thus be formalized as
follows:

$$L(u_i \in c)f = \frac{\sum_j B_{ij}\delta(c_j)}{|\{v \in c\}|} \tag{9}$$

where c is a community, $\delta(c_j) = 1$ if $v_j \in c$ or $\delta(c_j) = 0$ if $v_j \notin c$

The numerator in this expression is the same as the probabilistic function nu-
merator. Only the denominator is different.

4.4. *Reassignment Modularity function*

Reassigning node w from C_1 to C_2 either increases or decreases the modularity
defined in Equation (2). Such a change is referred to as Reassignment Modularity
($RM_{w:C_1 \to C_2}$).

The full development about this expression is exposed in Annex 2 (cf section 9).
After simplification this expression yields to:

$$RM_{w:C_1 \to C_2} = \frac{1}{m}(l_{w|2} - l_{w|1}) - \frac{1}{2m^2}[d_w^2 + d_w(d_{C_2} - d_{C_1})] \tag{10}$$

Reassignment is a very interesting measure. It allows detection of nodes that are
not properly assigned to a community. Since most community detection algorithms
are greedy algorithms some nodes may not be in a stable situation. The RM value
reveals unstable nodes and the community to which they should be assigned.

5. Experimentation

This section will consider several benchmarks from various sources. We begin by
applying our method to two simple graphs: the so-called "karate club" unipartite
graph from [35] shows friendship relations between members of a karate sport club;
and the "Southern Women" bipartite graph depicts relations between southern
American women participating in several events. Our method is then applied to

a medium-sized dataset extracted from a real-world situation. For this purpose, we consider a bipartite graph (people tagged on photos) drawn from a student's "Facebook" account containing an average number of photos and people. Lastly, this same method will be applied to human brain data in order to derive dependencies between several areas in the brain. We also applied our method on several well known unipartite and bipartite graph benchmarks as well as on big size benchmarks.

5.1. *Unipartite graph: Karate club*

The karate club graph [35] is a well-known benchmark showing friendship relations between members of a karate club; it is a unipartite graph on which many partitioning algorithms have been experimented. Consequently, this set-up makes it possible not only to verify that our method for bipartite graphs when applied to unipartite graphs meets expected results, but also to assess the additional knowledge extracted from overlapping.

We began by directly applying the Louvain algorithm to the original unipartite graph, represented by its adjacency matrix A. which yielded four separate communities (as shown in 2). These are the same communities extracted by other authors, e.g. [25]. During a second experiment, we considered that the adjacency matrix A is in fact a biadjacency matrix B which is representative of a bipartite graph whose corresponding objects are the club members and whose properties are also club members. An edge exists in the bipartite graph between a club member-object and a club member-property provided an edge is present between the two club members in the original unipartite graph. The new A' adjacency matrix is $A' = \begin{bmatrix} O_r & B \\ B^t & O_s \end{bmatrix}$, where $B = A + I$. and where I is the identity matrix (as explained in section 3.1.4). We once again apply the Louvain algorithm to A'.

Results. As expected, these same four communities identified in the unipartite graph have been extracted from the bipartite graph, with the same individuals appearing twice in each community (see Figure 2). This initial result confirms the absence of bias when transforming a unipartite graph into a bipartite one. The second result is more pertinent because it reveals an overlap between communities when considering legitimacy values. If we were to consider just the cell colorings in the figure, an overlap would be observable whenever at least one node from a community is linked to other nodes in another community. The legitimacy values that indicate the involvement of each node in each community offer an effective tool for identifying and analyzing new features. Some slight differences have been noted in works by other authors: for example, in page 2, Porter [29] placed node number 10 in the second community. In our case, this node has been placed in the first community, though the legitimacy value suggests that it should have been placed in the second one, in which case the situation would be reversed in the second community and node 10 would have a legitimacy value that alters its placement in the first community. Node 10 is thus in a hesitation mode between the two
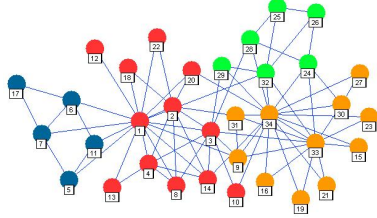
12   *Crampes, Plantié*

Fig. 1. Karate club graph with partitioned communities

node N°	1	2	3	4	8	10	12	13	14	18	20	22	9	15	16	19	21	23	27	30	31	33	34	24	25	26	28	29	32	5	6	7	11	17
Unipartite community N°	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4
overlapping community + legitimacy — 1	10/12	8/12	6/12	5/12	4/12	1/12	1/12	2/12	4/12	2/12	2/12	2/12	2/12	0	0	0	0	0	0	0	1/12	1/12	1/12	0	0	0	1/12	1/12	1/12	1/12	1/12	1/12	1/12	0
2	1/11	1/11	2/11	0	0	1/11	0	0	1/11	0	1/11	0	3/11	2/11	2/11	2/11	2/11	2/11	2/11	3/11	3/11	9/11	10/11	3/11	0	0	1/11	1/11	2/11	0	0	0	0	0
3	1/6	0	2/6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/6	0	2/6	4/6	2/6	3/6	2/6	1/6	3/6		0	0	0	0	0
4	4/5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2/5	3/5	3/5	2/5	2/5

Fig. 2. Karate club communities and legitimacy measures

communities.

To the best of our knowledge, this experiment represents the first time Karate communities are shown as separate and overlapping. Partitioning provides a practical way to observe communities; however, overlapping reveals the extent to which partitioning reduces the amount of initial information. With our method for example, it can be seen that some nodes actually straddle several communities, e.g. node 10 in our experiment.

5.2. *Unipartite graphs: other known benchmarks*

We have applied our method on several other well known unipartite graphs, such as Dolphins and other benchmark graphs such as those in [11]. As for the "Karate" case, we get the same community partitions as Newman algorithm [25]. [15] proposed a well known algorithm to generate benchmark graphs (also used by [29, 8, 12] and others) where communities are well identified. We used this algorithm to generate 30, 128, 500 and 1000 node such graphs to test our algorithms and show the efficiency of our method. We do find the same number of communities as Newman's algorithm since the modularity formula we use is directly derived from Newman's one and we get the same analysis and results as in [15]. However we provide a very interesting knowledge with supplementary data to observe node overlapping on these communities.

The modularity has a limited resolution that depends on the number of edges in the network [9]. We observed a main consequence of the resolution limit: the modules in large networks may have hidden substructures that require deeper investigations

to reveal.

5.3. *Bipartite graph: Southern Women*

This benchmark has been studied by most authors interested in checking their partitioning algorithm for bipartite graphs. The goal here is to partition, into various groups, 18 women who attended 14 social events according to their level of participation in these events. In his well-known cross-sectional study, [10] compared results from 21 authors, most of whom identified two groups.

Results. In Figure 3, the bipartite graph is depicted as a bi-layer graph in the middle with women at the top and events below; moreover, the edges between women and events represent woman-event participations. Three clusters with associated women and events have been found and eventually shown with red, blue and yellow colorings. This result is more accurate than the majority of results presented in [10]; only one author found three female communities. Beyond mere partitioning, Figure 3 presents overlapping communities using two overlapping functions, namely legitimacy and reassignment modularity (RM). Legitimacy and RM for women are placed just above female partitioning; for events, both are symmetrically shown below event partitioning. As expected, reassignment in the same community produces a zero RM value. The best values for legitimacy and RM have been underscored. Only the values of woman 8 and event 8 indicate that they could have been in another community. This is the outcome of early assignment during the first Louvain phase for entities with equal or nearly equal probabilities across several communities. It can be observed in [10] that woman 8's community is also debated by several authors; our results appear to be particularly pertinent in terms of both partitioning and overlapping.

The fact that women and events are correlated may be considered to cause a bias, such as in the number of communities. When comparing our results to those of other authors however, the merging of our blue and yellow communities produces their corresponding second community. In their trial designed to obtain a varying number of communities in both sets, Suzuki [32] found a large number of singletons. Their results were far from those presented in [10], while ours were compatible and more highly detailed.

In conclusion, results on the Southern Women's benchmark are particularly relevant. Moreover, our visualization enables observing community partitioning, overlapping and possible assignment contradictions. The application of reassignment for better modularity optimization will be tested in a subsequent work.

5.4. *Bipartite graph: Facebook account*

In a Facebook (FB) account, several types of informations may be extracted. We extracted and evaluated only data coming from FB photo albums with its tags. We did not use friendship relations. Three Facebook photo files were downloaded from various Facebook (FB) accounts. All these files were extracted with the consent
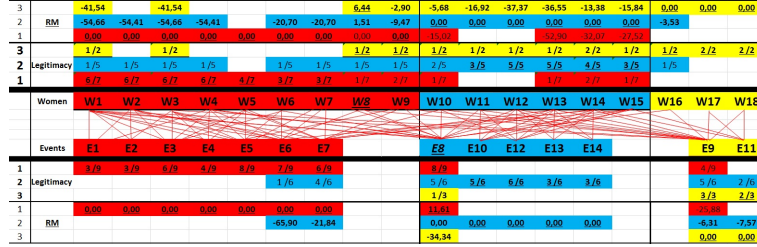
14   *Crampes, Plantié*

		W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16	W17	W18
3		-41,54		-41,54					6,44	-2,90	-5,68	-16,92	-37,37	-36,55	-13,38	-15,84	0,00	0,00	0,00
2	RM	-54,66	-54,41	-54,66	-54,41		-20,70	-20,70	1,51	-9,47	0,00	0,00	0,00	0,00	0,00	0,00	-3,53		
1		0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-15,02		52,90	-32,07	-27,52				
3		1/2		1/2					1/2	1/2	1/2	1/2	1/2	2/2	1/2	1/2	2/2	2/2
2	Legitimacy	1/5	1/5	1/5	1/5		1/5	1/5	1/5	1/5	2/5	3/5	5/5	5/5	4/5	3/5	1/5		
1		6/7	6/7	6/7	6/7	4/7	3/7	3/7	1/7	2/7	1/7		1/7	2/7	1/7				
	Women	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16	W17	W18
	Events	E1	E2	E3	E4	E5	E6	E7			E8	E10	E12	E13	E14			E9	E11
1		3/9	3/9	6/9	4/9	8/9	7/9	6/9			8/9							4/9	
2	Legitimacy						1/6	4/6			5/6	5/6	6/6	3/6	3/6			5/6	2/6
3											1/3							3/3	2/3
1		0,00	0,00	0,00	0,00	0,00	0,00	0,00			11,61							-25,88	
2	RM						-65,90	-21,84			0,00	0,00	0,00	0,00	0,00			-6,31	-7,57
3											-34,34							0,00	0,00

Fig. 3. Women Events communities with legitimacy and Reassignment modularity measures

of their owners, none of whom were members of the research team. A person was considered to be linked to a photo if he/she had been tagged in the photo. We then have a bipartite graph composed of two type of nodes : persons and photos. Community extraction using our method reveals some common features among the datasets. These features are shown in Figure 4 for one FB photo file, in which 274 people could be identified in a total of 644 photos.

Results. Communities are seldom overlapping, which supports the notion that the photos were taken at different times in the owner's life (this is to be confirmed in a forthcoming study). When the owner was asked to comment on the communities, two main observations were submitted. The various groups of people were indeed consistent, yet with one exception. The owner was associated in the partition with a group she had met on only a few occasions and not associated with other groups of close friends. An analysis of the results provided a good explanation, which is partially displayed in Figure 4. From this view, the FB account owner is in the first community on the left, yet she is also present in most of the other communities (see grey color levels in the first column). Although at first glance it might be assumed that she is not part of other communities, our visualization indicates that such is not the case. She is present in most communities, even though she is mainly identified in the first one. Three types of photos can be distinguished in this first community. More than 200 photos only contain the owner's tag, plus a few photos with unique tags of another community member; for every other person, at least one photo tags him/her with the owner. This first community has in fact been built from the first group with photos of unique owner's tags associated with the owner. The owner's tag thus encompasses photos containing two people, one of whom is the owner. It turns out that this group is predominantly the owner's group.

In conclusion, partitioning only the bipartite graph would have produced a major pitfall: the owner would have been isolated in a community that is not his/her top preference. With our method, merging partitioning and overlapping exposes better multiple regroupings with broader affinities. Other communities also showed high consistency when considering the photos: each community was associated with some particular event responsible for gathering a group of the FB account owner's friends.
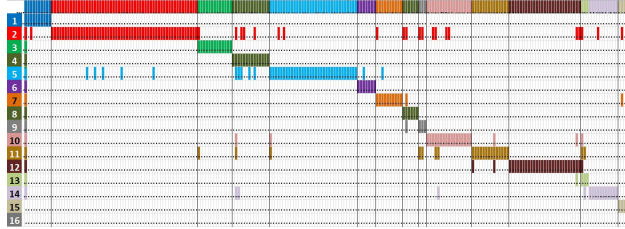
Fig. 4. Facebook account communities with overlapping

5.5. *Bipartite graph: Brain Data*

Our method was initially designed for human community detection and analysis. In this experiment, we have demonstrated how it can be applied to other data analysis techniques as well. The brain dataset was collected on a single patient by a research team affiliated with the "Human Connectome" project working on brain tractography techniques [4]. These techniques use Magnetic Resonance Imaging (MRI) and Diffusion Tensor Imaging (DTI) to explore white matter tracks between brain regions. Probabilistic tractography produces 'connectivity' matrices between Regions Of Interest (ROI) in the brain. For the case we studied, 'seed' ROIs were located in the occipital lobe and 'target' ROIs throughout the entire brain. The goal here was to detect possible brain areas in the occipital lobe through ROI clustering on the basis of similar track behavior. In [4], the research team used Spectral Clustering (SC) to combine ROIs. It is interesting to note that SC is one of numerous techniques that have traditionally been applied in social community detection, e.g. by Bonacich on the Southern Women's benchmark [10]. SC results are limited to community partitioning (though in theory overlapping could also be computed). The goal was to experiment with our method and produce both partitioning and overlapping analyzes of brain areas.

The original matrix contained 1,914 rows and 374 columns, with cells denoting the probabilities of linkage between ROIs. We considered this matrix as a bipartite graph biadjacency matrix with weighted values and then applied our community detection method. Figure 5 presents the results of ROI community partitioning and overlapping. Each color in the first row is associated with a community that gathers several ROIs. Each ROI is represented by a column that indicates its belonging to the other communities. When a cell is highlighted with a color, a nonzero overlapping value exists for both this ROI and the corresponding community (with community numbers being plotted on the left-hand side of the figure). This value has been computed with the legitimacy function, which has been extended to the weighted edges, i.e. the weighted sum of values from cerebral hemisphere zones (ELF) within the selected community. Each community is associated with a threshold value corresponding to the maximum weighted legitimacy above which the community would lose a full member. For each community, this threshold value is automatically com-
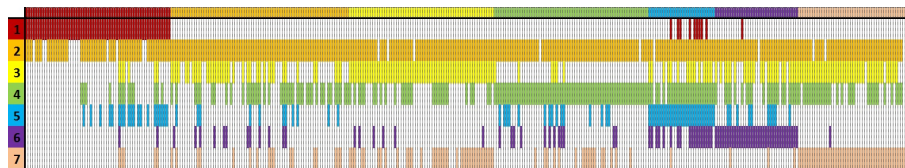
16   *Crampes, Plantié*

Fig. 5. Brain data communities with overlapping

puted in order to include all ROI members of the community.

Results. We found 7 communities when neurologists selected 8 clusters with SC and after choosing the most significant eigenvectors on a scree test. Let's observe that two communities overlap heavily on all others, which thus overlap to a lesser extent. Figure 5 confirms the strong interest in this set-up that simultaneously exhibits overlapping and non-overlapping data. These results have been taken into account by a team of neurological researchers as different observations recorded on brain parcellation.

5.6. *Bipartite graphs: others known benchmarks*

Bipartite graph datasets are not easy to find in litterature. We also tested our algorithms with bipartite networks used as benchmark networks in [1]. One of them is the network benchmark describing corporate interlocks in Scotland in the early twentieth century. The data set characterizes 108 Scottish firms during 1904-5, detailing the corporate sector, capital, and board of directors for each firm. The data set includes only those board members who held multiple directorships, totaling 136 individuals. Barber found "roughly" (sic) 20 communities, whereas we find 15 communities and provide very interesting knowledge about overlapping for these communities. We obtained a global modularity of 0.71038 whereas Barber found a smaller value of 0.56634.

To evaluate scalability on our method we tested a rather big co-authorship bipartite dataset to detect scientific communities extracted from the well known PubMed (http://www.ncbi.nlm.nih.gov/pubmed) biomedical scientific litterature online library. Our dataset was composed of 30,000 persons and more than 80,000 scientific papers. We extracted 184 communities of average 670 members in about 3 seconds, with interesting overlapping information. Regarding resolution limit mentionned earlier, the modularity method applied to bipartite graphs has a similar limit, with similar consequences.

6. Discussion and new perpectives

The above experiments show that our method is able to find overlapping communities in different types of graphs. Moreover, it is able to measure the degree of membership for each node to each community. We then get a first semantic interpretation of each node in terms of community membership. These results are obtained

through the use of the off-diagonal block square matrix $A^{'}$. Several other methods may compute modularity by using directly a graph structure without building any off-diagonal block matrix. For example LPA based methods [12, 18] which use Barber's modularity definition for bipartite graphs may work directly with the graph structure. However the results that are presented in these papers are different. They find 4 communities for the Women Events dataset instead of 3 in our case. Since Barber's modularity expression is different from ours, it is difficult to compare these different results.

Louvain algorithm which uses Newman's modularity formula is adapted to monopartite graphs. Since the approach with the block matrix requires more data and computation, we also tried applying Louvain algorithm directly on to the biadjacency matrix B. We tested it on the same two bipartite data sets: Women Events and Facebook. Surprisingly, we got the same results as those with the block square matrix $A^{'}$. This experiment suggests the possibility of directly applying unipartite graph models onto bipartite graph models with unipartite graph modularity. Moreover our method with the block matrix $A^{'}$ could be a good means for validating this possibility.

This counter intuitive conclusion needs more experiments and more theoretical proof. Particularly since other authors use Barber's model which is specifically adapted for bipartite graphs. Future work will deeper investigate this possibility of directly applying unipartite graph methods to bipartite graphs.

7. Conclusion

In this paper, we have demonstrated the feasibility of unifying bipartite graphs, directed graphs and unipartite graphs under a common unipartite graph model. It was then proved that any unipartite graph partitioning algorithm aiming at optimizing the standard unipartite modularity model leads to a bipartite graph partitioning, wherein both types of nodes are bound in the communities. In the special case of directed graphs, nodes appear twice in potentially different communities depending on their roles; for unipartite graphs, nodes are cloned and appear with their clones in the same communities.

We also introduced the possibility of unifying in a single view, the partitioning and the overlapping communities. This development is possible thanks to associating both types of nodes in the communities. Moreover, overlapping can be characterized through several functions presenting different interpretations. For instance, it is possible to identify those nodes that define the community cores, i.e. those who belong exclusively to just one community and, conversely, those who serve as bridges between different communities. We also introduced reassignment values which open up the possibility of improving partitioning results. Practically speaking, when applying our method to various benchmarks and datasets, we are able to extract meaningful communities and display surprising overlapping properties when other authors limit their goal to identifying communities. We extend far

18   *Crampes, Plantié*

beyond this point and provide tools for analyzing and interpreting results.

Lastly, we introduced an essential result after experimenting on real brain datasets, supplied by a research team from the Connectome project. Historically authors dealing with community detection problems used to borrow their methods from data or graph analysis such as hierarchical clustering, clique enumeration or spectral analysis. Recent community detection approaches based on modularity optimization use original methods (Louvain, label propagation). We showed that these methods could also be applied to data analysis with good results. Moreover these results can be obtained without the need to choose parameters such as the number of clusters, or a threshold value. It is of particular interest to note that after borrowing their methods from other scientific domains, community detection techniques are now enough mature for providing these domains with new original performing methods.

In the future we will continue exploring cross fertilization between community detection techniques and other scientific domains. In particular we will use Nash equilibrium for studying community stability through the reassignment value we introduced in this paper. Indeed we think that community stability could be another quality criteria along with modularity optimization for driving and assessing community detection algorithms' performances.

Acknowledgments

References

[1] Michael Barber. Modularity and community detection in bipartite networks. *Physical Review E*, 76(6):1–9, 2007.
[2] Michael J Barber and John W Clark. Detecting network communities by propagating labels under constraints. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, 80(2 Pt 2):026129, 2009.
[3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, October 2008.
[4] Marco Catani and Michel Thiebaut de Schotten. *Atlas of human brain connections.* Oxford University Press, 2012, 2012.
[5] Abhijnan Chakraborty, Saptarshi Ghosh, and Niloy Ganguly. Detecting overlapping communities in folksonomies. In *Proceedings of the 23rd ACM conference on Hypertext and social media HT 12*, page 213. ACM Press, 2012.
[6] Ernesto Estrada and Juan A Rodriguez-Velazquez. Complex Networks as Hypergraphs. *Systems Research*, page 16, 2005.
[7] T S Evans and R Lambiotte. Line Graphs, Link Partitions and Overlapping Communities. *Physical Review E*, 80(1):9, 2009.
[8] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):103, June 2009.

[9] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104(1):36–41, 2007.

[10] Linton C. Freeman. Finding social groups: A meta-analysis of the southern women data. In *Dynamic Social Network Modeling and Analysis. The National Academies*, pages 39—-97. Press, 2003.

[11] M. Girvan and M E J Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, 2002.

[12] Steve Gregory. Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12(10):103018, 2009.

[13] Roger Guimerà, Marta Sales-Pardo, and Luís Amaral. Module identification in bipartite and directed networks. *Physical Review E*, 76(3), September 2007.

[14] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, March 2009.

[15] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, 78(4 Pt 2):6, 2008.

[16] Conrad Lee, Fergal Reid, Aaron McDaid, and Neil Hurley. Detecting highly overlapping community structure by greedy clique expansion. *4th Workshop on Social Network Mining and Analysis SNAKDD10*, 10:10, 2010.

[17] E A Leicht and M E J Newman. Community structure in directed networks. *Physical Review Letters*, 100(11):118703, 2007.

[18] Liu Xin and Murata Tsuyoshi. An Efficient Algorithm for Optimizing Bipartite Modularity in Bipartite Networks. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 14(4):408–415, 2010.

[19] Michel Plantié and Michel Crampes. Mining social networks and their visual semantics from social photos. *International Journal of Computer science & Applications*, VIII(II):102–117, 2011.

[20] Tsuyoshi Murata. Modularities for bipartite networks. *Proceedings of the 20th ACM conference on Hypertext and hypermedia HT 09*, 90(6):245–250, 2009.

[21] Tsuyoshi Murata. Detecting communities from tripartite networks. *WWW*, pages 0–1, 2010.

[22] Neubauer Nicolas and Obermayer Klaus. Towards Community Detection in k-Partite k-Uniform Hypergraphs. In *Proceedings NIPS 2009 . . .* , 2009.

[23] Mark Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6), June 2004.

[24] Mark Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, 74(3 Pt 2):036104, 2006.

[25] Mark Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2), February 2004.

[26] Andreas Noack and Randolf Rotta. Multi-level algorithms for modularity clustering. *arXiv*, page 12, December 2008.

[27] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–8, June 2005.

[28] S Papadopoulos, Y Kompatsiaris, A Vakali, and P Spyridonos. Community detection in Social Media. *Data Mining and Knowledge Discovery*, 1(June):1–40, 2011.

[29] Mason A. Porter, Jukka-Pekka Onnela, and Peter J. Mucha. Communities in Networks, 2009.

[30] Camille Roth and Paul Bourgine. Epistemic Communities: Description and Hierarchic Categorization. *Mathematical Population Studies: An International Journal of Mathematical Demography*, 12(2):107–130, 2005.

[31] Sune Lehmann,Martin Schwartz,Lars Kai Hansen. Biclique communities. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 78(1 Pt 2), 2008.

[32] Kenta Suzuki and Ken Wakita. Extracting Multi-facet Community Structure from Bipartite Networks. *2009 International Conference on Computational Science and Engineering*, 4:312–319, 2009.

[33] Zhihao Wu, Youfang Lin, Huaiyu Wan, Shengfeng Tian, and Keyun Hu. Efficient overlapping community detection in huge real-world networks. *Physica A: Statistical Mechanics and its Applications*, 391(7):2475 – 2490, 2012.

[34] Bo Yang, Dayou Liu, Jiming Liu, and Borko Furht. *Discovering communities from Social Networks: Methodologies and Applications* . Springer US, Boston, MA, 2010.

[35] W W Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.

8. Annex 1

8.1. *New use of Newman modularity*

In this Annex, we will provide full details of the demonstration that yielded Equation (2).

For the sake of convenience, let's use the definition of unipartite graph modularity offered in Newman [17]. It is a function Q of matrix $A^{'}$ and the communities detected in G [25]:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A^{'}_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \qquad (11)$$

where $A^{'}_{ij}$ denotes the weight of the edge between i and j, $k_i = \sum_j A^{'}_{ij}$ is the sum of the weights of edges attached to vertex i, c_i is the community to which vertex i has been assigned, the Kronecker's function $\delta(u,v)$ equals 1 if $u = v$ and 0 otherwise and $m = 1/2 \sum_{ij} A^{'}_{ij}$.

In our particular case (i.e. where $A^{'}$ is the off-diagonal block adjacency matrix of a bipartite graph), we apply the following transformations:

Let's rename i_1 as index i when $1 \leq i \leq r$ and i_2 when $r < i \leq r+s$. Conversely, let's rename j_1 the index j when $1 \leq j \leq r$ and j_2 when $r < j \leq r+s$.

To avoid confusion between the $A^{'}$'s indices and B's indices let's rename B indices i_b and j_b : $1 \leq i_b \leq r$ and $1 \leq j_b \leq s$ (see a representation of A matrix below (Figure 12))

$$A' = \begin{array}{|c|c|c|c|c|}
\hline
A' \ indexes \downarrow\rightarrow & ....j_1.... & ....j_2.... & & \\
\hline
\begin{array}{c}...\\ i_1 \\ ...\end{array} & O_r & B & \begin{array}{c}...\\ i_b \\ ...\end{array} & r\ rows \\
\hline
\begin{array}{c}...\\ i_2 \\ ...\end{array} & B^t & O_s & \begin{array}{c}...\\ j_b \\ ...\end{array} & s\ rows \\
\hline
 & ....i_b.... & ....j_b.... & \leftarrow\uparrow B\ indexes & \\
\hline
 & r\ columns & s\ columns & & \\
\end{array} \qquad (12)$$

Let's call k_{i_b} the margin of row i_b in B and k_{j_b} the margin of column j_b in B.

$$k_{i_b} = \sum_{j_b} B_{i_b j_b} = \sum_{j_2} A'_{i_1 j_2} = \sum_{i_2} A'_{i_2 j_1}, \ where \ i_b = i_1 = j_1 \qquad (13)$$

$$k_{j_b} = \sum_{i_b} B_{i_b j_b} = \sum_{i_1} A'_{i_1 j_2} = \sum_{j_1} A'_{i_2 j_1}, \ where \ j_b = i_2 - r = j_2 - r \qquad (14)$$

$k_{i_b}$is the degree of node u_{i_b}, k_{j_b} is the degree of node v_{j_b}. Let's define $k_{i/j_1} = \sum_{j_1} A'_{ij_1}$ and $k_{i/j_2} = \sum_{j_2} A'_{ij_2}$. Conversely : $k_{j/i_1} = \sum_{i_1} A'_{ji_1}$ and $k_{j/i_2} = \sum_{i_2} A'_{ji_2}$. Hence : $k_i = \sum_j A'_{ij} = k_{i/j_1} + k_{i/j_2}$, $k_j = \sum_i A'_{ij} = k_{j/i_1} + k_{j/i_2}$.

By taking into account the structure and properties of A' in (13) and (14) for the indices we derive the following properties :

k_{i/j_1} has non-zero values only for $i = i_2$, with k_{j_b} the degree of node v_{j_b}:

$$k_{i/j_1} = k_{i_2/j_1} = \sum_{j_1} A'_{i_2 j_1} = \sum_{i_1} A'_{i_1 j_2} = k_{j_2/i_1} = k_{j_b} \qquad (15)$$

k_{i/j_2} has non-zero values only for $i = i_1$, with k_{i_b} the degree of node u_{i_b}:

$$k_{i/j_2} = k_{i_1/j_2} = \sum_{j_2} A'_{i_1 j_2} = \sum_{i_2} A'_{i_2 j_1} = k_{j_1/i_2} = k_{i_b} \qquad (16)$$

Moreover and more directly: k_{j/i_1} offers values only for $j = j_2$: $k_{j/i_1} = k_{j_2/i_1} = k_{i_2/j_1} = k_{j_b}$, the degree of node v_{j_b}. k_{j/i_2} offers values only for $j = j_1$: $k_{j/i_2} = k_{j_1/i_2} = k_{i_1/j_2} = k_{i_b}$, the degree of node u_{i_b}.

8.2. *Analyzing second part of Q in equation* (11)

Using these properties of matrix A', it is now possible to analyze $\sum_{ij} k_i k_j$. in equation (11).

Next, by developing k_i and k_j in A' we obtain: $\sum_{ij} k_i k_j = \sum_{ij} (k_{i/j_1} + k_{i/j_2})(k_{j/i_1} + k_{j/i_2})$

$$= \sum_{ij} k_{i/j_1} k_{j/i_1} + \sum_{ij} k_{i/j_2} k_{j/i_2} + \sum_{ij} k_{i/j_1} k_{j/i_2} + \sum_{ij} k_{i/j_2} k_{j/i_1}$$

$$= \sum_{i_2 j_2} k_{i_2/j_1} k_{j_2/i_1} + \sum_{i_1 j_1} k_{i_1/j_2} k_{j_1/i_2} + \sum_{i_2 j_1} k_{i_2/j_1} k_{j_1/i_2} + \sum_{i_1 j_2} k_{i_1/j_2} k_{j_2/i_1} \qquad (17)$$

Let's note that $\sum_{ij} k_{i/.} k_{j/.} = \sum_i k_{i/.} \sum_j k_{j/.}$ where the dot may take any value in i_1, i_2, j_1, j_2.

Let c be a community, in equation (11) summations $\sum_{ij} k_i k_j$ on indices i and j may only be applied under the condition $\delta(c_i, c_j) = 1$. Where an edge is present between two nodes u and v belonging to c: $\delta(c_i, c_j) = 1$ and $\delta(c_j, c_i) = 1$. Consequently for each row i representing a node belonging to c, a corresponding column j represents this same node belonging to c and *vice versa*.

From (15), (16) and the above observations:

$$\sum_{ij} k_{i/j_1} k_{j/i_1} \delta(c_i, c_j) = \sum_i k_{i/j_1} \sum_j k_{j/i_1} \delta(c_i, c_j) =$$
$$\sum_{i_2} k_{i_2/j_1} \sum_{j_2} k_{j_2/i_1} \delta(c_{i_2}, c_{j_2}) = \sum_{j_b} k_{j_b} \sum_{j_b} k_{j_b} = [\sum_{j_b} k_{j_b}]^2$$
$$\sum_{ij} k_{i/j_2} k_{j/i_2} \delta(c_i, c_j) = \sum_i k_{i/j_2} \sum_j k_{j/i_2} \delta(c_i, c_j) =$$
$$\sum_{i_1} k_{i_1/j_2} \sum_{j_1} k_{j_1/i_2} \delta(c_{i_2}, c_{j_2}) = \sum_{i_b} k_{i_b} \sum_{i_b} k_{i_b} = [\sum_{i_b} k_{i_b}]^2$$
$$\sum_{ij} k_{i/j_1} k_{j/i_2} \delta(c_i, c_j) = \sum_i k_{i/j_1} \sum_j k_{j/i_2} \delta(c_i, c_j) =$$
$$\sum_{i_2} k_{i_2/j_1} \sum_{j_1} k_{j_1/i_2} \delta(c_{i_2}, c_{j_1}) = \sum_{j_b} k_{j_b} \sum_{i_b} k_{i_b}$$
$$\sum_{ij} k_{i/j_2} k_{j/i_1} \delta(c_i, c_j) = \sum_i k_{i/j_2} \sum_j k_{j/i_1} \delta(c_i, c_j) =$$
$$\sum_{i_1} k_{i_1/j_2} \sum_{j_2} k_{j_2/i_1} \delta(c_{i_2}, c_{j_1}) = \sum_{i_b} k_{i_b} \sum_{j_b} k_{j_b}$$

where $j_b = i_2\text{–}r = j_2\text{–}r$, $i_b = i_1 = j_1$, $u_{i_b} \in c$ and $v_{i_b} \in c$ these last two conditions can also be formalized with $\delta(c_{i_b}, c_{j_b}) = 1$ if u_{i_b} and v_{i_b} belong to the same community c and $\delta(c_{i_b}, c_{j_b}) = 0$ otherwise.

This development yields :

$$\sum_{ij} k_i k_j = [\sum_{j_b} k_{j_b}]^2 + [\sum_{i_b} k_{i_b}]^2 + 2[\sum_{j_b} k_{j_b}][\sum_{i_b} k_{i_b}] = \sum_{i_b j_b} (k_{i_b} + k_{j_b})^2 \text{ and:}$$

$$\sum_{ij} k_i k_j \delta(c_i, c_j) = \sum_{i_b j_b} (k_{i_b} + k_{j_b})^2 \delta(c_{i_b}, c_{j_b}) \qquad (18)$$

Equation (18) can be rewritten using the degrees of nodes:

$\sum_{i_b} k_{i_b}$ is the sum of the degrees of nodes u_{i_b} belonging to c under the condition δ in equation (18). We denote this $d_{u|c}$.

$\sum_{j_b} k_{j_b}$ is the sum of the degrees of nodes v_{j_b} belonging to c under the condition δ in equation (18) and has been called $d_{v|c}$.

$$Then \sum_{ij} k_i k_j \delta(c_i, c_j) = (d_{u|c} + d_{v|c})^2 \qquad (19)$$

8.3. *Analyzing first part in equation* (11)

First part in Q is $\sum_{ij} A'_{ij}$. Let's examine what it represents in terms of B. It is possible to identify matrix B in A' using indices i_1 and j_2. Conversely B^t can be identified with indices i_2 and j_1:

For $i = i_1$ A'_{ij}s only produce values for $j = j_2$, moreover for $i = i_2$, A'_{ij}s only produce values for $j = j_1$ with $A'_{i_1 j_2} = B_{i_b j_b}$ and $A'_{i_2 j_1} = B^t_{i_b j_b}$ under typical conditions regarding indices.

Then $\sum_{ij} A'_{ij} = \sum_{i_1 j_2} A'_{i_1 j_2} + \sum_{i_2 j_1} A'_{i_2 j_1}$

And $\sum_{ij} A'_{ij} \delta(c_i, c_j) = \sum_{i_1 j_2} A'_{i_1 j_2} \delta(c_{i_1}, c_{j_2}) + \sum_{i_2 j_1} A'_{i_2 j_1} \delta(c_{i_2}, c_{j_1})$

The left-hand side of the sum equals the number of edges from nodes u to nodes v inside c. The right-hand side is the number of edges from these same nodes v and u inside c. This set-up then leads to:

$\sum_{i_1 j_2} A'_{i_1 j_2} \delta(c_{i_1}, c_{j_2}) = \sum_{i_2 j_1} A'_{i_2 j_1} \delta(c_{i_2}, c_{j_1})$ *with* $i_1 = j_2$ *and* $i_2 = j_1$

$$Then \sum_{ij} A'_{ij} \delta(c_i, c_j) = 2 \sum_{i_1 j_2} A'_{i_1 j_2} \delta(c_{i_1}, c_{j_2}) = 2 \sum_{i_b j_b} B_{i_b j_b} \delta(c_{i_b}, c_{j_b}) \tag{20}$$

This value can also be formalized using the number of edges:

$$\sum_{i_b j_b} B_{i_b j_b} \delta(c_{i_b}, c_{j_b}) = |(u_{i_{b|c}}, v_{j_{b|c}})| = |e_{i_{b|c}, j_{b|c}}| \; where \; e_{i_{b|c}, j_{b|c}} \in E \; \& \; u_{i_{b|c}}, v_{j_{b|c}} \in c$$

$$\tag{21}$$

For the entire matrix $A' : \sum_{ij} A'_{ij} = 2 \sum_{i_b j_b} B_{i_b j_b}$

From equation (11), $m = 1/2 \sum_{ij} A'_{ij}$

Let's now define $m_b = \sum_{i_b j_b} B_{i_b j_b} = |e_{i_b j_b}|$ where $e_{i_b j_b} \in E$

Then $m = \frac{1}{2} \times \sum_{ij} A'_{ij} = \frac{1}{2} \times 2 \times \sum_{i_b j_b} B_{i_b j_b} = m_b$

8.4. *Modularity for all graphs*

Lastly, by removing sub-index b, which had only been introduced to distinguish indices i and j when applied to A' or B, we can redefine the A' modularity in terms of B:

$$Q^B = \frac{1}{m} \sum_{ij} [B_{ij} - \frac{(k_i + k_j)^2}{4m}] \delta(c_i, c_j) \tag{22}$$

In terms of edges, by simplifying $e_{i_{b|c}, j_{b|c}}$ as e_c (where e_c has both ends in c) and by dropping sub-index b Equation (22) becomes:

$$Q^B = \sum_c [\frac{|e_c|}{m} - (\frac{(d_{u|c} + d_{v|c})}{2 \times m})^2] \tag{23}$$

This definition of modularity may be used for bipartite graphs since both types of nodes are bound. In previous sections, we have validated the above results on the basis of another author's graph modularity models. It can thus be concluded that equation (22) offers a good candidate for bipartite graph modularity that takes some specific characteristics into account.

9. Annex 2: Reassignment Modularity function

In this Appendix, we will provide full details of the demonstration that yielded Equation 10.

Reassigning node w from C_1 to C_2 either increases or decreases the modularity defined in Equation (2). Such a change is referred to as Reassignment Modularity ($RM_{w:C_1 \to C_2}$).

Let w be a node u or v. If w is withdrawn from C_1 and reassigned to C_2, then we can define $RM_{w:C_1 \to C_2} = Q^B_{w \in C_2} - Q^B_{w \in C_1}$

where Q^B is the modularity value in:

$$Q^B = \sum_c \left[\frac{|e_c|}{m} - \left(\frac{(d_{u|c} + d_{v|c})}{2 \times m} \right)^2 \right]. \tag{24}$$

Let $l_{w|i} = l_{w,w'|w' \in C_i}$ be the number of edges between a node w and all other nodes w' where $w' \in C_i$,

Let d_w be the degree of w, $|e_i|$ the number of edges in C_i and $d_{C_i} = d_{u|c_i} + d_{v|c_i}$.

We consider that the node w which belongs to C_1 is bound to be withdrawn from this community and assigned to the community C_2.

$Q^B_{w \in C_2}$ is $Q^B_{w \in C_1}$ with correction after w is reassigned. Then

$Q^B_{w \in C_1} = \left[\frac{1}{m}|e_1| - \frac{(d_{C_1})^2}{(2m)^2} + \frac{1}{m}|e_2| - \left(\frac{(d_{C_2})^2}{(2m)^2} \right) \right] + K_{others}$ where K_{others} is the contribution to modularity brought by other communities than C_1 and C_2. This last value does not change when reassigning a node from C_1 to C_2.

$Q^B_{w \in C_2} = \left[\frac{1}{m}(|e_1| - l_{w|1}) + \frac{1}{m}(|e_2| + l_{w|2}) - \left(\frac{(d_{C_1} - d_w)^2}{(2m)^2} + \frac{(d_{C_2} + d_w)^2}{(2m)^2} \right) \right] + K_{others}$,

then

$Q^B_{w \in C_2} - Q_{w \in C_1} = \left[\frac{1}{m}(|e_1| - l_{w|1}) + \frac{1}{m}(|e_2| + l_{w|2}) - \left(\frac{(d_{C_1} - d_w)^2}{(2m)^2} + \frac{(d_{C_2} + d_w)^2}{(2m)^2} \right) \right] - \left[\frac{1}{m}|e_1| - \frac{(d_{C_1})^2}{(2m)^2} + \frac{1}{m}|e_2| - \left(\frac{(d_{C_2})^2}{(2m)^2} \right) \right]$

and after simplification,

$$RM_{w:C_1 \to C_2} = \frac{1}{m}(l_{w|2} - l_{w|1}) - \frac{1}{2m^2}[d_w^2 + d_w(d_{C_2} - d_{C_1})] \tag{25}$$

This equation can be partly validated if after withdrawing w from C_1 we put it back into C_1 and expect no change for Q^B, i.e. $RM_{w:C_1 \to C_1} = 0$. Considering that C_2 is in fact C_1 without w, we get $d_{C_2} = d_{C_1} - d_w$, replacing d_{C_2} in equation (25) by its value yields $RM_{w:C_1 \to C_1} = 0$.

A second validation can be performed with Equation 5 in [33]. Although the authors' demonstration is limited, it can still be noticed that their final formula resembles ours with a slight difference (i.e. division by 2 in their case) due to their definition of modularity for overlapping communities. Moreover, in arguing that the right part of their equation is not meaningful for large graphs, the authors only considered $dEQ = \frac{l_2 - l_1}{2m}$ which is the equivalent of $\frac{1}{m}(l_{w|2} - l_{w|1})$ in our Reassignment Modularity definition. In our case, we do not limit reassignment to large graphs and we keep the whole value in Equation (25).