Towards realistic artificial benchmark for community detection algorithms evaluation

Günce Keziban Orman, Vincent Labatut, Hocine Cherifi

Towards realistic artificial benchmark for community detection algorithms evaluation

Günce Keziban Orman*

Faculté des Sciences Mirande,
LE2I UMR CNRS 6306,
Université de Bourgogne,
9, avenue Alain Savary BP 47870, 21078, Dijon, France
E-mail: gunceorman@gmail.com
*Corresponding author

Vincent Labatut

Computer Science Department,
Galatasaray University,
Çırağan Cad. No. 36, Ortaköy 34357, Istanbul, Turkey
E-mail: vlabatut@gsu.edu.tr

Hocine Cherifi

Faculté des Sciences Mirande,
LE2I UMR CNRS 6306,
Université de Bourgogne,
9, avenue Alain Savary BP 47870, 21078, Dijon, France
E-mail: hocine.cherifi@u-bourgogne.fr

Abstract: Many algorithms have been proposed for revealing the community structure in complex networks. Tests under a wide range of realistic conditions must be performed in order to select the most appropriate for a particular application. Artificially generated networks are often used for this purpose. The most realistic generative method to date has been proposed by Lancichinetti, Fortunato and Radicchi (LFR). However, it does not produce networks with some typical features of real-world networks. To overcome this drawback, we investigate two alternative modifications of this algorithm. Experimental results show that in both cases, centralisation and degree correlation values of generated networks are closer to those encountered in real-world networks. The three benchmarks have been used on a wide set of prominent community detection algorithms in order to reveal the limits and the robustness of the algorithms. Results show that the detection of meaningful communities gets harder with more realistic networks, and particularly when the proportion of inter-community links increases.

Keywords: community structure; topological properties; LFR benchmark; configuration model; preferential attachment.

Reference to this paper should be made as follows: Orman, G.K., Labatut, V. and Cherifi, H. (xxxx) 'Towards realistic artificial benchmark for community detection algorithms evaluation', *Int. J. Web Based Communities*, Vol. X, No. Y, pp.000–000.

Biographical notes: Günce Keziban Orman obtained her BSc and MS in Computer Engineering at Galatasaray University in Istanbul, Turkey, in 2005, 2010 respectively. She is currently a PhD student at the University of Burgundy, France and Research Assistant at Galatasaray University. Her main research interest deals with community structure in complex networks.

Vincent Labatut received his PhD in Computer Science and Artificial Intelligence from the University Paul-Sabatier in Toulouse, France, in 2003. He is presently an Assistant Professor in the Computer Science Department of Galatasaray University in Istanbul, Turkey. His research interest include complex data mining, especially network mining, complex networks analysis of real-world systems such as social or communication networks.

Hocine Cherifi received his MSc and PhD in Computer Science, both from the National Polytechnic Institute in Grenoble, France, in 1981 and 1984, respectively. He is presently a Professor in the Department of Computer Science, University of Burgundy in Dijon, France. His research activities are in the fields of computer vision, pattern recognition and complex networks.

This paper is a revised and expanded version of a paper entitled 'Qualitative comparison of community detection algorithms' presented at DICTAP 2011, Dijon, France, June 2011.

1 Introduction

Complex systems composed of a set of interacting entities can be described by graphs where entities are the nodes and mutual interactions are the links between nodes. The analysis of the graph representation of many real-world systems from different fields has revealed some common features. Among these characteristics, the community structure of these complex networks is an important issue. Indeed, it reveals the internal organisation of the network, and allows inferring special relationship between the nodes. Because of the spread of complex network applications, the community detection problem has been studied in many different areas such as computer science, biology, sociology, resulting in numerous algorithms based on a whole range of principles (Fortunato, 2010). In order to perform community detection on a specific real-world network, one needs to select the most appropriate. This choice is difficult because of the profusion of methods, and also of the variability of their performances according to the networks characteristics.

As most of these algorithms represent the community structure under the form of a node partition, their performance can be assessed by comparing the estimated partition with the real one. This requires the availability of networks whose community structure is known. Such real-world networks are very heterogeneous and not so numerous. It is thus difficult to select a network collection matching the topological properties of the targeted system.

Artificial networks seem to be an appropriate alternative. They are widely used to compare community detection algorithms performances (Danon et al., 2005; Orman et al., 2011a). Indeed, generative models allow producing easily and quickly large collections of such networks. Moreover, these models provide a control on some topological properties of the generated networks, making it possible to mimic the targeted

system features. The only point of concern is the level of realism of the generated networks, which is a prerequisite to obtain relevant test results. For this purpose, generative models are generally defined in order to reproduce known real-world networks properties. Of course, current knowledge regarding these properties may not be exhaustive, and we can consequently never be completely assured that the generated networks are perfectly realistic. For this reason, tests on artificial networks should be seen as complementary to tests on real-world networks.

Up to now, a few methods have been designed to generate networks with a community structure. The most popular one is certainly the model by Girvan and Newman (2002). Although widely used to test and compare algorithms (Donetti and Munoz, 2004; Duch and Arenas, 2005; Girvan and Newman, 2002; Radicchi et al., 2004) it is limited in terms of realism (Lancichinetti et al., 2008). The generated networks are rather small compared to most real-world networks. Furthermore, all nodes have roughly the same degree and all communities have the same size. Yet, typically both the community size distribution and the degree distribution of real-world complex networks follow a power law (Da Fontura Costa et al., 2008; Guimerà et al., 2003). To tackle this problem, several variants of this model have been defined, producing larger networks, and communities with heterogeneous sizes (Danon et al., 2006; Fortunato, 2010; Pons and Latapy, 2005).

More recently, a different approach appeared, based on rewiring. First, an initial network with desired properties is randomly generated, then virtual communities are drawn, and finally some links are rewired so that these communities emerge in the network. The method introduced by Bagrow (2008) uses the Barabasi-Albert (BA) model (Barabasi and Albert, 1999) to generate the initial network. It produces small networks with a power law degree distribution. As all the communities have the same size, this algorithm does not capture an essential property of real-world networks. The method by Lancichinetti et al. (LFR) (2008) is based on the configuration model (CM) (Molloy and Reed, 1995), which generates networks with power law degree distribution too. However, unlike Bagrow's method, the network size is not constrained and the community size distribution is a power law. Although LFR exhibits the most realistic properties, it also has some noticeable limitations. Previous experiments show that generated networks exhibit a low transitivity and close to zero degree correlation (Orman and Labatut, 2009), while real-world networks usually have a clearly non-zero degree correlation, and their transitivity is relatively high (Newman, 2003).

Previous study demonstrates that community detection algorithms can be very sensitive to topological properties variation of the benchmarks in terms of accuracy. Indeed, performance degradation was observed when authors switched from equal-sized communities to heterogeneous distributions (Danon et al., 2006; Pons and Latapy, 2005). The introduction of a power law degree distribution also made the benchmarks more discriminatory, allowing to highlight differences between algorithms whose performances were considered similar before (Lancichinetti et al., 2008).

In this work, we propose and evaluate two modifications of the LFR method to improve the realism of the generated networks. The realism level is appreciated by comparing popular topological properties of synthetic networks with reference values commonly observed in real ones. In order to assess the influence of the realism level variation, 11 widespread community detection algorithms are tested with artificial networks generated by the original and modified LFR methods. Preliminary results have been presented in Orman and Labatut (2010). In this paper, we investigate more deeply

the effect of network realism by conducting an exhaustive comparison of a wider range of community detection algorithms.

The rest of this article is organised as follows. Section 2 describes the topological properties used to characterise complex networks. Section 3 is dedicated to the LFR method presentation and the proposed modifications. Section 4 is a brief description of the community detection algorithms under test. We present and justify the choice of selected parameters values to generate the benchmarks in Section 5. Topological properties of the datasets are compared with reference to the typical values of real-world networks in Section 6. The effect of the realism level of the benchmarks on community detection accuracy is investigated in Section 7. Finally, in Section 8, we highlight our contributions.

2 Topological properties

Undirected real-world networks are known to share some common properties. In this section, we present the most prominent ones: small-worldness, transitivity, degree-related properties and centrality-related properties.

- *Small-worldness*. The small word phenomenon is typical of many real-world networks where shortcuts connecting different areas of the networks allow reducing the distance between any two nodes of the networks. In social networks context, it has been popularised by the famous theory of 'six degree of separation' between people. Specifically, small world networks have a low average distance (i.e., the length of the shortest path between pair of nodes). Furthermore, it grows logarithmically with the number of nodes (Newman, 2003). This property is important, because it is related to the network efficiency to propagate information.
- *Transitivity*. The transitivity or clustering reflects the tendency for link formation between neighbouring nodes. The local clustering coefficient of a node is defined as the number of triangles in which the node participates, normalised by the maximum possible number of such triangles (Watts and Strogatz, 1998). In a human interaction network, it measures how well the friends of a person know each other. If none of them relates together, its value is zero while if they are all friends, its value is one. To characterise the global clustering coefficient, two different measures have been introduced. The higher the coefficient, the higher the probability to observe a link between two nodes sharing the same neighbour. Transitivity is known to be higher in real-world networks as compared to a purely random one with the same number of nodes and links. It usually takes high values for social networks.
- *Degree distribution*. Most real-world networks have a highly inhomogeneous degree distribution with few nodes linked to many other nodes and a large number of poorly connected nodes. Nodes with high degree are called hubs, because they have a more central role in the network. This inegalitarian structure is well described by a power law distribution. In other words, the probability for a node to have a degree k is $p_k \sim k^{-\gamma}$. Such networks are called scale-free, because their degree distribution does not depend on their size. Experimental studies showed that the γ coefficient usually ranges from 2 to 3 (Barabasi and Albert, 2002; Boccaletti et al., 2006; Newman, 2003).

- *Average and maximum degree*. In a real-world network, the average and maximal degrees generally depend on the number of nodes it contains. For a scale-free network, it is estimated to be $\langle k \rangle \sim k_{\max}^{-\gamma+2}$ (Barabasi and Albert, 2002; Boccaletti et al., 2006) and $k_{\max} \sim n^{1/(\gamma-1)}$ (Newman, 2003), respectively.
- *Degree correlation*. It indicates how a node is related to its neighbours according to its degree. Indeed, hubs might associate preferentially either with other hubs or with low-degree. In assortative networks, the nodes tend to associate with their connectivity peers and the degree correlation is positive. In disassortative networks, hubs tend to associate with low-degree nodes and the degree correlation is negative. Real-world networks usually exhibit a non-zero degree correlation value. Social networks tend to be assortative, while other kinds of networks are generally disassortative (Newman, 2003).
- *Centrality*. Centrality determines how influential a node is within a network. Among the various ways to define such a characteristic, degree, closeness and betweenness centrality are the most widely used (Freeman, 1979). *Degree centrality* measures the involvement of a node in the network by the number of nodes connected to it. This local definition does not take into account the position of the node in the network and therefore cannot measure its ability to reach others quickly. *Closeness centrality* based on the inverse sum of shortest distances to all the other nodes of the network capture this feature. *Betweenness centrality* asserts the ability of a node to play a 'broker' role in the network by measuring how well it lies on the shortest paths connecting other nodes.

  While centrality is a measure of the leadership of a node, centralisation is a global feature of the network. It measures the degree to which the network is focused around a few central nodes. A very centralised network is dominated by one or a few very central nodes. It is therefore very sensitive to these central node failures or attacks while a less centralised is more resilient. Centralisation measures are based on the differences between the centrality scores of the most central point and those of all other points. Its definition is general, so it can be based on any centrality measure. For the three centrality concepts presented, its value ranges from 0 to 1. A value of 0 is obtained on all three measures for a 'complete' graph while a value of 1 is achieved for a 'star' or 'wheel' graph.

3 Network generation

The LFR method produces networks with non-overlapping communities using a two steps process. First, a scale-free network is created by using a random model. Second, a community structure is randomly drawn, and the network is rewired to control the proportion of inter-community links. In this study, we propose to use another random model compared to the original one used in the first step of this algorithm. Two alternative models are considered.

3.1 Original LFR method

In the first step, the CM (Molloy and Reed, 1995) is used in order to generate a network containing n nodes, with average degree $\langle k \rangle$, maximum degree k_{max} and a power law distributed degree with exponent γ. Then, the second step is applied in two phases. First, the communities are randomly drawn, so that their distribution size follows a power law with exponent β. These are just virtual communities, i.e., groups of nodes, and the topology of the network does not reflect them for now. Second, an iterative process takes place to rewire certain links without changing the node degrees. The rewiring aims to control the proportion of inter-community links of each node according to a user defined parameter called the mixing coefficient μ. It is generally not possible to meet this constraint exactly, and the mixing coefficient is therefore only approximated in practice. Its value determines how clearly the communities are defined. For small μ values, the communities are distinctly separated because they share only a few links, whereas when μ increases, the proportion of inter-community links becomes higher, making community identification a difficult task. The network has no community structure for a limit value of the mixing coefficient given by: $\mu \min > (n - n_c^{\max}) / n$, where n and $n_c^{\max}$ are the number of nodes in the network and in the biggest community, respectively (Lancichinetti and Fortunato, 2009b). The LFR method guaranties to obtain several realistic properties: size of the network, power law distributed degrees and community sizes. Moreover, some parameters give the user a direct control on these properties: network size (n), degree distribution (γ, k_{max}, $\langle k \rangle$), community size distribution structure (β) and community visibility (μ).

3.2 Modified LFR method

The CM is very flexible as it is able to produce networks with any size and degree distribution. Nevertheless, it is known to generate networks with zero correlation (Serrano and Boguñá, 2005) and low transitivity when degrees are power law distributed (Newman, 2003). To overcome these drawbacks, we propose to use more realistic models. We considered the BA preferential attachment model (Barabasi and Albert, 2002) and one of its variants called evolutionary preferential attachment (Poncela et al., 2008). Both models generate scale-free networks with desirable size and average degree. Furthermore, as we still use the second step of the original LFR algorithm, community size is also power law distributed with exponent β. The BA preferential attachment model (BA) (Barabasi and Albert, 1999) was designed as an attempt to explain the power law degree distribution observed in real-world networks by the building process of these networks. Starting from an initial network containing m_0 connected nodes, a realistic iterative process is applied to simulate growth. At each iteration, one node is added to the network, and is randomly connected to m existing nodes ($m \leq m_0$). These m nodes are selected with a probability which is a function of their current degree In other words, nodes accumulate new edges in proportion to the number they have already, leading to a multiplicative process which is known to give power-law distributions. The 'rich get richer' mechanism of preferential attachment goes by many other names such as the 'Matthew effect' in sociology, the 'cumulative advantage' in scientometrics. The exponent γ of the power law cannot be controlled though, and tends towards 3 (Barabasi and Albert, 1999). The average degree depends directly on the parameter $m(\langle k \rangle = 2m)$

(Newman, 2003). The average distance is always less than in same-sized Erdős-Rényi networks, so it has the small world property (Barabasi and Albert, 2002). Transitivity is greater than in Erdős-Rényi networks, but nevertheless decreases with network size following a power law $\sim n^{-0.75}$ (Barabasi and Albert, 2002).

The evolutionary preferential attachment (EV) (Poncela et al., 2008) model is a variant of the BA model. It also uses the preferential attachment and growth mechanisms. The attachment probabilities are not based on the current degree value, but on some nodal dynamic property, updated using the prisoner's dilemma game. In every iteration, each node plays either cooperation or defection against all its neighbours. It gets a total score depending on the individual results: 0 for unilateral cooperation or bilateral defection, 1 for bilateral cooperation, and b for unilateral defection, with $b > 1$. The first move is randomly chosen, whereas the next one depends on the respective results of the considered node and a randomly picked neighbour. If the neighbour's score is better, the node might switch its strategy, with a probability depending on the difference between their scores. Nodes with higher scores are more attractive to a node added to the network, because by being connected to them, it may use a strategy which proved to be successful. According to its authors, this process is more realistic and leads to networks with high transitivity and degree correlation values. Besides the parameters already needed by BA (n, m_0, and m), EV uses two more parameters: the points scored for unilateral cooperation (b) and the selection pressure (ε). The latter allows modulating the influence of the preferential attachment mechanism. All nodes are equiprobable when $\varepsilon = 0$, whereas the nodes scores are fully considered for $\varepsilon = 1$.

As the generating processes differ only in the first step of the LFR algorithm for simplicity matters, we will thereafter refer to the network generators by using the name of the model employed during the first step. Consequently, LFR-CM will correspond to the original LFR method, whereas LFR-BA and LFR-EV are modified versions based on the corresponding models.

4 Community detection algorithms

Over the years, many methods have been devised to provide efficient community discovery algorithms. As the spectrum is wide, building a taxonomy of solutions is not easy. In this section, we present the most influential categories summarising the various solutions existing in the literature and the representative set of algorithms that we selected for evaluation.

4.1 Link-centrality-based algorithms

The algorithms based on link-centrality measures rely on a hierarchical divisive approach. Initially, the whole network is seen as a single community, i.e., all nodes are in the same community. The most central links are then repeatedly removed. The underlying assumption is that these particular links are located between the communities. After a few steps, the network is split into several components which can be considered as communities in the initial network. Iterating the process, one can split each discovered community again, resulting in a finer community structure. Algorithms of this category differ in the way they select the links to be removed. The first and most known algorithm using this approach was proposed by Newman (Girvan and Newman, 2002), and relies on

the *edge-betweenness* measure. It estimates the centrality of a link by considering the proportion of shortest paths going through it in the whole network. As the complexity of this algorithm is high, it is not well suited for very large networks. Radicchi et al. (2004) proposed a variation called *Radetal*, based on a local measure instead of the global *edge-betweenness*. This measure called *link transitivity* is defined as the number of triangles to which a given link belongs, divided by the number of triangles that might potentially include it. Its lower complexity makes it more appropriate for large networks. It is used as the representative of the link centrality-based approach.

4.2 Modularity optimisation algorithms

Modularity is a prominent measure of the quality of a community structure introduced by Newman and Girvan (2004). It measures internal connectivity of identified communities with reference to a randomised null model with the same degree distribution. Modularity optimisation algorithms try to find the best community structure in terms of modularity. They diverge on the optimisation process they are based on. As this approach is very influential in the community detection literature, we consider three algorithms for investigation.

FastGreedy developed by Newman (2004) relies on a greedy optimisation method applied to a hierarchical agglomerative approach. The agglomerative approach is symmetrical to the divisive one described in the previous subsection. In the initial state, each node constitutes its own community. The algorithm merges those communities step by step until only one remains, containing all nodes. The greedy principle is applied at each step, by considering the largest increase (or smallest decrease) in modularity as the merging criterion. Because of its hierarchical nature, *FastGreedy* produces a hierarchy of community structures like the divisive approaches. The best one is selected by comparing their modularity values.

Louvain is another optimisation algorithm proposed by Blondel et al. (2008). It is an improvement of *FastGreedy*, introducing a two-phase hierarchical agglomerative approach. During the first phase, the algorithm applies a greedy optimisation to identify the communities. During the second phase, it builds a new network whose nodes are the communities found during the first phase. The intra-community links are represented by self-loops, whereas the inter-community links are aggregated and represented as links between the new nodes. The process is repeated on this new network, and stops when only one community remains.

Spinglass by Reichardt and Bornholdt (2006) relies on an analogy between a very popular statistical mechanic model called *Potts spin glass*, and the community structure. It applies the simulated annealing optimisation technique on this model to optimise the modularity.

4.3 Spectral algorithms

Spectral algorithms take advantage of various matrix representations of networks. Classic spectral graph partitioning techniques focus on the eigenvectors of the Laplacian matrix. They were designed to find the partition minimising the links lying in-between node groups. The methods that we selected are variants adapted to complex networks analysis.

Leading eigenvector is proposed by Newman (2006). It applies the classic graph partitioning approach, but to the modularity matrix instead of the Laplacian. Doing so, it

performs an optimisation of the modularity instead of the objective measures used in classic graph partitioning, such as the minimal cut.

Commfind is developed by Donetti and Munoz (2005). It combines the analysis of the Laplacian matrix eigenvectors used in classic graph partitioning with a cluster analysis step. Instead of using the best eigenvector to iteratively perform bisections of the network, it takes advantage of the *m* best ones. Communities are obtained by a cluster analysis of the projected nodes in this *m*-dimensional space.

4.4 Random-walk-based algorithms

Several algorithms use random walks in various ways to partition the network into communities. They rely on the intuition that random walks tend to get trapped into densely connected parts corresponding to communities. We retain two of them in our comparisons.

Walktrap by Pons and Latapy (2005) uses a hierarchical agglomerative method like *FastGreedy* but with a different merging criterion. Unlike *FastGreedy*, which relies on the modularity measure, it uses a node-to-node distance measure to identify the closest communities. This distance is based on the concept of random-walk. If two nodes are in the same community, the probability to get to a third one located in the same community through a random walk, should not be very different for both of them. The distance is constructed by summing these differences over all nodes, with a correction for the degree.

MarkovCluster simulates a diffusion process in the network to detect communities (van Dongen, 2008). This method relies on the network *transfer matrix*, which describes the transition probabilities for a random walker evolving in this network. Two transformations (expansion and inflation) are iteratively applied on this matrix until convergence. The final matrix can be interpreted as the adjacency matrix of a network with disconnected components representing the communities.

4.5 Information-based algorithms

The main idea of those approaches is to take advantage of the community structure in order to represent the network using less information than that encoded in the full adjacency matrix. We selected two algorithms from this category.

Infomod was proposed by Rosvall and Bergstorm (2007). It is based on a simplified representation of the network focusing on the community structure trough a community matrix and a membership vector. The former is an adjacency matrix defined at the level of the communities (instead of the nodes), and the latter associates each node to a community. The authors use the mutual information measure to quantify the amount of information from the original network contained in the simplified representation. They obtain the best partition by considering the representation associated to the maximal mutual information.

Infomap is another algorithm developed by Rosvall and Bergstorm (2008). The community structure is represented through a two-level nomenclature based on Huffman coding. One is used to distinguish communities in the network and the other to distinguish nodes in a community. The problem of finding the best partition is expressed as minimising the quantity of information needed to represent some random walk in the network using this nomenclature. With a partition containing few inter-community links,

the walker will probably stay longer inside communities, therefore only the second level will be needed to describe its path, leading to a compact representation. The authors optimise their criterion using simulated annealing.

4.6 Other algorithms

A number of algorithms do not fit in the previously described approaches. We selected the *label propagation* algorithm by Raghavan et al. (2007), which uses the concept of node neighbourhood and simulates the diffusion of some information in the network to identify communities. Initially, each node is labelled with a unique value. Then an iterative process takes place, where each node takes the label which is the most spread in its neighbourhood (ties are broken randomly). This process goes on until convergence, i.e., each node has the majority label of its neighbours. Communities are then obtained by considering groups of nodes with the same label. By construction, one node has more neighbours in its community than in the others.

5 Benchmark generation

In order to investigate the effect of the generative parameters on the uncontrolled topological properties of the networks, it is necessary to consider an appropriate range of values for each parameter. Since we want realistic networks, these values must be, as much as possible, consistent with what is observed in real-world networks. For this matter, we used descriptions of real-world networks measurement from the literature (Barabasi and Albert, 2002; Boccaletti et al. 2006; Da Fontura Costa et al., 2008; Newman, 2003). As we could not find all the information needed to setup the models, we also based our choices on previous experiments in artificial networks generation (Lancichinetti et al., 2008; Orman and Labatut, 2009).

Since the LFR-CM is the only network generator that makes it possible to control the degree distribution exponent value γ, it is necessary to analyse the influence of this parameter on uncontrolled topological properties. For this purpose, we performed an extensive experimentation for a wide range of the controlled parameter values and γ values ranging from 2 to 3. Figure 1 illustrates the typical behaviour resulting from these experiments. It clearly demonstrates that the power law degree distribution exponent has a negligible effect on uncontrolled properties. The fact that this parameter value is fixed for LFR-BA and LFR-EV is therefore not a problem for this study.

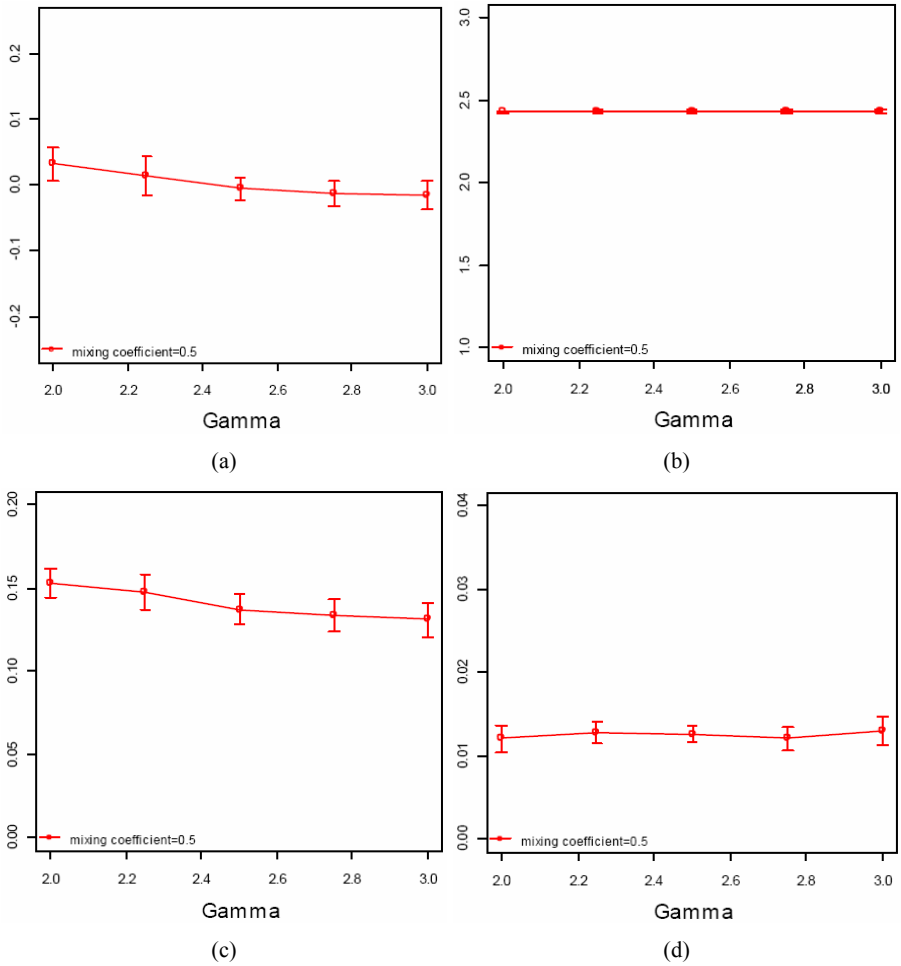
Furthermore, one need to know if and how changes in the other controlled parameters affect the uncontrolled topological properties. In a previous analysis of LFR-CM, it has been shown that variations of the exponent value β in a realistic range have a negligible effect on uncontrolled properties (Orman and Labatut, 2009). Moreover, results indicate that the mixing coefficient is the most influential parameter.

Since the uncontrolled network topological properties are not significantly sensitive to the variation of β and γ in their realistic range, we use a single value for each parameter ($\beta = 2$ and $\gamma = 3$). The later has been chosen in order to have a fair comparison of LFR-CM with competing alternatives. Recall that preferential attachment does not give any control on γ, which tends towards 3 by construction.

The average degree is directly related to the network size and in the case of scale-free networks, to the degree distribution exponent. However, this dependence is quite loose.

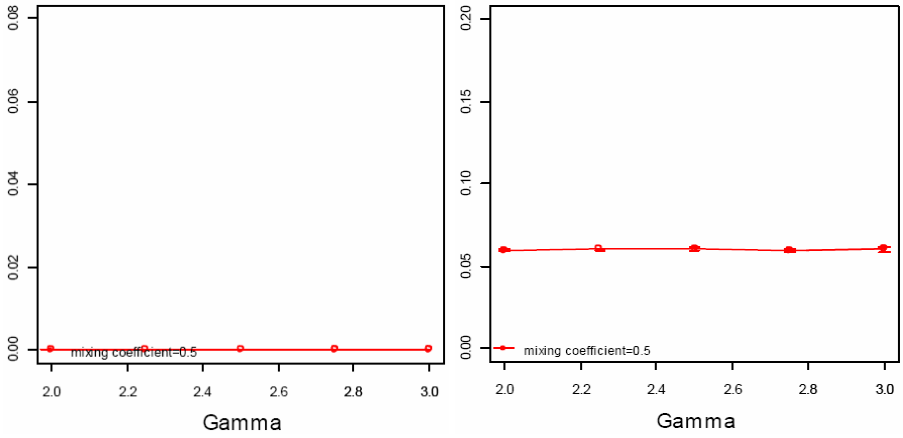
We consequently selected two consensual values for the average degree ($\langle k \rangle = 15; 30$). In LFR-CM, this constraint is enforced directly, whereas in LFR-BA and LFR-EV, we used $m = 7; 15$ to reach the same result. All three models allow controlling the average degree, but only LFR-CM lets the user specify the maximal degree. In order to get comparable networks, we tuned this parameter to make its values similar to what was observed in networks generated by LFR-BA and LFR-EV. We finally used the following sets of values for LFR-CM $\{\langle k \rangle = 15, k_{max} = 45\}$ and $\{\langle k \rangle = 30, k_{max} = 90\}$.

Figure 1 Influence of the degree distribution exponent on the measured properties, (a) degree correlation, (b) average distance (c) transitivity (d) betweenness centralisation (e) clonesness centralisation (f) degree centralisation (see online version for colours)



Note: Networks were generated with parameters $n = 5{,}000$, $\mu = 0.5$, $\langle k \rangle \approx 30$ using the original LFR method with CM.

Figure 1 Influence of the degree distribution exponent on the measured properties, (a) degree correlation, (b) average distance (c) transitivity (d) betweenness centralisation (e) clonesness centralisation (f) degree centralisation (continued) (see online version for colours)



(e) (f)

Note: Networks were generated with parameters n = 5,000, μ = 0.5, $\langle k \rangle \approx 30$ using the original LFR method with CM.

Additionally, LFR-EV allows controlling transitivity, and we found out that score and selection pressure $\varepsilon = 0.99$ give the highest transitivity.

The network size n has a direct effect on the processing time, not only regarding the generation of networks, but even more importantly concerning the community detection task. For this reason, we selected a size of n = 5,000 nodes, which is at the same time reasonably large and computationally tractable.

As the mixing coefficient μ is the most influential parameter on uncontrolled topological properties, networks are generated for different values ranging from 0.05 to 0.95 with a 0.05 step. In order to overcome the statistical discrepancies, 25 networks have been generated for each combination of parameters.

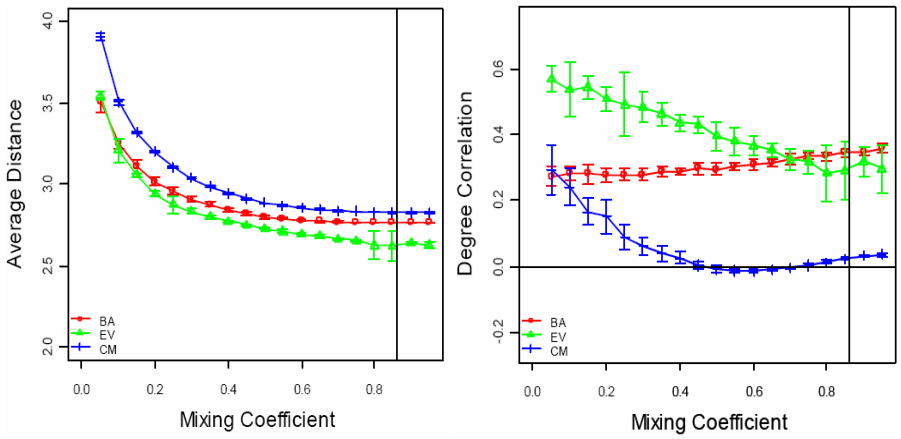
6 Generated networks properties

In this section, we present the uncontrolled topological properties of the generated networks and discuss their realism. Figure 2 shows the results for average distance, degree correlation and transitivity. Results were very similar for $\langle k \rangle$ = 15 and 30, so we only present the latter here, but comments apply to both. The largest communities in the generated networks have around 700 nodes, so communities are supposed to be structurally well-defined for μ = 0.86. Beyond this limit, represented on the plots under the form of a vertical line, properties values have little interest because the generated networks have no community structure.

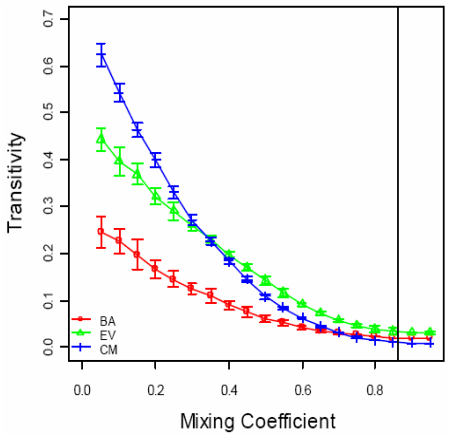
The average distance plots are rather similar for all three models. Nevertheless, the average distance is always slightly lower for LFR-BA and LFR-EV than for LFR-CM. It decreases monotonically as μ increases until an asymptotic value is reached around

μ = 0.3. Indeed, for small values of the mixing coefficient, communities are well separated and there are few links between communities. This results in longer paths between nodes belonging to different communities. When the number of inter-community links increases, the number of shortest paths between nodes also increases, so that the influence of community structure becomes negligible. Except for well separated communities, 'small-worldness' property is very robust to mixing proportion variations. This is an interesting behaviour when comparing algorithms performances.

Figure 2 Influence of the mixing coefficient μ on the measured properties, (a) average distance, (b) degree correlation and (c) transitivity (see online version for colours)



(a) (b)



(c)

Notes: Networks were generated with parameters n = 5,000, $\gamma \approx 3$, $\beta = 2$, $\langle k \rangle \approx 30$ and using the LFR method on three different generative models: CM, BA model and evolutionary preferential attachment model (EV). Each point corresponds to an average over 25 generated networks. The vertical lines at $\mu = 0.86$ represents the average limit above which communities stop being clearly defined.

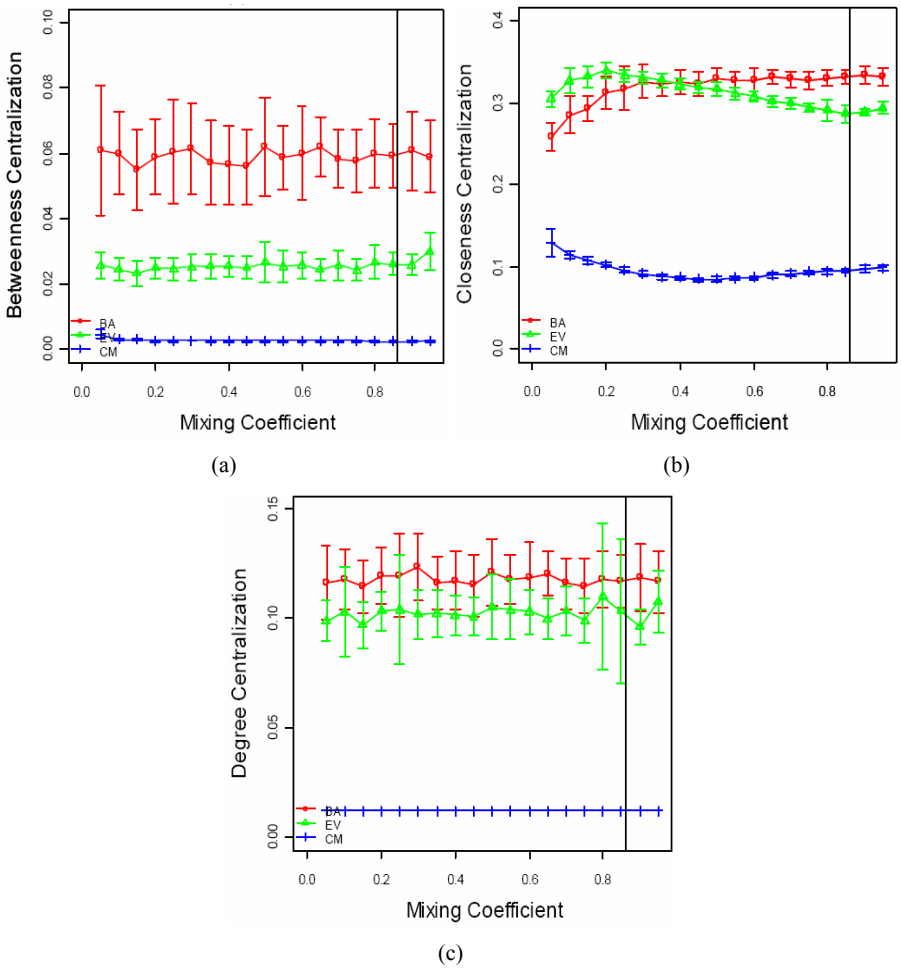
LFR-CM has the highest transitivity, with values around 0.6 for well separated communities, but it decreases rapidly reaching almost zero transitivity for μlim. We observe the same behaviour for the other methods, but the variation range is much smaller, mainly because their values for $\mu \approx 0$ are significantly smaller (around 0.25 for LFR-BA and 0.45 for LFR-EV). So surprisingly enough, LFR-EV benchmark do not exhibit a higher transitivity than LFR-CM benchmark, at least for small values of the mixing coefficient. However, due to its lesser sensitivity to the mixing coefficient variations, its transitivity is higher for $\mu > 0.3$. Note that in the literature, real-world networks with a transitivity greater than 0.3 are considered highly transitive (Da Fontura Costa et al., 2008), so we can state that all three models exhibit realistic transitivity for low μ values. The issue is more about their sensitivity to the mixing coefficient, leading to non-realistic values for high μ values. This behaviour could be linked to the rewiring process as it is common to the three models.

Considering the degree correlation, there is a clear difference between LFR-CM and the other models. LFR-CM generates networks with realistic degree correlation values for well separated communities but it decreases rapidly and oscillates around zero for $\mu > 0.4$. The LFR-EV benchmark exhibits the highest degree correlation, with values greater than 0.5 for $\mu \approx 0$. It decreases linearly when μ increases, resulting in values close to 0.25 for $\mu \approx 1$. Finally, unlike other models, the degree correlation of the LFR-BA benchmark increases linearly with μ, ranging approximately from 0.25 ($\mu \approx 0$) to 0.35 ($\mu \approx 1$). It is also noteworthy that the statistical variations for this algorithm are much lower than for the two others.

Figure 3 displays the evolution of the different centralisations. Whatever the definition used, the centralisation is always higher for LFR-BA and LFR-EV as compared to LFR-CM. We can conclude that the latter does not produce networks that include influential nodes because, except for the closeness centralisation, the measured values are very close to zero. The higher centralisation values observed for both LFR-BA and LFR-EV may be linked to the preferential attachment process used in these models. It tends to generate nodes highly connected to their neighbours. Naturally, the presence of these hubs increases the degree and closeness centralisation values. Note that centralisation values are very stable relatively to the mixing coefficient. From their evolution, we can suppose that the rewiring process affect slightly the central nodes. This makes sense, at least for degree centrality, since this property is directly based on the degree, and the rewiring process is supposed to preserve the degree distribution. Overall, in regards to centralisation, we can conclude that LFR-BA is the most suited model to mimic real-world networks. Indeed, it generates networks with degree and betweenness centralisation values with the same order of magnitude that typical social networks.

To summarise, we can state that LFR-EV and LFR-BA produce more realistic networks than LFR-CM. Indeed, their topological properties are closer to those encountered in real networks. Small world property is slightly more pronounced and both algorithms exhibit realistic values for the degree correlation as compared to LFR-CM. From the perspective of centralisation, the generated networks are significantly different, yet with a clear advantage for LFR-BA followed by LFR-EV. Networks generated with LFR-CM are nevertheless more transitive, at least when the communities are well separated. This advantage is reduced as the proportion of inter-community links increases. Indeed, the transitivity decreases monotonically when the mixing coefficient value increases for the three algorithms but more drastically for LFR-CM. This drawback seems to be linked to the rewiring process as they all exhibit the same behaviour.

Figure 3 Influence of the mixing coefficient μ on the measured properties: (a) betweenness centralisation, (b) degree centralisation, (c) closeness centralisation (see online version for colours)



Notes: Networks were generated with parameters $n = 5{,}000$, $\gamma \approx 3$, $\beta = 2$, $\langle k \rangle \approx 30$ and using the LFR method on three different generative models: CM, BA model and evolutionary preferential attachment model (EV). Each point corresponds to an average over 25 generated networks. The vertical lines at $\mu = 0.86$ represents the average limit above which communities stop being clearly defined.

If we compare these algorithms in terms of the robustness of uncontrolled topological properties with respect to variations of controlled properties, LFR-EV is the most effective algorithm. LFR-CM is clearly the most sensitive, showing the largest range of values for both transitivity and degree correlation, whereas LFR-BA is the most stable.

7 Community detection performances

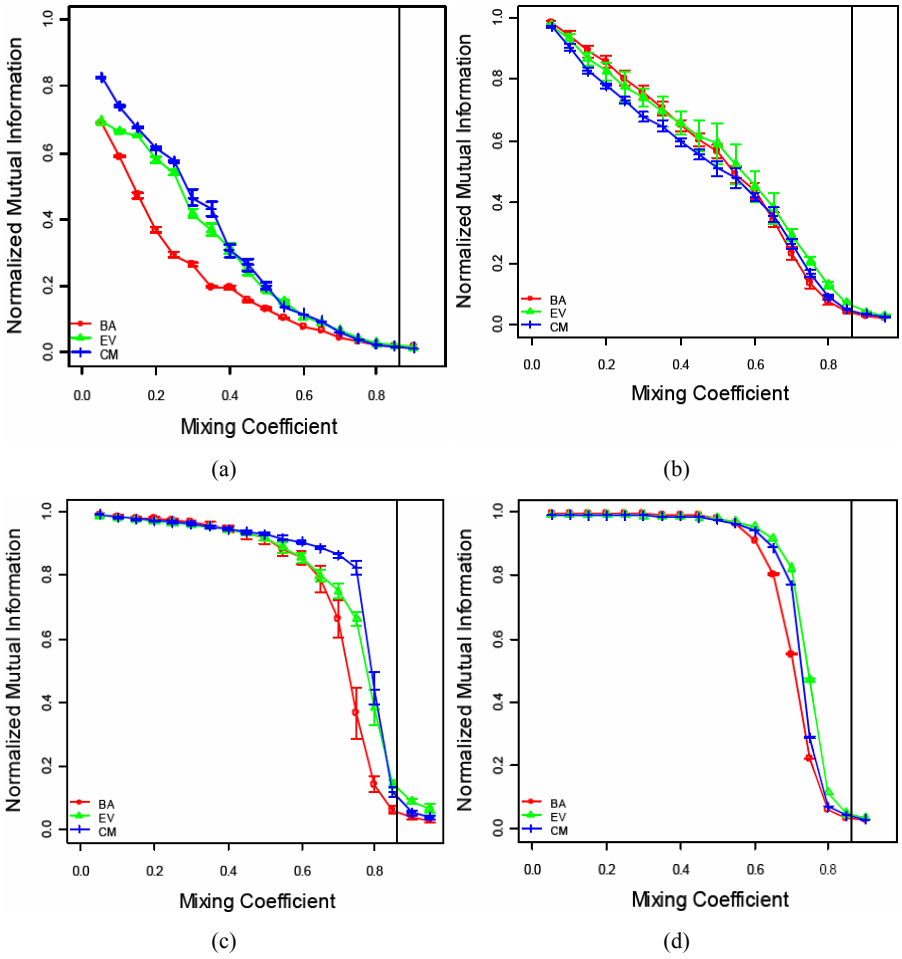
The community detection algorithms presented in Section 4 have been tested on all the generated networks. To measure the performances, we use the normalised mutual information (NMI) as it is commonly used to assess community detection performance (Danon et al., 2005; Lancichinetti et al., 2008; Lancichinetti and Fortunato, 2009a, 2009b). Although we tested the algorithms on networks with average degree $\langle k \rangle = 15$ and 30, there was no relevant difference. Indeed, the performances were uniformly slightly better for $\langle k \rangle = 30$ than for $\langle k \rangle = 15$. For the sake of clarity, we only present the results for $\langle k \rangle = 30$ in the remaining. Figure 4 presents the performances of the 11 algorithms under test on the three benchmarks. Each curve shows the evolution of the NMI with the mixing coefficient μ. Generally, as expected, the accuracy decreases with increasing community interactions. Overall, we can distinguish three types of similar behaviour when we observe the plots. In the first type of plots, for low values of the mixing coefficient, the algorithms manage to successfully identify the real communities. We do not observe any difference between the three benchmarks. In other words, in this area, the realism level of the generated networks has no influence on the performance of community detection algorithms. When the mixing coefficient increases, however, the performance deteriorates in a sharp way and the differences between the various benchmarks appear. This behaviour is characteristic of *Louvain*, *Spinglass*, *Walktrap*, *Markov Cluster* and *label propagation*. Nevertheless, the range of the mixing coefficient for which this phenomenon is observed is not the same for all the algorithms. Consequently, we can order the algorithms in terms of robustness against the variations induced by the three generative models. To do so, we compare the values of the mixing coefficient for which the accuracy differences appear. The less sensitive is *Infomap* followed by *Spinglass* and *Walktrap*. Indeed, for these algorithms, performance differences can be observed for a mixing coefficient above 0.6 approximately. *Louvain* starts to be sensitive to the model deviation when the mixing coefficient reaches 0.5. For *Markov cluster* and *label propagation*, the differences appear when the mixing coefficient is around 0.2. Note that *label propagation* is the most sensitive to the benchmarks differences. Furthermore, it is very sensitive to statistical fluctuations and its performances drop drastically compared to *Markov cluster*. Except for *label propagation*, performances are always worse with the benchmark generated with LFR-BA. Overall, there is not a clear difference between LFR-CM and LFR-EV. Indeed, very slight performance differences for *Spinglass*, *Walktrap* and *Louvain* are observed. *Infomap* and *label propagation* perform better on LFR-CM benchmark while *Markov cluster* is more efficient on LFR-EV benchmark.

In the remaining plots, whatever the mixing coefficient value, one can always observe differences on performances for the three different benchmarks. Nevertheless, they can also be distinguished in two categories by the general shape of the curves.

The second category includes *Commfind* and *Fast Greedy*. For both algorithms, the performances decrease monotonically when the mixing coefficient increases. The NMI varies almost linearly with the mixing coefficient for *Fast Greedy* while it evolves exponentially for *Commfind*. As the differences observed are not statistically significant, we can conclude that *Fast Greedy* is not very sensitive to the generative models differences. While very similar using benchmarks generated with LFR-CM and LFR-EV, *Commfind* performances deteriorate with data from LFR-BA. Note that performances are not very impressive for both algorithms for mixing coefficient values above 0.2. In other

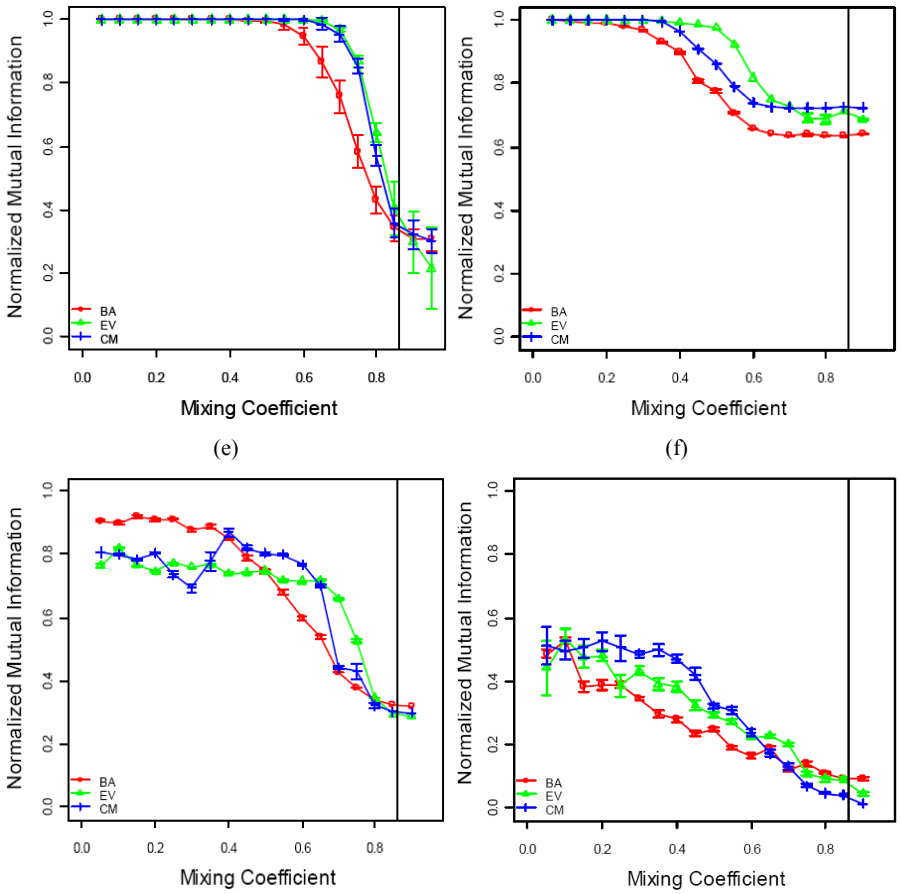
words, these algorithms are more suited for situations where communities are well separated.

Figure 4 Performances of the community detection algorithms on the three benchmarks, (a) Commfind (b) Fastgreedy (c) Louvain (d) Spinglass (e) Walktrap (e) Walktrap (g) Radicchi (h) eigenvector (i) Infomod (j) Infomap (k) label propagation (see online version for colours)



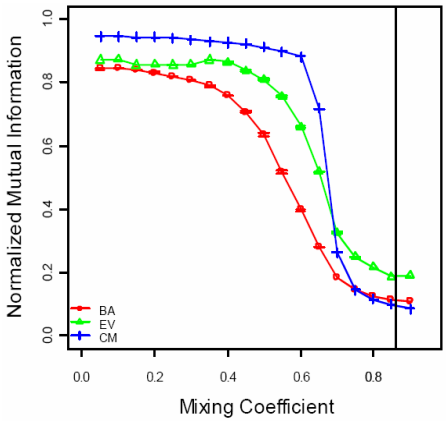
Notes: Networks were generated with parameters $n = 5{,}000$, $\gamma \approx 3$, $\beta = 2$, $\langle k \rangle \approx 30$ using three different generative models: CM, BA model and evolutionary preferential attachment model (EV). Each point corresponds to an average over 25 generated networks. The vertical lines at $\mu = 0.86$ represents the average limit above which communities stop being clearly defined.

Figure 4 Performances of the community detection algorithms on the three benchmarks, (a) Commfind (b) Fastgreedy (c) Louvain (d) Spinglass (e) Walktrap (f) Markov cluster (g) Radicchi (h) eigenvector (i) Infomod (j) Infomap (k) label propagation (continued) (see online version for colours)

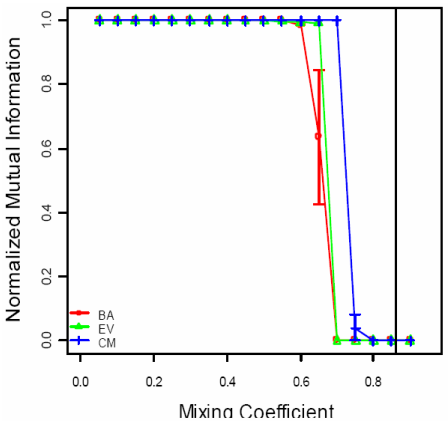


Notes: Networks were generated with parameters $n = 5{,}000$, $\gamma \approx 3$, $\beta = 2$, $\langle k \rangle \approx 30$ using three different generative models: CM, BA model and evolutionary preferential attachment model (EV). Each point corresponds to an average over 25 generated networks. The vertical lines at $\mu = 0.86$ represents the average limit above which communities stop being clearly defined.

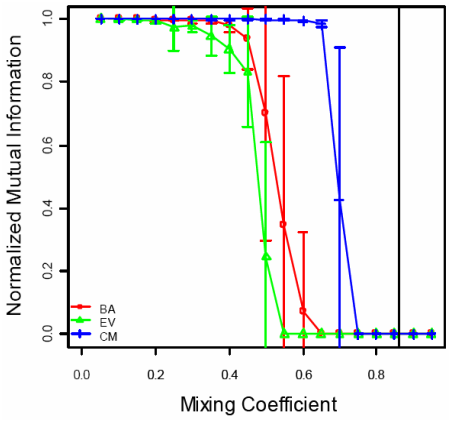
Figure 4 Performances of the community detection algorithms on the three benchmarks, (a) Commfind (b) Fastgreedy (c) Louvain (d) Spinglass (e) Walktrap (f) Markov cluster (g) Radicchi (h) eigenvector (i) Infomod (j) Infomap (k) label propagation (continued) (see online version for colours)



(i)



(j)



(k)

Notes: Networks were generated with parameters $n = 5{,}000$, $\gamma \approx 3$, $\beta = 2$, $\langle k \rangle \approx 30$ using three different generative models: CM, BA model and evolutionary preferential attachment model (EV). Each point corresponds to an average over 25 generated networks. The vertical lines at $\mu = 0.86$ represents the average limit above which communities stop being clearly defined.

In a decreasing order of efficiency, *Infomod*, *Radetal* and *leading eigenvector* are the three members of the last category. They all are sensitive to the realism level of the networks whatever the mixing coefficient value. For *Infomod* and *leading eigenvector*, it is more difficult to identify the real communities when the data are from LFR-EV instead of LFR-CM and this is even harder if the data come from LFR-BA. *Radetal*'s behaviour is original as its performances increase for LFR-BA benchmark as compared to LFR-CM and LFR-EV, when the mixing coefficient value is less than 0.5. Above this value, it

exhibits a more typical behaviour. Indeed, performances are always worse for LFR-BA generated data than for the two other alternatives.

From these results, we can conclude that the classification based on the underlying principles used by the community detection algorithms is not relevant to explain the sensitiveness to the level of realism. Indeed, algorithm from different categories can exhibit the same behaviour while others in the same category can behave quite differently. For example, *Infomap* and *Infomod* are both information theory-based algorithms but *Infomap* performances are insensitive to the benchmarks variation while this is not the case for *Infomod*. As the proportion of inter-community links is not too high, the most efficient algorithms are hardly influenced by the realism level of the networks. However, when the boundaries between communities become looser, this effect becomes significant. Globally, networks generated with the LFR-BA model are the most difficult to process, whereas those generated by LFR-CM are associated to the highest performances. The LFR-EV model lies somewhere in between. Even if it is not easy to explain the observed differences, we can say that LFR-BA is the most reliable generator because of the stability of uncontrolled topological properties.

8 Conclusions

In this paper, we investigate the effect of the realism level of artificially generated complex networks on the performance of community detection algorithms. In order to improve the realism level of the LFR method, we propose to use the BA or the evolutionary preferential attachment (EV) model instead of the original CM. An extensive evaluation of the topological properties of the three different benchmarks demonstrate that LFR-BA and LFR-EV produce networks with lower average distance, more realistic degree correlation, lower transitivity, and higher centralisation, when compared to the original LFR-CM method. For these properties, globally, LFR-EV exhibits better absolute values but LFR-BA is less sensitive to the mixing coefficient variation.

In order to analyse the effect of these modifications on the community detection process, a wide range of algorithms has been tested. Overall, we observe that performances deteriorate more or less when the degree of realism of the networks increases. The highest performances are in general obtained when applied to the LFR-CM benchmark, whereas the lowest correspond to LFR-BA data. More precisely, the algorithms can be categorised in three classes according to their sensitiveness to benchmark variation. In the first class, differences appear only when the proportion of intercommunity links is high, making the community detection problem a difficult task. Among these algorithms, *Infomap* is the most robust and efficient algorithm followed by *Spinglass* and *Walktrap*. It should be preferred by practitioners to detect communities on real networks. The algorithms from the second and the third class are always sensitive to the benchmarks variations whatever the proportion of the inter-community links. The shape of the performance curves is the main characteristic allowing to distinguish them. While in the second class performances decrease monotonically when the proportion of inter-community links increases, in the third class, performances are very stable up to a certain value of the proportion of inter-community links after which it drops sharply.

Among the three generation models, LFR-BA is the most appropriate in order to evaluate community detection algorithms. Indeed, its topological properties are closer to

those encountered in real networks compared to LFR-CM. Furthermore, the relative stability of its topological properties for the all range of mixing proportion allows to consistently compare the algorithms.

Apart from the comparison of algorithms, the proposed modifications of the LFR algorithm can have a big impact for all studies requiring the simulation of real networks, and this, especially in the field of social networks. Indeed, a realistic model to cover social network properties should generate small world networks, with positive degree correlation and high transitivity and centralisation values. LFR-BA and LFR-EV exhibits these characteristics, especially when the communities are well-separated. LFR-CM is one step behind with its low centralisation and degree correlation values. Of course, both LFR-BA and LFR-EV should be further improved to be more realistic. A more effective rewiring process that respects not only the degree of the nodes, but also a predefined local transitivity could enhance even more their realism level.

References

Bagrow, J.P. (2008) 'Evaluating local community methods in networks', *Journal of Statistical Mechanics-Theory and Experiment*, No. 5, p.05001.

Barabasi, A. and Albert, R. (1999) 'Emergence of scaling in random networks', *Science*, Vol. 286, No. 5439, p.509.

Barabasi, A.L. and Albert, R. (2002) 'Statistical mechanics of complex networks', *Reviews of Modern Physics*, Vol. 74, No. 1, pp.47–96.

Blondel, V.D., Guillaume, J-L., Lambiotte, R. and Lefebvre, E. (2008) 'Fast unfolding of communities in large networks', *J. Stat. Mech.*, No. 10, p.10008.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D. (2006) 'Complex networks: structure and dynamics', *Physics Reports*, Vol. 424, Nos. 4–5, pp.175–308.

Da Fontura Costa, L., Oliveira, O.N., Jr., Travieso, G., Rodrigues, F.A., Villas Boas, P.R., Antiqueira, L., Viana, M.P. and Da Rocha, L.E.C. (2008) 'Analyzing and modeling real-world phenomena with complex networks: a survey of applications', arXiv 0711.3199.

Danon, L., Diaz-Guilera, A. and Arenas, A. (2006) 'The effect of size heterogeneity on community identification in complex networks', *J. Stat. Mech*., No. 11, p.11010.

Danon, L., Díaz-Guilera, A., Duch, J. and Arenas, A. (2005) 'Comparing community structure identification', *J. Stat. Mech*., No. 9, p09008.

Donetti, L. and Munoz, M.A. (2004) 'Detecting network communities: a new systematic and efficient algorithm', *J. Stat. Mech*., No. 10, p.10012.

Donetti, L. and Munoz, M.A. (2005) 'Improved spectral algorithm for the detection of network communities', arXiv physics/0504059v1.

Duch, J. and Arenas, A. (2005) 'Community detection in complex networks using extremal optimization', *Phys Rev E*., Vol. 72, No. 2, p.027104.

Fortunato, S. (2010) 'Community detection in graphs', *Physics Reports*, Vol. 486, Nos. 3–5, pp.75–174.

Freeman, L.C. (1979) 'Centrality in social networks I: conceptual clarification', *Social Networks*, Vol. 1, No. 3, pp.215–239.

Girvan, M. and Newman, M.E.J. (2002) 'Community structure in social and biological networks', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 99, No. 12, pp.7821–7826.

Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F. and Arenas, A. (2003) 'Self-similar community structure in a network of human interactions', *Phys. Rev. E*, Vol. 68, No. 6, p.065103.

Lancichinetti, A. and Fortunato, S. (2009a) 'Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities', *Phys. Rev. E*, Vol. 80, No. 1, p.016118.

Lancichinetti, A. and Fortunato, S. (2009b) 'Community detection algorithms: a comparative analysis', *Phys. Rev. E*, Vol. 80, No. 5, p.056117.

Lancichinetti, A., Fortunato, S. and Radicchi, F. (2008) 'Benchmark graphs for testing community detection algorithms', *Phys. Rev. E*, Vol. 78, No. 4, Pt. 2, p.046110.

Molloy, M. and Reed, B. (1995) 'A critical point for random graphs with a given degree sequence', *Random Structures and Algorithms*, Vol. 6, Nos. 2/3, pp.161–179.

Newman, M.E.J. (2003) 'The structure and function of complex networks', *SIAM Review*, Vol. 45, No. 2, pp.167–256.

Newman, M.E.J. (2004) 'Fast algorithm for detecting community structure in networks', *Physical Review E*, Vol. 69, No. 6, p.066133.

Newman, M.E.J. (2006) 'Finding community structure in networks using the eigenvectors of matrices', *Physical Review E*, Vol. 74, No. 3, p.036104.

Newman, M.E.J. and Girvan, M. (2004) 'Finding and evaluating community structure in networks', *Phys. Rev. E*., Vol. 69, No. 2, p.026113.

Orman, G.K. and Labatut, V. (2009) 'A comparison of community detection algorithms on artificial networks', *Lecture Notes in Artificial Intelligence*, Vol. 5808, pp.242–256.

Orman, G.K. and Labatut, V. (2010) 'The effect of network realism on community detection algorithms', in *ASONAM*, Odense, DK.

Orman, G.K., Labatut, V. and Cherifi, H. (2011a) 'On accuracy of community structure discovery algorithms', *Journal of Convergence Information Technology*, Vol. 6, No. 11, pp.283–292.

Poncela, J., Gomez-Gardeñes, J., Floría, L.M., Sanchez, A. and Moreno, Y. (2008) 'Complex cooperative networks from evolutionary preferential attachment', *PLoS ONE*, Vol. 3, No. 6, p.e2449.

Pons, P. and Latapy, M. (2005) 'Computing communities in large networks using random walks', arXiv physics/0512106.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. and Parisi, D. (2004) 'Defining and identifying communities in networks', *PNAS*, Vol. 101, No. 9, pp.2658–2663, USA.

Raghavan, U.N., Albert, R. and Kumara, S. (2007) 'Near linear time algorithm to detect community structures in large-scale networks', *Phys. Rev. E*, Vol. 76, No. 3, p.036106.

Reichardt, J. and Bornholdt, S. (2006) 'Statistical mechanics of community detection', *Phys. Rev. E*, Vol. 74, No. 1, p.016110.

Rosvall, M. and Bergstrom, C. (2007) 'An information-theoretic framework for resolving community structure in complex networks', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 104, No. 18, pp.7327–7331.

Rosvall, M. and Bergstrom, C. (2008) 'Maps of random walks on complex networks reveal community structure', *Proceedings of the National Academy of Sciences*, Vol. 105, No. 4, p.1118.

Serrano, M. and Boguñá, M. (2005) 'Weighted configuration model', in *AIP Conference*.

Van Dongen, S. (2008) 'Graph clustering via a discrete uncoupling process', *SIAM J. Matrix Anal. Appl*., Vol. 30, No. 1, pp.121–141.

Watts, D.j. and Strogatz, S.H. (1998) 'Collective dynamics of 'small-world' networks', *Nature*, Vol. 393, No. 6684, pp.409–410.