On the Evaluation Potential of Quality Functions in Community Detection for Different Contexts

Jean Creusefond[1(✉)], Thomas Largillier[1], and Sylvain Peyronnet[2]

[1] Normandy University, Caen, France
jean.creusefond@unicaen.fr
[2] Qwant and ix-labs, Rouen, France

Abstract. Due to nowadays networks' sizes, the evaluation of a community detection algorithm can only be done using quality functions. These functions measure different networks/graphs structural properties, each of them corresponding to a different definition of a community. Since there exists many definitions for a community, choosing a quality function may be a difficult task, even if the networks' statistics/origins can give some clues about which one to choose.

In this paper, we apply a general methodology to identify different `contexts`, *i.e.* groups of graphs where the quality functions behave similarly. In these `contexts` we identify the best quality functions, *i.e.* quality functions whose results are consistent with expectations from real life applications.

Keywords: Quality functions · Social networks · Community detection

1 Introduction

Every community detection algorithm is justified by the search for particular substructures, *i.e.* communities, defined by a particular purpose in a particular network. This combination of structured data and purpose makes the field complex and fuzzy, but drives research to unravel the different meanings that the word "community" bears.

As a result, a large number of desirable properties of communities have been discovered. To measure them, many works aimed at designing functions quantifying these properties in order to evaluate the goodness of a community. Called quality functions, these mathematical tools are not only useful for evaluation purposes but can also be used in greedy algorithms as community detection methods directly.

However, evaluating an algorithm may be difficult because it implies choosing between quality functions that often output contradictory results. The structural properties of the network and of the communities being looked for may strongly differ from one case to the other. It is then of the utmost importance to identify the right quality function for each graph. In order to do that we define the notion of `context` which is a group of graphs where quality functions behave similarly.

A. Wierzbicki et al. (Eds.): NetSci-X 2016, LNCS 9564, pp. 111–125, 2016.
DOI: 10.1007/978-3-319-28361-6_9

One then only needs to identify the right quality function for a `context` and the means to identify which `context` a graph is part of.

In this paper we identify some `contexts` for community detection, and select quality functions that feature behavior that is coherent with real-world data. To achieve this goal, we compare 10 functions from relatively recent literature, using 10 datasets featuring ground-truth, 7 community detection algorithms and 2 extrinsic evaluation functions. We look at the correlation between quality functions and real-world data: do they rank higher clusterings that are close to the ground-truth, and conversely? We then identify `contexts` when quality functions rank different graphs in the same way.

2 Related Work

The rise of community detection as a research field has inevitably given birth to a variety of works on meta analysis. They feature a wide range of methods, but all of them are aimed to identify quality functions with desirable properties.

Van Laarhoven and Marchiori [28] designed six axioms that qualify intuitive good behavior of quality functions. They show that modularity does not satisfy two of them, partly because of the resolution limit [14].

Yang and Leskovec [30] studied 12 quality functions that could be applied at cluster level. They classified them into four groups depending on how they were correlated when applied to real-world graphs, and these groups corresponded to the measured structural property. They designed "goodness metrics" that measure only one property of a cluster and compared how the quality functions fared in order to identify what property were measured by which function.

Almeida *et al.* [2] compared the result of 5 quality functions when applied to 5 real-world graphs. They applied 4 parameterized algorithms on these networks and changed the parameters to get different number of communities. They observed that some metrics have the tendency to favor bigger clusters while others favor the opposite.

Our approach differs from previous works by the scale and purpose of our work: to the best of our knowledge, we are the first to focus on the identification of `contexts` for quality functions. Chakraborty *et al.* [9] have already applied part of this methodology in the context of community detection in order to experimentally demonstrate the efficiency of the quality function they proposed.

3 Quality Functions

Throughout the rest of this paper we use the following notations.

A quality function is an application $f(G, \mathcal{C}) \to \mathbb{R}$, whose purpose is to quantify the quality of a clustering on a graph. For brevity, we omit the graph input. Note that quality functions are different from comparison methods, the latter comparing two clusterings.

In order to ease comparisons, we normalize some quality functions. We categorize the functions depending on the locality of information they use. We

General	
$G = (V, E)$	Undirected graph (set of vertices, edges)
n, m	# of vertices (= nodes) and edges
$N_{v \in V}$	Set of neighbors of a node v
$k_{v \in V}$	# of neighbors (degree) of a node v
k_m	Median degree
Set-specific	
$m(S \subseteq V)$	# of internal edges of S
$m(S \subseteq V, S' \subseteq V)$	# of edges with one end in S and another in S'
$N_{v \in V, S \subseteq V}$	Internal neighborhood of v in S
$k_{S \subseteq V}$	Size of a cluster S
$Vol(S \subseteq V)$	Volume of a cluster S (sum of the degrees of the vertices)
$diam(S \subseteq V)$	Internal diameter of a cluster S
Clusterings	
$\mathcal{C}, \mathcal{L}$	Clustering, set of sets of nodes whose union is V
$C(v \in V), L(v \in V)$	Set of clusters in which a node v belongs to in $\mathcal{C}/\mathcal{L}$

identify three classes of locality: vertex-level, community-level and graph-level. The formula for each quality function can be found in Table 1.

Vertex-Level Quality Functions. Compute a quality for every node in the graph and output the average as the total quality of the clustering on the graph. Let $v \in V$ be the considered node, and $C \in \mathcal{C}$ be the community of v.

The **Local internal clustering coefficient** [29] (called clustering coefficient from now on) of a node is the probability that two of his neighbors that are in the same community are also neighbors. The clustering property of communities is actually one of the most well-known in the field, and is explained by the construction of social networks by homophily.

This property is included in **Permanence** [9], where it is combined with a notion of equilibrium for the nodes concerning their membership to their community. A node has a lower Permanence if there is another community than its own that highly attracts it, *i. e.* to which it is very connected compared to its connection to its community.

The **Flake-ODF** [13] compares internal to external degree. It is similar to the **Fraction Over Median Degree** [30] (FOMD), that compares internal degree and the median degree in the whole graph.

Community-Level Quality Functions. Compute a score for each cluster and output the sum as the quality of the clustering.

The **Conductance** [17] and the **Cut-ratio** are concerned with the external connectivity of the community. The Cut-ratio normalizes it with the number of potential edges between the individuals of the community and the remainder of

the network. On the other hand, the Conductance is normalized by the same potential number of edges but takes into account the degrees in the community (few edges may reach a community of consisting of a few nodes with small degree). We weight these local measures with the size of the community so that each vertex in the networks has the same level of participation in the measure.

The **Compactness** [11] measures the potential speed of a diffusion process in a community. Starting from the most eccentric node, the function captures the number of edges reached per time step by a perfect transmission of information. The underlying model defines community as a group of people within which communication quickly reaches everyone.

Modularity [21] is the difference between the number of internal edges of the community versus the expected number of edges. This expectancy is expressed using the configuration model, a graph model guaranteeing the same degree distribution as the original one but with randomized edges. Assuming that this model ignores community structure, a high difference between expectancy and reality would indicate an abnormal density, *ergo* community structure.

Graph-Level Quality Functions. Output the score of the whole graph. **Surprise** [1] (in its asymptotical approximation [26]) and **Significance** [27] are based on the computation of an asymmetric difference, the Kullback-Leibler divergence, between two-points probability distributions (only one event and its complement). Considering x/y indifferently as one of the two probabilities featured in the reference/non-reference distribution, the divergence is: $D(x||y) = x\log(x/y) + (1-x)\log((1-x)/(1-y))$.

The reference distribution of Surprise models the probability that an edge is internal to a community, and the non-reference is the event that a couple of nodes are inside the same community. Significance features one reference distribution per community that corresponds to the event that a random couple of nodes inside the same community are linked by an edge. The non-reference distribution is the same value for the whole graph.

4 Networks with Ground-Truth

To identify `contexts` for community detection, we need some real-life information on what a community actually is. We therefore pulled 10 networks with known community structure from literature. To compare them with the algorithms that classify all nodes, vertices with no ground-truth communities are removed and only the largest connected component is considered. We note $\rightarrow$ for directed networks and $\textcircled{\scriptsize\cap}$ for overlapping communities.

Collaboration Networks (cf Table 2). The networks represent people working together in certain organizations. They have a strong underlying bipartite structure.

The Computer Science (CS) network comes from the same source as the DBLP network, but it features only computer scientists and a different kind of

Table 1. Quality functions

Name	Function
Local clustering coefficient	$f_{clus}(v,C) = \frac{2 * \lvert\{u \in N_{v,C}, w \in N_{v,C} \backslash \{u\}, (u,w) \in E\}\rvert}{\lvert N_{v,C}\rvert(\lvert N_{v,C}\rvert - 1)}$
Permanence	$f_{perm}(v,C) = \frac{m(v,C)}{max_{C' \in \mathcal{C} \backslash \{C\}}(m(v,C')) \times k_v} + f_{clus}(v,C) - 1$
1-Flake-ODF	$f_{flak}(v,C) = \begin{cases} 1 & \text{when } m(v,C) > m(v, V \backslash C) \\ 0 & \text{otherwise} \end{cases}$
FOMD	$f_{FOMD}(v,C) = \begin{cases} 1 & \text{when } m(v,C) > d_m \\ 0 & \text{otherwise} \end{cases}$
1-Cut ratio	$f_{cut}(C) = \left(1 - \frac{m(c, V \backslash C)}{k_C(n - k_C)}\right) \times \frac{k_C}{n}$
1-Conductance	$f_{cond}(C) = \left(1 - \frac{m(v, V \backslash C)}{Vol(C)}\right) \times \frac{k_C}{n}$
Compactness	$f_{comp}(C) = \frac{m(C)}{diam(C)}$
Modularity	$f_{mod}(C) = \frac{m(C)}{m} - \left(\frac{Vol(C)}{2m}\right)^2$
Surprise	$f_{surp}(\mathcal{C}) = D\left(\frac{\sum_{C \in \mathcal{C}} m(C)}{m} \middle\| \frac{\sum_{C \in \mathcal{C}} \binom{k_C}{2}}{\binom{n}{2}}\right)$
Significance	$f_{sign}(\mathcal{C}) = \sum_{C \in \mathcal{C}} \binom{k_C}{2} D\left(\frac{m(C)}{\binom{k_C}{2}} \middle\| \frac{m}{\binom{n}{2}}\right)$

Table 2. Collaboration networks

Name	n	m	Nodes	Edges	Communities
DBLP[a] [30]	129981	332595	authors	co-authorships	publication venues ⓒ
CS [7,8]	400657	1428030	authors	co-authorships	publication domains ⓒ
Actors (imdb)[b] [5]	124414	20489642	actors	co-appearances	movies ⓒ
Github[b,c]	39845	22277795	developers	co-contributions	projects ⓒ

[a] http://snap.stanford.edu/data/
[b] http://konect.uni-koblenz.de
[c] https://github.com/blog/466-the-2009-github-contest

ground-truth. Furthermore, the actors and github networks are constructed from bipartite graphs, and therefore form cliques inside of the communities.

Online Social Networks (OSNs) (cf Table 3). Most of these networks are originally directed but due to the high reciprocity the original authors considered safe to set all links as undirected.

Table 3. Online social networks

Name	n	m	Nodes	Edges	Communities
LiveJournal[a] [30]	1143395	16880773	bloggers	following →	explicit groups ◎
Youtube[a] [30]	51204	317393	youtubers	following →	explicit groups ◎
Flickr [20]	368285	11915549	users	following →	explicit groups ◎

[a]http://snap.stanford.edu/data/

Table 4. Social-related networks

Name	n	m	Nodes	Edges	Communities
Amazon[a] [30]	147510	267135	products	frequent co-purchases	categories
Football [15]	115	613	football teams	>1 one disputed match	divisions
Cora[b] [25]	23165	89.156	scientific papers	citations →	categories

[a]http://snap.stanford.edu/data/
[b]http://konect.uni-koblenz.de

Social-Related Networks (cf Table 4). Nodes in these networks do not represent people, but their connections are created by social interaction.

Artificial Benchmarks. We use the Lancichinetti-Fortunato-Radicchi (LFR) [19] benchmark as a validation method for our methodology.

On this benchmark, we may chose the number of nodes, the average degree ($\hat{k}$), the maximum degree (k_{max}), the mixing parameter (μ), the coefficients of the power laws of degree and community size distributions (respectively t_1 and t_2), the average clustering coefficient ($\hat{cc}$), the number of nodes belonging to multiple communities (on) and the number of communities they belong to (om).

Table 5. The parameters of the five classes of synthetic LFR networks

Name	n	$\hat{k}$	k_{max}	μ	t_1	t_2	$\hat{cc}$	on	om
LFRa	10 000	50	1 000	0.1	2.5	2.5	0.2	8 000	4
LFRb	100 000	50	2 500	0.1	2.5	2.5	0.2	8 000	4
LFRc	10 000	100	500	0.4	2.1	2.0	0.1	8 000	5
LFRd	10 000	50	1 000	0.1	2.5	2.5	0.2	0	0
LFRe	10 000	100	500	0.4	2.1	2.0	0.1	0	0

As presented in Table 5, we have 5 classes of networks. The a class represents a standard social network, with common values for each parameter. We note that the mixing parameter is quite low (the communities should be well-cut) and the communities are overlapping. The b class is the same as the a class but with ten times more nodes. The c class is however completely different, with all its parameters changed except size (but it is still overlapping). The $d(e)$ class is the same as the $a(c)$ class but without any overlapping community.

5 Comparison Methods

A comparison method (or extrinsic clustering evaluation metric [3]) is an application $f(\mathcal{C}, \mathcal{L}) \rightarrow [-1; 1]$, whose purpose is to evaluate the closeness of two clusterings.

The **Normalized Mutual Information (NMI)** measures the quantity of information gained by the knowledge of one clustering compared to the other. The version that we use was introduced by Lancichinetti *et al.* [18].

The **F-BCubed (fb3)** [4] precision measures for each element e the proportion of its associates (*e.g.* individuals that are in the same cluster) in $\mathcal{C}$ that are still its associates in $\mathcal{L}$, and takes the average among all e. Amigó *et al.* [3] extended this metric for overlapping clustering, taking into account the number of clusters in common that e and its associates have. They define BCubed overlapping precision and recall as follows:

$$prec(C, L) = Avg_e \left[Avg_{\substack{e' \\ C(e) \cap C(e') \neq \emptyset}} \left(\frac{min(|C(e) \cap C(e')|, |L(e) \cap L(e')|)}{|C(e) \cap C(e')|} \right) \right] \quad (1)$$

$$recall(C, L) = prec(L, C) \quad (2)$$

$$F\text{-}BCubed(C, L) = \frac{1}{\frac{1}{2 * prec(C, L)} + \frac{1}{2 * recall(C, L)}} \quad (3)$$

Amigó *et al.* [3] also gave an extensive comparison of evaluation metrics by designing intuitive properties of goodness. Their conclusion was that the F-BCubed measure satisfied all of them, while the other common metrics fail at least on of these axioms.

6 Experimental Setup

In this section, we describe our experiments. We first cover the methodology, then present the community detection algorithms used to generate clusterings and finally tools we used to keep tractable the number of operations.

6.1 Methodology

The methodology has two goals: to identify `contexts` in which quality functions behave in the same way, and to identify the best quality functions for each `context`. For each graph with ground-truth communities (cf Sect. 4), we execute the following steps:

1. Apply various community detections methods on the base graph (cf Sect. 6.2).
2. Compute quality functions over the resulting clusterings (cf Sect. 3).
3. Compare the communities found to the ground-truth, creating a gold standard value for each clustering (cf Sect. 5).

4. Compare for each graph the ranking of the clusterings given by the gold standard value to the ranking of clusterings measured by quality functions with Spearman's coefficient. For each graph, each quality now have a score.
5. For each couple of graphs, compute the correlation of the previous scores using Spearman's coefficient.

The rationale behind step 4 is that a quality function fits a ground-truth if the clusterings that are the closest to the ground-truth are highly ranked with the quality, and conversely. Therefore, at this step we can conclude which quality function is the best for each graph. We also need to go through step 5 in order to identify `contexts`: the graphs are compared on their ranking from the quality functions, and `contexts` may be identified as sets of graphs that are highly correlated.

6.2 Community Detection Algorithms

Since we consider large graphs, we decided to use community detection algorithms that have sub quadratic time and space complexity. We chose several methods, based on their availability, efficiency, originality and/or spread.

We classify the algorithms we use in three groups:

- Modularity optimization: Louvain [6], Clauset [10].
- Random walks: MCL [12], Infomap [23].
- Heuristics: LexDFS [11], 3-core [24], label propagation [22].

6.3 Computation Time Management

Three kinds of measures are computation-heavy in our experimental setup: triangle computation, diameter and fb3. Fb3 needs $O(|C|^2)$ operations to compute the values for the community C. f_{clus} and f_{perm} need the computation of all internal triangles, which is very demanding for highly clustered graphs.

We therefore sample our dataset and average these two values over the sample. We use the Hoeffeding bound [16] (our samples are *i.i.d* and in the $[0, 1]$ interval) to get the number of samples t needed ensure that there is a small probability p that the error resulting in our sampling is not bounded by ϵ.

$$P(|\overline{X} - E[X]| < \epsilon) = p \geq 2e^{-2n\epsilon^2} \Leftrightarrow n \geq \frac{ln(p/2)}{-2\epsilon^2} \tag{4}$$

We use 5000 samples, meaning that $p \leq 5\,\%$ and $\epsilon \leq 0.02$. Of course, the bound is a worst-case: in practice, we observe errors of about 10^{-4}, which is too small to disturb the rankings.

The diameter computation, needed by f_{comp}, is in $O(|C|^2)$. We use the standard approximate algorithm based on two BFSs to compute it in near-linear time. The first BFS starts at a random point of the community, and the last node visited by this BFS is used as the origin of another BFS. This heuristic searches for an eccentric point which is likely to feature at the end of a maximum-distance path.

Due to the process of ranking quality functions and comparison methods, even bounded errors may have unbounded impact on the results if the approximated values are too close to each other. On top of that, some of the community detection algorithms make nondeterministic choices, which implies an incontrollable potential difference in results. To gain confidence that the randomness of the processes involved does not influence the results too much, we ran the whole process multiple times. We obtained very close results in every run.

7 Experimental Results

7.1 Correlations in LFR

We first study the results of the methodology when applied to LFR graphs. In order to assess its stability, we create three benchmark graphs from different random seeds for each class of LFR graphs described in Sect. 4. In order to judge behavioral similarity of quality functions between graphs, we compute Spearman's coefficient of each couple of graphs (as presented in Sect. 6.1, step 5) and report the results in Table 6 (resp. Table 7) for NMI (resp. for fb3).

In Table 6, we see that quality functions of the same class behave in the same way when compared to NMI. However, this positive view is tarnished by some exceptions: c1 seems to relate more to graphs of the a class than from its own class, and the same can be said from e3. We assume that these exceptions are due to the random nature of the generative model, which might produce networks that have some structural properties that vary significantly enough to disturb comparison with NMI.

Table 6. The correlation between the ranking of quality functions (with NMI ranking) for synthetic graphs (A colored version is available on the authors' webpage).

file\file	a1	a2	a3	b1	b2	b3	c1	c2	c3	d1	d2	d3	e1	e2	e3
a1	-	1.00	1.00	0.98	0.99	0.45	0.40	-0.23	-0.09	0.53	0.53	0.53	0.69	0.31	0.95
a2	-	-	1.00	0.98	0.99	0.45	0.40	-0.23	-0.09	0.53	0.53	0.53	0.69	0.31	0.95
a3	-	-	-	0.98	0.99	0.45	0.43	-0.22	-0.07	0.53	0.53	0.53	0.69	0.31	0.94
b1	-	-	-	-	0.99	0.48	0.35	-0.15	0.01	0.53	0.53	0.53	0.69	0.31	0.92
b2	-	-	-	-	-	0.46	0.34	-0.21	-0.07	0.53	0.53	0.53	0.69	0.31	0.94
b3	-	-	-	-	-	-	0.29	0.03	0.21	0.15	0.15	0.15	0.67	0.56	0.42
c1	-	-	-	-	-	-	-	0.07	0.15	0.27	0.27	0.27	0.30	0.25	0.41
c2	-	-	-	-	-	-	-	-	0.90	0.34	0.34	0.34	-0.12	0.05	-0.32
c3	-	-	-	-	-	-	-	-	-	0.22	0.22	0.22	-0.03	0.20	-0.16
d1	-	-	-	-	-	-	-	-	-	-	1.00	1.00	0.47	0.12	0.47
d2	-	-	-	-	-	-	-	-	-	-	-	1.00	0.47	0.12	0.47
d3	-	-	-	-	-	-	-	-	-	-	-	-	0.47	0.12	0.47
e1	-	-	-	-	-	-	-	-	-	-	-	-	-	0.76	0.70
e2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.35
e3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 7. The correlation between the ranking of quality functions (with FB3 ranking) for synthetic graphs

file\file	a1	a2	a3	b1	b2	b3	c1	c2	c3	d1	d2	d3	e1	e2	e3
a1	-	0.99	1.00	0.95	0.94	0.96	-0.23	-0.50	-0.44	0.72	0.72	0.72	0.89	0.89	0.88
a2	-	-	0.99	0.94	0.93	0.94	-0.26	-0.48	-0.42	0.70	0.70	0.70	0.89	0.87	0.88
a3	-	-	-	0.95	0.94	0.96	-0.23	-0.50	-0.44	0.72	0.72	0.72	0.89	0.89	0.88
b1	-	-	-	-	1.00	1.00	-0.26	-0.49	-0.44	0.80	0.80	0.80	0.90	0.92	0.93
b2	-	-	-	-	-	1.00	-0.27	-0.48	-0.43	0.80	0.80	0.80	0.90	0.92	0.93
b3	-	-	-	-	-	-	-0.25	-0.49	-0.43	0.80	0.80	0.80	0.90	0.91	0.92
c1	-	-	-	-	-	-	-	0.62	0.76	-0.27	-0.27	-0.27	-0.42	-0.38	-0.37
c2	-	-	-	-	-	-	-	-	0.90	-0.26	-0.26	-0.26	-0.49	-0.51	-0.37
c3	-	-	-	-	-	-	-	-	-	-0.31	-0.31	-0.31	-0.47	-0.47	-0.39
d1	-	-	-	-	-	-	-	-	-	-	1.00	1.00	0.75	0.78	0.79
d2	-	-	-	-	-	-	-	-	-	-	-	1.00	0.75	0.78	0.79
d3	-	-	-	-	-	-	-	-	-	-	-	-	0.75	0.78	0.79
e1	-	-	-	-	-	-	-	-	-	-	-	-	-	0.983	0.963
e2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.96
e3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

It was expected that the a class would be rated in the same way as the b and d class, and would be different from the other two. If the correct similarities are observed, surprisingly, the e class seems to behave similarly to the a class, and this is even clearer with the fb3 measure (cf Table 7). This is probably because the distribution difference and the mixing parameter have less influence in the structural properties of the network than the overlapping nature. We conclude that the comparison with NMI is globally efficient, but it is very sensitive to noise and overlapping difference.

In Table 7, we see that the comparison with fb3 is much more clear-cut than the one with NMI: there is no value between -0.2 and 0.6, which would indicate medium to weak correlations. It is also very clear that the c class is considered as differently ranked for fb3 than the other ones. As stated above, the fb3 measure does not identify the model difference in the generation of the e class.

We note that the c1 and the e3 networks that did not behave like the others when compared with NMI measure behave in the same way when looking at the fb3 measure. We conclude that comparing networks through the measure is less sensitive than NMI to random variations due to network generation processes, the downside being that it may show resemblance between two networks that are actually very different.

7.2 Correlations in Real World Data

Just as with the LFR benchmark, we start by identifying groups of networks where quality functions behave approximately in the same way. Unlike LFR, the only classification available for these networks is their representation of reality, and not the underlying model.

Table 8. Spearman's coefficient of the rows of Table 10 (NMI, Real-world)

file\file	CS	actors	amazon	cora	dblp	flickr	football	github	lj	youtube
CS	-	0.923	0.281	0.972	0.302	0.103	0.245	0.014	0.253	-0.187
actors	-	-	0.264	0.899	0.276	0.168	0.318	0.105	0.262	-0.077
amazon	-	-	-	0.280	0.965	-0.231	0.523	-0.033	-0.269	-0.336
cora	-	-	-	-	0.327	0.115	0.213	0.052	0.334	-0.191
dblp	-	-	-	-	-	-0.238	0.453	0.031	-0.241	-0.357
flickr	-	-	-	-	-	-	0.180	0.367	0.808	0.759
football	-	-	-	-	-	-	-	0.350	-0.191	0.117
github	-	-	-	-	-	-	-	-	0.329	0.549
lj	-	-	-	-	-	-	-	-	-	0.587
youtube	-	-	-	-	-	-	-	-	-	-

Real-life data are less clear-cut than controlled benchmark networks. However, we see in Tables 8 and 9 that the connections (cora, CS) and (lj, youtube, flickr) appear with both comparison methods as high, which means that these networks are consistently close with the ranking of their ground-truth. This observation is consistent with our knowledge of these networks. Cora and CS both correspond to scientific publication and their ground-truthes both correspond to publication domains. Interestingly, neither the overlapping nature of CS nor the size difference seem to affect this outcome, which comforts us in the robustness of the method. Youtube, flickr and lj have similar connection (someone follows someone) and ground-truth (explicit membership) mechanics. The other correlation relationships differ given the considered comparison method.

NMI: We notice first that the tuple (cora, CS) is extended to **(cora, CS, actors)**, which brings another collaboration network close to the first two. We note, however, that the github network is not correlated with them. We notice that the structural difference with github, where an individual belongs to more

Table 9. Spearman's coefficient of the rows of Table 11 (FB3, Real-world)

file\file	CS	actors	amazon	cora	dblp	flickr	football	github	lj	youtube
CS	-	-0.070	0.920	0.970	0.502	0.224	-0.434	-0.344	-0.351	-0.035
actors	-	-	-0.052	-0.157	0.472	-0.091	0.776	0.774	0.434	0.227
amazon	-	-	-	0.935	0.411	0.189	-0.455	-0.378	-0.316	-0.105
cora	-	-	-	-	0.409	0.358	-0.456	-0.381	-0.266	0.040
dblp	-	-	-	-	-	0.250	0.163	0.187	0.143	0.456
flickr	-	-	-	-	-	-	0.154	0.156	0.533	0.790
football	-	-	-	-	-	-	-	0.911	0.719	0.497
github	-	-	-	-	-	-	-	-	0.654	0.414
lj	-	-	-	-	-	-	-	-	-	0.760
youtube	-	-	-	-	-	-	-	-	-	-

Table 10. Spearman's coefficient of the NMI (ground truth, algorithms) compared to the results of quality functions. Real-world dataset

file\quality	cc	fb3	mod	nmi	perm	sign	cond	FOMD	comp	cut_ratio	f-odf	sur
CS	0.00	0.82	-0.25	1.00	0.00	-0.14	0.14	0.61	-0.04	0.32	0.00	-0.46
actors	-0.54	0.46	-0.89	1.00	-0.21	-0.50	-0.21	-0.21	-0.39	-0.21	-0.32	-0.57
amazon	-0.30	0.03	-0.97	1.00	-0.97	-0.12	-0.97	-0.87	0.03	-0.96	-0.97	-0.44
cora	0.06	0.69	-0.06	1.00	0.06	-0.06	0.19	0.69	0.06	0.44	0.19	-0.06
dblp	-0.43	0.89	-0.96	1.00	-0.96	-0.32	-0.89	-0.57	0.18	-0.88	-0.86	-0.46
flickr	0.00	-0.71	0.75	1.00	0.61	0.07	0.61	0.07	0.14	-0.01	0.61	-0.75
football	0.38	0.38	0.10	1.00	0.88	0.38	-0.37	0.56	0.38	-0.87	-0.33	0.38
github	-0.29	-0.36	-0.11	1.00	-0.07	0.11	-0.11	-0.04	-0.43	-0.14	0.07	-0.04
lj	-0.21	-0.86	0.43	1.00	0.21	-0.18	0.50	0.25	0.32	0.35	0.50	-0.32
youtube	0.36	-0.89	0.96	1.00	0.79	0.39	0.68	0.07	0.11	0.31	0.68	0.61

Table 11. Spearman's coefficient of the fb3 (ground truth, algorithms) compared to the results of quality functions. Real-world dataset

file\quality	cc	fb3	mod	nmi	perm	sign	cond	FOMD	comp	cut_ratio	f-odf	sur
CS	-0.50	1.00	-0.14	0.82	0.14	-0.75	0.39	0.75	-0.61	0.59	0.18	-0.93
actors	0.29	1.00	-0.07	0.46	-0.79	0.43	-0.79	-0.79	0.18	-0.79	-0.64	0.36
amazon	-0.86	1.00	-0.04	0.03	-0.04	-0.89	0.00	0.25	-0.93	-0.01	0.00	-0.79
cora	-0.64	1.00	0.04	0.69	0.29	-0.75	0.50	0.89	-0.75	0.79	0.50	-0.75
dblp	-0.68	1.00	-0.79	0.89	-0.79	-0.57	-0.64	-0.32	-0.07	-0.67	-0.61	-0.71
flickr	0.18	1.00	-0.21	-0.71	0.07	0.04	0.07	0.46	0.39	0.60	0.07	0.29
football	1.00	1.00	0.68	0.38	0.25	1.00	-0.96	-0.05	1.00	-0.21	-0.93	1.00
github	-0.29	1.00	0.39	-0.36	-0.57	0.71	-0.61	-0.79	0.71	-0.57	-0.93	0.71
lj	0.29	1.00	-0.54	-0.86	-0.14	0.39	-0.46	-0.36	-0.11	-0.38	-0.46	0.32
youtube	0.04	1.00	-0.86	-0.89	-0.61	-0.07	-0.54	0.04	0.14	-0.19	-0.54	-0.32

groups than actors (7.8 compared to 3.8), resembles the difference between LFR a and c class, which was demoted by NMI.

An unexpected correlation is **(dblp, amazon)**: quality functions behave in similar ways in a co-purchase network and in a co-authorship network. As observed in Sect. 7.3, this result is due to the erratic behavior of the correlation between qualities with very low correlation values.

FB3: We notice a surprising correlation of the co-purchase network with scientific networks **(amazon, cora, CS)**.

The networks that are strongly defined by the underlying bipartite network, **(football, actor, github)**, are correlated with fb3. They have a similar structure, with a particularly high clustering coefficient inside of the communities.

We observe that the **(lj, football, github)** tuple appears as close to each other. It could be explained by the underlying bipartite model of lj (and the other two OSNs) that creates a weak correlation with the other networks that are structurally more defined by it.

7.3 Quality Functions in Contexts

We analyze the correlations between the quality functions and the comparison methods. Our aim is to find quality functions that give a consistent ranking that is highly correlated with the ground truth.

In the context of OSNs (flickr, youtube, lj), we see in Table 11 that the fb3 does not give us an answer on the best quality function to use since no satisfying correlation is observed. However, we see in Table 10 that NMI tells us that Modularity gives a consistently correlated score, while Permanence also behaves well while being more inconsistent (notably with lj).

Concerning scientific collaboration networks (cora, CS), the average FOMD consistently shows a strong correlation with the ground-truth, close to the Cut-ratio. This tendency is coherent with both comparison methods.

The networks with strong bipartite underlying structure (football, github, actor) do not show any particular outlier when compared with NMI with very weak correlations. However, fb3 outlines the performance of Signature and Surprise.

The last two networks, amazon and dblp, do not show any satisfying correlation with the selected quality functions. We suspect the quality functions that we use are not adapted to the `contexts` of these graphs.

8 Conclusion

In this paper, we introduced fb3 as a clustering comparison method for community detection algorithms. We gave evidence that quality functions are `context`-dependant. The application of a quality function comparison methodology resulted in the identification of three `contexts` and of the relevant quality functions. We also provided evidence that the methodology clearly differentiate `contexts`.

The methodology that has been presented here may very well be applied to overlapping/weighted quality functions that would measure the efficiency of overlapping/weighted community detection algorithms.

We are currently in the process of integrating all the functionalities presented in this paper in a tool that will be made available shortly to the public.

References

1. Aldecoa, R., Marn, I.: Surprise maximization reveals the community structure of complex networks. Scientific reports 3, January 2013
2. Almeida, H., Guedes, D., Meira Jr., W., Zaki, M.J.: Is there a best quality metric for graph clusters? In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011, Part I. LNCS, vol. 6911, pp. 44–59. Springer, Heidelberg (2011)
3. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Inf. Retrieval **12**(4), 461–486 (2009)

4. Bagga, A., Baldwin, B.: Entity-based cross-document coreferencing using the vector space model. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, pp. 79–85. Association for Computational Linguistics (1998)
5. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
6. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theor. Exp. **2008**(10), P10008 (2008)
7. Chakraborty, T., Sikdar, S., Tammana, V., Ganguly, N., Mukherjee, A.: Computer science fields as ground-truth communities: their impact, rise and fall. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 426–433. IEEE (2013)
8. Chakraborty, T., Sikdar, S., Ganguly, N., Mukherjee, A.: Citation interactions among computer science fields: a quantitative route to the rise and fall of scientific research. Soc. Netw. Anal. Mining **4**(1), 1–18 (2014)
9. Chakraborty, T., Srinivasan, S., Ganguly, N., Mukherjee, A., Bhowmick, S.: On the permanence of vertices in network communities. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2014, pp. 1396–1405. ACM, New York (2014). http://doi.acm.org/10.1145/2623330.2623707
10. Clauset, A., Newman, M., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**(6), 066111 (2004)
11. Creusefond, J., Largillier, T., Peyronnet, S.: Finding compact communities in large graphs. In: 2015 Proceedings of SOMERIS, Workshop of Advances in Social Networks Analysis and Mining (ASONAM). IEEE (2015)
12. van Dongen, S.: Graph clustering by flow simulation. Ph.D. thesis (2000)
13. Flake, G.W., Lawrence, S., Giles, C.L.: Efficient identification of web communities, pp. 150–160. ACM Press (2000)
14. Fortunato, S., Barthelemy, M.: Resolution limit in community detection. Proc. Natl. Acad. Sci. **104**(1), 36–41 (2007)
15. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. PNAS **99**(12), 7821–7826 (2002)
16. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**(301), 13–30 (1963)
17. Kannan, R., Vempala, S., Vetta, A.: On clusterings: good, bad and spectral. J. ACM (JACM) **51**(3), 497–515 (2004)
18. Lancichinetti, A., Fortunato, S., Kertsz, J.: Detecting the overlapping and hierarchical community structure in complex networks. New J. Phys. **11**(3), 033015 (2009)
19. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E **78**(4), 046110 (2008)
20. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC 2007), San Diego, CA (2007)
21. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**(2), 026113 (2004)
22. Raghavan, U., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E **76**(3), 036106 (2007)
23. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. **105**(4), 1118–1123 (2008)

24. Seidman, S.B.: Network structure and minimum degree. Soc. Netw. **5**(3), 269–287 (1983)
25. Šubelj, L., Bajec, M.: Model of complex networks based on citation dynamics. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 527–530 (2013)
26. Traag, V.A., Aldecoa, R., Delvenne, J.C.: Detecting communities using asymptotical surprise. Phys. Rev. E **92**(2), 022816 (2015)
27. Traag, V.A., Krings, G., Van Dooren, P.: Significant Scales in Community Structure. Scientific reports 3, October 2013
28. Van Laarhoven, T., Marchiori, E.: Axioms for graph clustering quality functions. J. Mach. Learn. Res. **15**(1), 193–215 (2014)
29. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-worldnetworks. Nature **393**(6684), 440–442 (1998)
30. Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. Knowl. Inf. Syst. **42**(1), 181–213 (2012)