Some genes have many partners: Detecting biological modularity
Progress Report II
(version 0)

Antoine Allard
antoine.allard.1@ulaval.ca
http://dynamica.phy.ulaval.ca

August 19, 2011

Abstract

This report describes the results obtained following the recommendations from the document *Thoughts and comments I* (sent on August 5$^{\text{th}}$).

1 Theoretical considerations

We recall a few theoretical concepts useful to understand this report as well as present the community density definitions that we use to analyze the results of the community detection.

1.1 Basic concepts

We consider simple undirected weighted networks where the edge between nodes i and j, e_{ij}, has a weight of $w_{ij} = w_{ji} \in [0, 1]$. Table 1 recalls a few quantities that will be used throughout this document.

N	Total number of nodes in the network
M	Total number of edges in the network
$\langle w \rangle$	Average weight of edges of the network
$\mathcal{C}$	Partition of the network into edge communities
$\mathcal{E}$	Set of all edges in the network ($\lvert\mathcal{E}\rvert = M$)
n_c	Number of nodes in community c
m_c	Number of edges in community c
$\langle w \rangle_c$	Average weight of edges in community c

Table 1: Definitions of quantities that are used in this report.

The similarity of two contiguous edges (sharing a node) is quantified by the Tanimoto coefficient, which is defined for edges e_{ik} and e_{kj} as

$$T(e_{ik}, e_{kj}) = \frac{\boldsymbol{a_i} \cdot \boldsymbol{a_j}}{|\boldsymbol{a_i}|^2 + |\boldsymbol{a_j}|^2 - \boldsymbol{a_i} \cdot \boldsymbol{a_j}} \tag{1}$$

where $\boldsymbol{a_i}$ is an N-element vector that contains information about how node i is connected to its neighbors. It is mostly the i-th row of the adjacency matrix, except for its i-th element that is defined according to two different definitions:

$$[\boldsymbol{a_i}]_j = w_{ij} + \frac{\delta_{ij}}{k_i} \sum_{l \in n(i)} w_{il} \qquad \text{Ahn } \textit{et al.} \tag{2}$$

$$[\boldsymbol{a_i}]_j = w_{ij} + \delta_{ij} \qquad \text{Kalinka } \textit{et al.} \tag{3}$$

where $n(i)$ is the set of nodes that are neighbors of node i.

1.2 Density of communities

We used the following definitions to compute the average density of communities[1][2]:

$$D^{(1)} = \frac{1}{M} \sum_c m_c \left(\left[\frac{m_c - (n_c - 1)}{n_c(n_c-1)/2 - (n_c - 1)} \right] \right) \tag{4}$$

$$D^{(2)} = \frac{1}{M'} \sum_c m_c \left(\left[\frac{m_c - (n_c - 1)}{n_c(n_c-1)/2 - (n_c - 1)} \right] \right) \tag{5}$$

$$D^{(3)} = \frac{1}{M\langle w \rangle} \sum_{c \in \mathcal{C}} m_c \langle w \rangle_c \left(\left[\frac{m_c - (n_c - 1)}{n_c(n_c-1)/2 - (n_c - 1)} \right] \right) \tag{6}$$

$$D^{(4)} = \frac{1}{M} \sum_{c \in \mathcal{C}} m_c \left(\langle w \rangle_c \left[\frac{m_c - (n_c - 1)}{n_c(n_c-1)/2 - (n_c - 1)} \right] \right) \tag{7}$$

$$D^{(5)} = \frac{1}{M\langle w \rangle} \sum_{c \in \mathcal{C}} m_c \langle w \rangle_c \left(\langle w \rangle_c \left[\frac{m_c - (n_c - 1)}{n_c(n_c-1)/2 - (n_c - 1)} \right] \right) \tag{8}$$

The first two definitions are the one introduced by Ahn *et al.*; the second one being an add-on introduced in their C++ code. They only differ in the way that the average is computed: the denominator in (5) explicitly excludes the communities of size two (*i.e.* links) that do not contribute to the average[3]. The last three definitions are attempts to generalize (4) to weighted networks where the weight of edges intervene either in the density of the communities themselves [(6)], or in the way individual densities are averaged over $\mathcal{C}$ [(7)], or both [(8)].

To have a better understanding of (6)–(8), let us recall that

$$\langle w \rangle_c = \frac{1}{m_c} \sum_{e_{ij} \in c} w_{ij} \tag{9}$$

$$\langle w \rangle = \frac{1}{M} \sum_{e_{ij} \in \mathcal{E}} w_{ij} \ . \tag{10}$$

[1] M' is the number of edges in the network that have *not* been assigned to communities of size two.

[2] The term in parentheses corresponds to the density of the community c itself.

[3] Communities of size $n_c = 2$ have $m_c = 1$ edge and therefore have a density of zero according to all definitions (4)–(8).

where $\mathcal{E}$ is the set of all edges in the network ($|\mathcal{E}| = M$). We thus conclude that $m_c\langle w\rangle_c$ and $M\langle w\rangle$ are the sum of the weights of edges in c and in the whole network, respectively.

2 The networks

Table 2 compares a few properties of the networks that we have used so far in this project.

	cutoff	N	M	$\langle k\rangle$	K
gBg	0.000	2400	1 565 816	1304.85	2 330 018 007
	0.010	2261	621 791	550.01	380 660 639
	0.100	2260	71 907	63.63	6 684 488
cBc	0.000	1495	1 115 538	1492.36	1 664 418 474
	0.100	n/a	n/a	n/a	n/a

Table 2: Properties of the gene-by-gene (gBg) and condition-by-condition (cBc) networks. A threshold of zero refers to the whole network (*i.e.* all edges). N, M, $\langle k\rangle$ and K are respectively the number of nodes, the number of edges, the average degree of nodes and the number of contiguous edges in the networks. Data that are indicated as non-available (n/a) have just not been computed yet.

3 Results

We present preliminary results obtained by following the suggestions from the document *Thoughts and comments I.*

3.1 Disclaimer

Let us recall that the computation of the average density is done considering *every* communities and therefore that an absolute maximum does not necessarily correspond to the best similarity threshold value to look at for significantly correlated communities. It must rather be considered as an indicator of how well connected are *every* communities on average. We therefore suggest to use the range corresponding to high density as a starting point to look for significant communities. Further investigation should then consider higher similarity threshold values since the tightly connected core of these communities is the last to be dismantled as "less-similar" edges are peeled off. In other words, the similarity threshold should be used as a zoom that allows to look at the communities from various points of view.

3.2 gBg_cutoff_0p01

We considered the *gene-by-gene* network in which edges with a weight lower than 0.01 have been removed. As shown in Tab. 2, approximately 40% of the original edges have thereby been spared. As a result, we may expect:

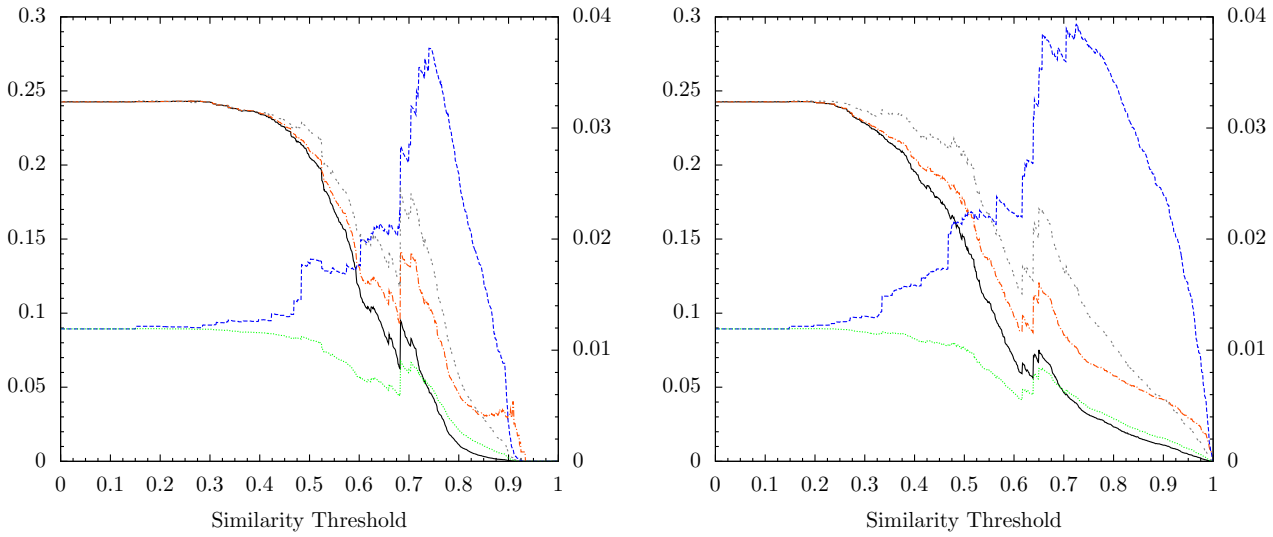
1. a shorter computation time, which is a function of $\langle k\rangle$ and K (both have been drastically reduced);

2. a more efficient and accurate community detection as the "noise" created by low-weighted edges has mostly been reduced. This is expected to "help" the algorithm to focus on the significantly correlated communities that we are looking for (which are likely to be composed of high-weighted edges).

In order to test this last hypothesis, as well as to compare the different definitions of density [(4)–(8)] and of the Tanimoto coefficient [(2)–(3)], we have run the algorithm on the `gBg_cutoff_0p01 network` using all combinations of those different definitions.

Fig. 1a and 1b compare definitions (4)–(8) as obtained on gBg_cutoff_0p01 network using definitions (2) and (3) respectively. Definitions (6) and (8) – the ones that have an extra $\langle w \rangle_c$ term that is not "canceled out" by the $\langle w \rangle$ term at the denominator – have been plotted using the right-hand side y axis to facilitate the comparison between all density definitions[4]. See the figure's caption for more details.

We first see that definitions (6) and (8) are not as affected by the high density of the network (at lower values of the similarity threshold) as the other definitions. We also see that all definitions peak approximately at the same values of similarity threshold, which is not too surprising as all definitions use a similar kernel (the term in brackets). Definition (8) is however slightly pushed towards higher similarity threshold values, which could reveal interesting features in the community structure. If this proves to be true, we would suggest definition (8) to be the weighted definition for community density that we are looking for, as it is so far the only one that has a peak that is easily optimizable.



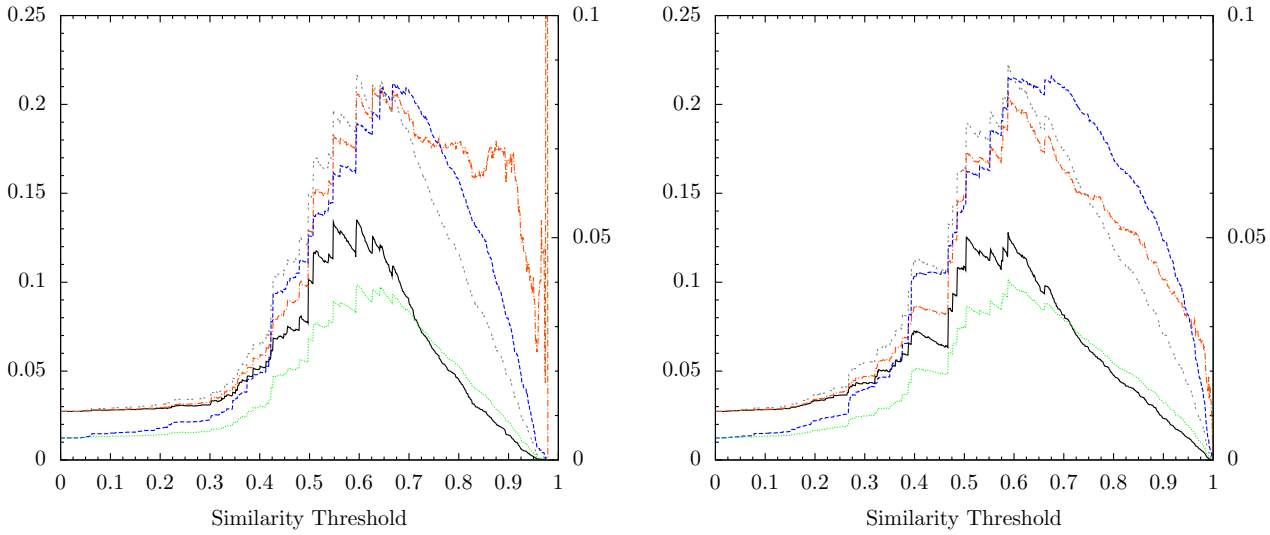
(a) Ahn *et al.* (b) Kalinka *et al.*

Figure 1: Average density of the communities detected on the `gBg_cutoff_0p01` network. The black, orange, gray, green and blue curves were computed using definitions (4)–(8), respectively. Definitions (4), (5) and (7) are plotted using the left-hand side y axis and definitions (6) and (8) are plotted using the right-hand side y axis. The Tanimoto coefficients were calculated using a) the definition of Ahn *et al.* [(2)] and b) the definition of Kalinka *et al.* [(3)].

[4]Let us recall that the absolute value of densities does not matter. It is rather its value compared to itself for different values of similarity threshold that is significant.

3.3 gBg_edgelist_01

We have run the community detection algorithm using the definition of Kalinka *et al.* (*i.e.* the one used in the `linkcomm` package) on the gene-by-gene network `gBg_edgelist_cutoff_01` in order to recover the communities that had been detected while in CSSS 2011. We also have run the community detection algorithm using the definition of Ahn *et al.* for the purpose of comparison of the methods. Fig. 2 present the results obtained.



(a) Ahn *et al.* (b) Kalinka *et al.*

Figure 2: Average density of the communities detected on the `gBg_edgelist_cutoff_01` network. The black, orange, gray, green and blue curves were computed using definitions (4)–(8), respectively. Definitions (4), (5) and (7) are plotted using the left-hand side y axis whereas definitions (6) and (8) are plotted using the right-hand side y axis. The Tanimoto coefficients were calculated using a) the definition of Ahn *et al.* [(2)] and b) the definition of Kalinka *et al.* [(3)].

Looking at Fig. 2a and 2b, we make the following remarks:

- Although peaks are not as sharp as the ones obtained for the `gBg_cutoff_0p01` network, we see that all density functions are high in the same region of the similarity threshold.
- All curves on Fig. 2b have a maximum at 0.588.
- The `linkcomm` package computes the "dissimilarity" which is defined as one minus the similarity. Hence one must take this into account when comparing the communities detected via the C++ algorithm with the ones detected with the `linkcomm` package.
- There are two sharp peaks on the orange curve in Fig. 2a at similarity thresholds of 0.966 and 0.975. While these two could be artifacts due to the fact that the denominator in (5) become very small, it could also mean that only very tightly – and perhaps significant – communities remain at these threshold values, and that all other communities were dismantled into single edge communities. When looking at the rough data, we believe that the peak at 0.966 might be significant as $M' = 230$ nontrivial communities were still detected, and that the peak at 0.975 might simply be an artifact as only $M' = 9$ nontrivial communities were still detected.

3.4 Recommendations

Following the recommendations made in the document *Thoughts and comments I*, we suggest the following tasks to be done:

1. Retreive the communities found using the `linkcomm` package in the `gBg_edgelist_cutoff_01` network using the definition of Kalinka *et al.* (Fig. 2b).
2. Try to find the same communities in the `gBg_edgelist_cutoff_01` network using the definition of Ahn *et al.* (Fig. 2a) to see whether the choice of the definition influences the detection of significantly correlated communities or not.
3. Use both definitions – or only one, depending on the results of step 2 – for the Tanimoto coefficient to look for the communities in the larger network `gBg_cutoff_0p01` to see if we can detect more complete communities.

Steps 1 and 2 allows us to fulfill the suggestions 1.2.1 and 1.2.2 of the document *Thoughts and comments I* while step 3 is a first attempt to tackle suggestion 1.2.3.