Expected Nodes: A Quality Function for the Detection of Link Communities

Noé Gaumont, François Queyroi, Clémence Magnien, Matthieu Latapy

Expected Nodes: a quality function for the detection of link communities

Noé Gaumont[1], François Queyroi[2], Clémence Magnien[1] and Matthieu Latapy[1]

[1] Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France
CNRS, UMR 7606, LIP6, F-75005, Paris, France. noe.gaumont@lip6.fr,
[2] Géographie-Cités, CNRS - Univ Paris 01/07.

Abstract. Many studies use community detection algorithms in order to understand complex networks. Most papers study node communities, *i.e.* groups of nodes, which may or may not overlap. A widely used measure to evaluate the quality of a community structure is the *modularity*. However, sometimes it is also relevant to study link partitions rather than node partitions. In order to evaluate a link partition, we propose a new quality function: *Expected Nodes*. Our function is based on the same inspiration as the modularity and compares, for a given link group, the number of incident nodes to the expected one. In this short note, we discuss the advantages and drawbacks of our quality function compared to other ones on synthetics graphs. We show that *Expected Nodes* is able to pass some fundamental sanity criteria and is the one that best identifies the most relevant partition in a more realistic context.

Keywords: complex networks, community detection, link partition, quality measure

1 Introduction

In the past years, complex networks were extensively studied because of the broad range of systems they can model, from protein-protein interactions to social networks. One question of interest is the detection of communities. Despite the important literature that covers the detection of classical, overlapping or even dynamic communities, most works focus on grouping nodes. On the other hand, the question of link communities has received less attention [4,1,8]. Intuitively, partitioning a network's links is very relevant in some contexts. For example, in a social network, most individuals belong to multiple communities such as families, friends, and co-workers, while the links between individuals usually exist for a dominant reason. In this context, a link community would be a group of interactions on one topic.

In this paper, we address the problem of evaluating the quality of a link partition. After a review of previous works (Section 2), we introduce a novel measure: *Expected Nodes* (Section 3). It is based on the assumption that a link community corresponds to less individuals than expected while its surroundings links correspond to more individuals than expected. We use several test cases (Section 4) to study how this measure behaves when compared to other quality functions.

2 Related Work

We introduce some notations used throughout this paper. Let $G = (V,E)$ ge a graph, $d(u)$ denotes the degree of node u in G. A link partition in k groups is noted $\mathscr{L} = (L_1, L_2, \ldots, L_k)$ with $L_i \subseteq E \ \forall i$, $L_i \cap L_j = \emptyset \ \forall i \neq j$ and $\bigcup_i L_i = E$. For a given link group $L \in \mathscr{L}$, let $V_{in}(L) = \{u \in V, \exists (u,v) \in L\}$ be the group of nodes inside L and $V_{out}(L) = \{u \in V \setminus V_{in}(L), (u,v) \in E \setminus L \wedge v \in V_{in}(L)\}$ be the nodes adjacent to L.

Ahn *et al.* [1] were among the first to propose a method to detect link communities. Their method *link clustering* is a hierarchical clustering method constructing a dendrogram by iteratively merging groups of links according to a similarity measure based on the Jaccard index. To decide where to cut the dendrogram in order to create a partition, they use a density based measure: the *partition density*. For a given link partition $\mathscr{L}$, the *partition density* is given by:

$$D(\mathscr{L}) = \frac{2}{|E|} \sum_{L \in \mathscr{L}, |L|>2} |L| \frac{|L| - (|V_{in}(L)| - 1)}{(|V_{in}(L)| - 1)(|V_{in}(L)| - 2)}. \tag{1}$$

However, the *partition density* cannot be easily generalized to weighted networks. An attempt in this direction has been made by Kim [5].

Evans *et al.* [4] propose three quality functions to evaluate link partitions. Their quality functions can be computed and optimized on the original graph but also on specific weighted line graphs (LG_1, LG_2, LG_3) using existing algorithms such as the *Louvain* method [3]. A line graph of an undirected graph is a graph where each node represents a link from the original graph and two nodes are connected if the corresponding links share a node.

To define these particular line graphs LG_1, LG_2 and LG_3, let B denote the incidence matrix of a network G: the elements $B_{i\alpha}$ of this $|V| \times |E|$ matrix are equal to 1 if link α is connected to node i and 0 otherwise. Matrices LG_1, LG_2 and LG_3 are defined as:

	$x=1$	$x=2$	$x=3$
$LG_x(\alpha,\beta)$	$B_{i\alpha}B_{i\beta}(1-\delta_{\alpha\beta})$	$\sum_{i \in V, d_G(i)>1} \frac{B_{i\alpha}B_{i\beta}}{d(i)-1}$	$\sum_{i,j \in V, d(i)d_G(j)>0} \frac{B_{i\alpha}A_{ij}B_{j\beta}}{d(i)d(j)}$

Let $k_x(\alpha) = \sum_\beta LG_x(\alpha,\beta)$ be α's weighted degree in LG_x and $W_x = \sum_{\alpha,\beta \in |E|} LG_x(\alpha,\beta)$. For $x \in \{1,2,3\}$, the quality function $Evans_x$ is:

$$Evans_x(\mathscr{L}) = \frac{1}{W_x} \sum_{L_i \in \mathscr{L}} \sum_{e_1,e_2 \in L_i^2} LG_x(e_1,e_2) - \frac{k_x(e_1)k_x(e_2)}{W}. \tag{2}$$

Kim *et al.* [6] explored the extension of the concept of Minimum Length Description introduced by Rosvall *et al.* [10] which is an information-theoretic framework. This extension directly considers link partitions. An advantage of their method is the ability to compare link and node partitions.

3 Our quality function: *Expected Nodes*

One commonly accepted assertion for node communities is: a community should have more internal connections than the expected number of connections in a random null model where no communi

ty structure exists. This assertion is at the core of the modularity introduced by Newman and Girvan [9]. In the same way, to evaluate a link community, we compare the number of nodes to its expected number of nodes. Like *modularity*, the measure can be decomposed, for each group L, into two components: internal quality and external quality. Like *modularity*, we use the configuration model [2] for a null model. In this model, the links are created by choosing random pairs of half-link (or stubs), each node having as many stubs as is degree in the original graph.

We start by describing the internal quality. Intuitively, a group of links L is a relevant community if it consists of a large number of links adjacent to few nodes, *i.e.* if V_{in} is small compared to what would be expected in the configuration model. By definition, a node is an internal node of L if one of its stubs (half-links) is in L. Therefore, to compute the expected number of internal nodes in the configuration model, we choose randomly $2|L|$ stubs among a total of $2|E|$ stubs. A node u has therefore $d(u)$ ways to be picked. The expected number of internal nodes for a given link group L, denoted by $\mu_G(|L|)$, is then:

$$\mu_G(|L|) = \sum_{u \in V} \mathbb{P}(u \textit{ picked at least once}) = \sum_{u \in V} 1 - \frac{\binom{2|E|-d(u)}{2|L|}}{\binom{2|E|}{2|L|}}. \tag{3}$$

Note that the function μ_G only depends on the degree sequence $\{d(v)\}_{v \in V}$. Note also that if $|L| = 1$, then $\mu_G(|L|) \leq 2$; this is because the configuration model allows self loops. A group L has a good internal quality if it has less internal nodes than expected. We therefore choose to define the internal quality function Q_{in} for a given group L as the variation between the actual number of internal nodes and its expectation:

$$Q_{in}(L) = \frac{\mu_G(|L|) - |V_{in}(L)|}{\mu_G(|L|)}. \tag{4}$$

With this definition, for a given $|L|$, the fewer nodes a group of links involves, the higher Q_{in} will be.

We now describe the external quality of a group L. The process to evaluate the neighbourhood $V_{out}(L)$ of a group L is similar to the process for the internal nodes. However in this case, we consider that L has a bad neighbourhood if it has fewer external nodes than expected. Indeed if there are many external links and few external nodes, these external links should be included in the community. Let $\bar{d}(L,u) = \sum_{v \in V} \mathbb{1}_{(u,v) \in E \setminus L}$ be the degree of u restricted to links not in L and $\bar{d}(L) = \sum_{u \in V_{in}(L)} \bar{d}(L,u)$. The expectation of the number of adjacent nodes is evaluated as the number of nodes that are picked when $\bar{d}(L)$ stubs are chosen randomly in the configuration model where the links of L have been removed. The corresponding degree sequence is $\{d_{G \setminus L}(u)\}_{u \in V}$ where $G \setminus L = (V, E \setminus L)$. Only one half link is chosen randomly because the other half has to remain attached to an internal node of L. Thus, we have the following equation:

$$\mathbb{E}[\bar{d}(L)] = \mu_{G \setminus L}(\bar{d(L)}/2). \tag{5}$$

Since we are interested in penalizing groups that have few external nodes, but do not consider that a group is particularly good if it has a large number of external nodes, we bound the external quality by 0:

$$Q_{ext}(L) = min\left(0, \frac{|V_{out}(L)| - \mu_{G\setminus L}(\bar{d}(L)/2)}{\mu_{G\setminus L}(\bar{d}(L)/2)}\right). \tag{6}$$

Finally, we define *Expected Nodes* for a group L as:

$$Q(L) = 2\frac{|L|Q_{in}(L) + |L_{out}|Q_{ext}(L)}{|L| + |L_{out}|}. \tag{7}$$

Notice that the trivial group containing all links has a null quality because Q_{in} and Q_{out} will be equal to 0. The other trivial decomposition where each link belongs to its own group has a negative quality. However, in some cases a group containing a single link might be the best choice. It is the case when the link is a bridge between dense groups. Finally, we define *Expected Nodes* for a given link partition $\mathscr{L}$ as the weighted sum of the quality of each group:

$$Q_G(\mathscr{L}) = \frac{\sum_{L\in\mathscr{L}} |L|Q(L)}{|E|}. \tag{8}$$

4 Comparison with existing methods

In order to study the relevance of *Expected Nodes*, we use two test cases.We also compare it to acknowledged quality functions: *partition density* [1] and the quality functions developed by Evans *et al.* [4] denoted by $Evans1$, $Evans2$ and $Evans3$.

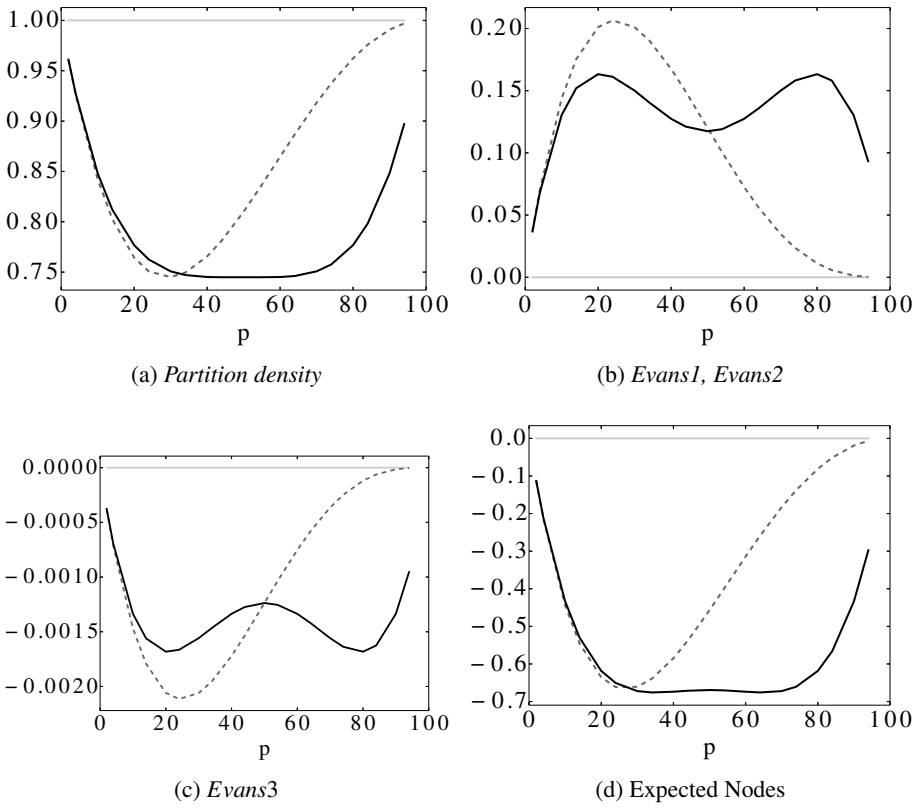
4.1 Complete graph

We start with a simple case in order to check that *Expected Nodes* satisfies some important and fundamental properties. We study a complete graph of 100 nodes (we obtained similar results with different sizes). On this graph, we define the trivial partition with one group containing all links, and two partitions families: one with two groups and one with three groups. Given a parameter $p < |V|$, let V' be a set of p nodes. Both partitions place all links in $V' \times V'$ in one group. The 2-groups partition places all other links in the second group. The 3-groups partition places all links in $V \times V \setminus V'$ in a second group and all remaining links in the third. These assignment rules are illustrated in Figure 1

As the graph is a single complete graph, the best solution is to capture only one group with all the links, *i.e.* the trivial partition should have the highest ranking. Figure 2 shows the results. For each value of p and each quality function, we present the values for the corresponding partitions in 2 and 3 groups and for the trivial partition. Surprisingly, quality functions $Evans1$ and $Evans2$ fail this simple test because they evaluate the 2- or 3-groups partitions as better than the trivial one. According to *partition density*, *Expected Nodes* and E_3, the trivial partition is best. The quality function $Evans3$ differs because of its small amplitude ($\approx 10^{-3}$).

(a) (b)

Fig. 1: Two different link partitions of a complete graph with $p = 5$: (a) in two link groups and (b) in three link groups. The dark nodes corresponds to V' and the color of a link corresponds to its group.



(a) *Partition density* (b) *Evans1, Evans2*

(c) *Evans*3 (d) Expected Nodes

Fig. 2: Evaluation of 5 quality functions on a complete graph of 100 nodes for 3 different partitions. The tested partitions are presented in Section 4.1. The results for quality functions *Evans*1 and *Evans*2 are identical. The gray line, black line and dashed line represent respectively the 1-group partition, the 2-groups partition and the 3-groups partition.

4.2 Overlapping LFR benchmark

We now discuss results obtained by comparing the quality functions on random networks with a known community structure. To the best of our knowledge, there is no graph generator based on link partitions. We use the benchmark proposed by Lancichinetti *et al.* [7] which generates graphs based on a known node cover. We introduce two transformations of this overlapping community structure into link partitions denoted by TA and TB (see Figure 3). Given $u, v \in V$, let $C_{u,v}$ denote the intersection between the communities of u and v in the node cover and $U_{u,v}$ their union. We define the group of a link $(u,v) \in E$ in the partitions as follows:

intra-community if $|C_{u,v}| = 1$ then (u,v) is in community $C_{u,v}$;
inter-community if $|C_{u,v}| = 0$ then in TA, (u,v) belongs to its own community. In TB it belongs to community $U_{u,v}$ which contains all links (u',v') such that $U_{u',v'} = U_{u,v}$;
overlapping if $|C_{u,v}| > 1$ then (u,v)'s community is chosen randomly in $C_{u,v}$.

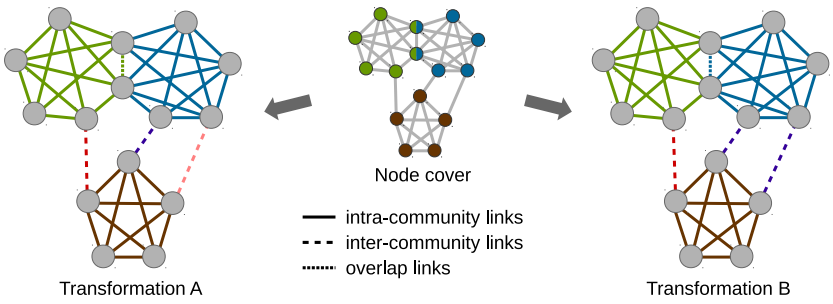


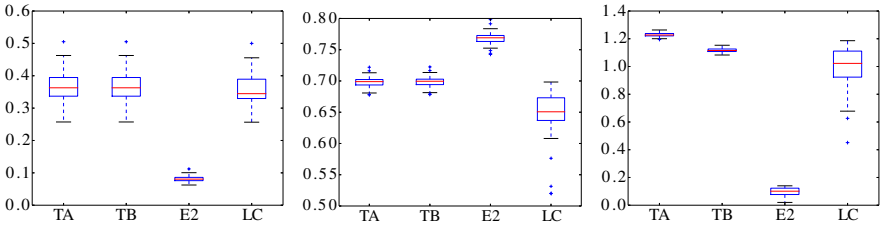
Fig. 3: Construction of TA and TB from a node cover. Link colours denote groups.

We describe the results averaged over 30 graph generations with 500 nodes, an average degree of 25, 10 overlapping nodes and a mixing parameter of 0.1[3]. There are on average 5620 intra-community links, 625 inter-community links and only 5 overlapping links. For each generation, the partition TA, TB, the partition LC found by *link clustering* [1] and the partition $E2$ found by the second method of Evans *et al.* [4] (based on the optimization of $Evans2$)[4] are evaluated using *Partition Density*, $Evans2$ and *Expected Nodes*.

In TA (resp. TB), there are 650 (resp. 70) groups on average. Manual investigations show that the $E2$ partitions are very close to the ground truth (TA or TB) if inter-community links are not considered. Indeed in $E2$, inter-community links are randomly distributed among adjacent larger link communities. The LC partitions contain

[3] Remaining parameters with original notations: $k_{max} = 50$, $t_1 = -2$, $t_2 = -1$, $C_{min} = 20$, $C_{max} = 100$.

[4] The results are similar for the algorithms using $E1$ and $E3$.



(a) Partition density (b) Evans2 (c) Expected Nodes

Fig. 4: Boxplots of the three quality functions values for the different link partitions. The box shows lower and upper quartiles and the median. The whiskers extend to 1.5 time the interquartile range. Flier points are those past the end of the whiskers.

720 groups on average and intra-community links are split into many small groups. Notice that neither TA nor TB get the best evaluation according to *Evans*2 and *Partition density* even though they are considered as ground truth.
The following observations can be made. First, *Expected Nodes* (Fig. 4c) behaves differently than both other measures (Fig. 4a and 4b). This shows that our measure brings something new to the picture. Moreover, its values are usually higher for TA and TB than for the partitions found using the two algorithms, which corresponds to our expectations. Second, the *Expected Nodes* values are significantly different for TA and TB. It is not the case for quality functions $E2$ and *Partition density*. Indeed, external links between the same community are likely to form a group of isolated links in TB. This situation is highly penalized by our measure. It also explains why *Expected Nodes* evaluates LC partitions better than $E2$ partitions. For those reasons, we believe that maximizing *Expected Nodes* would result in partitions close to TA in this benchmark.

4.3 Conclusion

In this paper, we propose a new quality function, *Expected Nodes*, to evaluate the quality of a link partition of a graph[5]. It compares the number of nodes adjacent to a link group to its expectation, in the same way as the modularity evaluates the relevance of a node group by comparing the number of adjacent links to its expected value. To show the relevance of *Expected Nodes*, we compared it to existing quality functions. The main perspective of our work is to design an algorithm for maximizing *Expected Nodes* in order to detect relevant link partitions. More detailed comparisons between quality functions may also be performed. For instance, it would be interesting to evaluate their behaviour to detect whether they are likely to present local maxima such as the one observed in Figure 2c or not.
Acknowledgements : This research was supported by a DGA-MRIS scholarship, by a grant from the French program "*PIA – Usages, services et contenus innovants*" under grant number $O18062-44430$ and by the CODDDE project ANR-13-CORD-0017-01.

[5]The code used to compute each quality function is available: `https://github.com/ksadorf/ExpectedNodes`

References

1. Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, Aug. 2010.
2. E. A. Bender and E. Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, May 1978.
3. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, Oct. 2008.
4. T. S. Evans and R. Lambiotte. Line graphs, link partitions, and overlapping communities. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 80(1):016105, July 2009.
5. S. Kim. *Community Detection in Directed Networks and its Application to Analysis of Social Networks*. PhD thesis, Ohio State University, 2014.
6. Y. Kim and H. Jeong. Map equation for link communities. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 84(2):026110, Aug. 2011.
7. A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 80:016118, 2009.
8. S. Lim, S. Ryu, S. Kwon, K. Jung, and J.-G. Lee. LinkSCAN*: Overlapping community detection using the link-space transformation. In *2014 IEEE 30th International Conference on Data Engineering*, pages 292–303. IEEE, Mar. 2014.
9. M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 69, 2004.
10. M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4):1118–23, Jan. 2008.