Durán *et al. Appl Netw Sci* (2021) 6:29
https://doi.org/10.1007/s41109-021-00370-x

Applied Network Science

RESEARCH Open Access

Geometrical inspired pre-weighting enhances Markov clustering community detection in complex networks

Claudio Durán[1†], Alessandro Muscoloni[1†] and Carlo Vittorio Cannistraci[1,2*]

*Correspondence: kalokagathos.agon@gmail.com
†Claudio Durán and Alessandro Muscoloni contributed equally to this work.
[1] Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany
Full list of author information is available at the end of the article

Abstract

Markov clustering is an effective unsupervised pattern recognition algorithm for data clustering in high-dimensional feature space. However, its community detection performance in complex networks has been demonstrating results far from the state of the art methods such as Infomap and Louvain. The crucial issue is to convert the unweighted network topology in a 'smart-enough' pre-weighted connectivity that adequately steers the stochastic flow procedure behind Markov clustering. Here we introduce a conceptual innovation and we discuss how to leverage network latent geometry notions in order to design similarity measures for pre-weighting the adjacency matrix used in Markov clustering community detection. Our results demonstrate that the proposed strategy improves Markov clustering significantly, to the extent that it is often close to the performance of current state of the art methods for community detection. These findings emerge considering both synthetic 'realistic' networks (with known ground-truth communities) and real networks (with community metadata), and even when the real network connectivity is corrupted by noise artificially induced by missing or spurious links. Our study enhances the generalized understanding of how network geometry plays a fundamental role in the design of algorithms based on network navigability.

Keywords: Complex networks, Community detection, Markov clustering, Network similarity, Clustering

Introduction

Markov clustering (MCL) (van Dongen 2000) is an effective algorithm for data clustering in high-dimensional feature space and several studies in physics, computer science and bioinformatics (Vlasblom and Wodak 2009; Lancichinetti and Fortunato 2009; Papadopoulos et al. 2012a; Xie and Szymanski 2013; Jia et al. 2012; Satuluri and Parthasarathy 2009) tested its performance also for community detection. In network science, community detection refers to the well-known task in which complex networks are partitioned into communities. A community is an ensemble of nodes that are more likely to be interconnected rather than to out-connect to other groups of nodes (Girvan and Newman 2002). However, the previous studies (Vlasblom and Wodak 2009; Lancichinetti

and Fortunato 2009; Papadopoulos et al. 2012a; Xie and Szymanski 2013; Jia et al. 2012; Satuluri and Parthasarathy 2009) that applied MCL for community detection obtained results far from the ones provided by the state of the art methods (Yang et al. 2016; Hric et al. 2014) such as Infomap and Louvain. These preliminary experiments have been followed by further attempts to gain performance with MCL, trying to adopt deterministic partitions (Xie and Szymanski 2013) and regularization procedures (Satuluri and Parthasarathy 2009), with promising but not consistent outcomes. Recently, it was demonstrated that latent geometry inspired (LGI) measures can significantly improve affinity propagation (Cannistraci and Muscoloni 2018)—a clustering algorithm based on a message passing procedure—for community detection in undirected and unweighted networks with non-overlapping communities. Here we argue that the problem should be addressed using a similar strategy, with a pre-weighting of the input MCL adjacency matrix. In the next section, after describing the procedure behind MCL, we recall a collection of network science notions, at the interface between network topology and network geometry (Serrano et al. 2008; Krioukov et al. 2010; Papadopoulos et al. 2012b; Muscoloni et al. 2017; Muscoloni and Cannistraci 2018a; Jalili and Perc 2017), based on which we propose a rationale that can guide to design similarity measures to boost algorithms based on network navigability protocols. Here, our aim is to investigate the extent to which LGI measures can be employed to improve also MCL community detection. For this reason, we call our proposed methodology: latent geometry inspired MCL (LGI-MCL). The analyses performed in this study compare the LGI-MCL variants against the original MCL and the state of the art methods Infomap and Louvain. After presenting the results of wide evaluations both on real networks, real networks with noisy information and on a large benchmark of synthetic 'realistic' networks, we will finally discuss advantages and limitations of the LGI-MCL approach.

Methods

Markov clustering for community detection

MCL is an algorithm for data clustering based on simulations of stochastic flows (random walks) in graphs. It takes as input an unweighted or weighted network, where the weights are interpreted as similarities, and it works with an iterative process by applying in alternation two operators called expansion and inflation, which update a stochastic matrix representing the probabilities of random walks. The expansion operator corresponds to the computation of random walks of higher length (many steps), associating new probabilities between each pair of nodes. Since it is more frequent to have higher length paths within clusters rather than between different clusters, the probabilities related with node pairs from the same cluster will in general be larger. The inflation operator will then have the effect of increasing intra-cluster probabilities and lowering inter-cluster walks (van Dongen 2000). The iteration of expansion and inflation eventually leads to the separation of the graph into segments without paths between them, which is interpreted as the clustering result. The inflation operator has a parameter, which serves to detect clustering patterns on different scales of granularity. In our simulations, given the correct number of communities to detect, we implemented a binary search in the range [1.1, 20] in order to choose the inflation parameter value that produces a number of communities as close as possible to the correct one.

LGI-MCL: proposed rationale and relative similarity measures engineering

The proposed rationale states that, in order to favour the simulation of random walks, the graph similarities (or dissimilarities) should approximate the closeness (or distances) on the hidden nonlinear manifold that characterizes the graph geometry (Papadopoulos et al. 2012b; Boguñá et al. 2008). Indeed, in many networks the information can efficiently flow according to a greedy routing procedure because their topology is emerging from this hidden geometry (Boguñá et al. 2008), whose hyperbolic and tree-like structure facilitates the greedy propagation (Papadopoulos et al. 2012b; Boguñá et al. 2008; Cannistraci et al. 2010, 2013; Muscoloni and Cannistraci 2019). Recently, Muscoloni et al. (2017) and Muscoloni and Cannistraci (2018a) proposed two latent geometry based pre-weighting techniques (one local and one global) as valuable strategies for approximating the pairwise geometrical distances between connected nodes of an unweighted network. In a later study of the same authors, the clustering algorithm affinity propagation was applied to the community detection task adopting two related dissimilarity matrices, containing dissimilarity values both for connected and disconnected nodes, which proved to simulate a more navigable geometry than other kernels previously designed for this purpose (Cannistraci and Muscoloni 2018). Here, in accordance with the MCL algorithm requirements, we converted the previous pre-weighting techniques in similarity measures. They contain and merge two fundamental properties that characterize the hidden geometry of many real complex networks and thus might serve to improve stochastic flow simulations: node similarity (proximity or homophily), related with the network clustering and the concept of local attraction between common neighbours, and node popularity (centrality), related with the node degree (Papadopoulos et al. 2012b).

The first approach—which is called the repulsion-attraction rule (RA) (Muscoloni et al. 2017; Muscoloni and Cannistraci 2018a)—assigns an edge weight adopting only the *local* information related to its adjacent nodes (neighbourhood topological information). The repulsive part behind RA involves that adjacent nodes with a high external degree (where the external degree is computed considering the number of neighbours not in common) should be geometrically far. Indeed, they represent hubs without neighbours in common, which—according to the theory of navigability of complex networks presented by Boguñá et al. (2008)—tend to dominate geometrically distant regions. On the contrary, the attractive part of RA exploits that adjacent nodes sharing a high number of common neighbours should be geometrically close because most likely they have many things in common and therefore are similar. Thus, the RA (see below for the precise mathematical formula) is a simple and efficient approach that quantifies the trade-off between *hub repulsion* and *common-neighbours-based attraction* (Muscoloni et al. 2017; Muscoloni and Cannistraci 2018a). The algorithm to compute the RA similarity for each link (i, j) in the network is the following (note that the dissimilarity value is marked with an asterisk):

(a) Compute the RA pre-weighting (equivalent to the already published rule) (Muscoloni et al. 2017; Muscoloni and Cannistraci 2018a):

$$RA^*_{ij} = \frac{1 + e_i + e_j}{1 + cn_{ij}}$$

e_i is the number of external links of the node i (links that do not connect either to common neighbours with j or to j), e_j is the same for the node j; cn_{ij} is the number of common neighbours of the link (i, j).

(b) Convert into a similarity value:

$$RA_{ij} = 1 + \frac{1}{1 + RA^*_{ij}}$$

Although inspired by the same rationale, the second similarity is global (exploits the entire network topology to compute each similarity value between pairs of nodes), in fact as a first step it makes a global-information-based pre-weighting of the links, using the edge-betweenness-centrality (EBC) to approximate distances between nodes and regions of the network (Muscoloni et al. 2017). EBC is indeed a global topological network measure that assigns to each link a value of centrality related to its importance in propagating information across different regions of the network. The assumption is that central edges are bridges that tend to connect geometrically distant regions of the network, while peripheral edges tend to connect nodes in the same neighbourhood. The higher the EBC value of a network link, the more information will pass through that link. The algorithm to compute the EBC similarity for each link (i, j) in the network is the following:

(a) Compute the EBC pre-weighting (Muscoloni et al. 2017):

$$EBC^*_{ij} = \sum_{s,t} \frac{\sigma\left(s, t|e_{ij}\right)}{\sigma\left(s, t\right)}$$

s,t is any combination of network nodes; $\sigma(s,t)$ is the number of shortest paths between s and t; $\sigma(s,t\backslash e_{ij})$ is the number of shortest paths between s and t passing through the link (i, j).

(b) Convert into a similarity value:

$$EBC_{ij} = 1 + \frac{1}{1 + EBC^*_{ij}}$$

Here, we also introduce a novel similarity measure (ER) that merges the previous ones (EBC and RA) for each link (i, j) in the network as follows:

(a) Compute the pre-weightings RA^*_{ij} and EBC^*_{ij}.

(b) Convert into a unique similarity value:

$$ER_{ij} = 1 + \frac{1}{1 + RA^*_{ij}} + \frac{1}{1 + EBC^*_{ij}}$$

State of the art community detection methods: Infomap and Louvain

The community detection algorithms Infomap (Rosvall and Bergstrom 2011) and Louvain (Blondel et al. 2008) are two state of the art approaches that have been shown to provide high performances on synthetic benchmarks (Lancichinetti and Fortunato 2009; Yang et al. 2016; Orman and Labatut 2009). Recently, they have been tested also on small-size and large-size real networks, resulting overall among the best performing on recovering ground-truth communities associated to metadata (Hric et al. 2014).

The Infomap algorithm (Rosvall and Bergstrom 2011) finds the community structure by minimizing the expected description length of a random walker trajectory using the Huffman coding process. It uses the hierarchical map equation, a further development of the map equation, to detect community structures on more than one level. The hierarchical map equation indicates the theoretical limit of how concisely a network path can be specified using a given partition structure. In order to calculate the optimal partition (community) structure, this limit can be computed for different partitions and the community annotation that gives the shortest path length is chosen. We used the C implementation released by the authors at http://www.mapequation.org/code.html.

The Louvain algorithm (Blondel et al. 2008) is separated into two phases, which are repeated iteratively. At first every node in the (weighted) network represents a community in itself. In the first phase, for each node i, it considers its neighbours j and evaluates the gain in modularity that would take place by removing i from its community and placing it in the community of j. The node i is then placed in the community j for which this gain is maximum, but only if the gain is positive. If no gain is possible node i stays in its original community. This process is applied until no further improvement can be achieved. In the second phase the algorithm builds a new network whose nodes are the communities found in the first phase, whereas the weights of the links between the new nodes are given by the sum of the weight of the links between nodes in the corresponding two communities. Links between nodes of the same community lead to self-loops for this community in the new network. Once the new network has been built, the two phase process is iterated until there are no more changes and a maximum of modularity has been obtained. The number of iterations determines the height of the hierarchy of communities detected by the algorithm. We used the R function *multilevel.community*, an implementation of the method available in the *igraph* package (Csárdi and Nepusz 2006). For each hierarchical level there is a possible partition to compare to the ground-truth annotation. In this case, the hierarchical level considered is the one that guarantees the best match, therefore the detected partition that gives the highest NMI value. We let notice that most of this Methods section is equivalent to an analogous Methods section present in other studies of the authors (Cannistraci and Muscoloni 2018; Muscoloni et al. 2017).

Community detection evaluation by normalized mutual information

Different similarity measures have been developed for evaluating the matching between two partitions (the communities detected by the method and the ground-truth). They are mainly based on three categories: pair counting, cluster matching and information theory (Fortunato and Hric 2016). Although there is not yet one measure without any

drawback, the most adopted in community detection studies is the Normalized Mutual Information (NMI) (Danon et al. 2005).

The entropy can be defined as the information contained in a distribution p(x) in the following way:

$$H(X) = \sum_{x \in X} p(x) \log p(x)$$

The mutual information is the shared information between two distributions:

$$I(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p_1(x) p_2(y)} \right)$$

To normalize the value between 0 and 1 the following formula can be applied:

$$NMI = \frac{I(X, Y)}{\sqrt{H(X) H(Y)}}$$

If we consider a partition of the nodes in communities as a distribution (probability of one node falling into one community), the previous equations allow us to compute the matching between the annotations obtained by the community detection algorithm and the ground-truth communities of a network. We used the MATLAB implementation available at http://commdetect.weebly.com. As suggested in the code, when $\frac{N}{C} \leq 100$, where N represents the number of nodes and C the number of communities, the NMI should be adjusted in order to correct for chance (Vinh et al. 2010). We let notice that most of this Methods section is equivalent to an analogous Methods section present in other studies of the authors (Cannistraci and Muscoloni 2018; Muscoloni et al. 2017).

Real networks datasets

The community detection methods have been tested on 8 real networks, which represent differing systems: Karate; Opsahl_8; Opsahl_9; Opsahl_10; Opsahl_11; Polbooks; Football; Polblogs. The networks have been transformed into undirected, unweighted, without self-loops and only the largest connected component has been considered. The information of some basic statistics are available in Table 1. N is the number of nodes. E is the number of edges. The parameter m refers to half of the average node degree and it is also equal to the ratio E/N. Cl is the average clustering coefficient, computed for each node as the number of links between its neighbours over the number of possible links (Watts and Strogatz 1998). The parameter γ is the exponent of the power-law degree distribution, fitted from the observed degree sequence using the maximum likelihood procedure developed by Clauset et al. (2009) and released at http://tuvalu.santafe.edu/~aaronc/powerlaws/. C is the number of ground-truth communities.

The first network is about the Zachary's Karate Club (Zachary 1977), it represents the friendship between the members of a university karate club in US. The communities are formed by a split of the club into two parts, each following one trainer.

Table 1 Statistics of real networks

	N	*E*	*m*	*Cl*	*γ*	*C*
karate	34	78	2.3	0.59	2.1	2
opsahl 8	43	193	4.5	0.61	8.2	7
opsahl 9	44	348	7.9	0.68	5.9	7
opsahl 10	77	518	6.7	0.66	5.1	4
opsahl 11	77	1088	14.1	0.72	4.9	4
polbooks	105	441	4.2	0.49	2.6	3
football	115	613	5.3	0.40	9.1	12
polblogs	1222	16,714	13.7	0.36	2.4	2

N number of nodes, *E* number of edges, *m* half of average node degree, *Cl* clustering coefficient, *γ* power-law degree distribution exponent, *C* number of communities

The networks from the second to the fifth are intra-organisational networks from Cross and Parker (2004) and can be downloaded at https://toreopsahl.com/datasets/#Cross_Parker. Opsahl_8 and Opsahl_9 come from a consulting company and nodes represent employees. In Opsahl_8 employees were asked to indicate how often they have turned to a co-worker for work-related information in the past, where the answers range from: 0—I don't know that person; 1—Never; 2—Seldom; 3—Sometimes; 4—Often; 5—Very often. Directions were ignored. The data was turned into an unweighted network by setting a link only between employees that have at least asked for information seldom (2).

In the Opsahl_9 network, the same employees were asked to indicate how valuable the information they gained from their co-worker was. They were asked to show how strongly they agree or disagree with the following statement: "In general, this person has expertise in areas that are important in the kind of work I do." The weights in this network are also based on the following scale: 0—Do Not Know This Person; 1—Strongly Disagree; 2—Disagree; 3—Neutral; 4—Agree; 5—Strongly Agree. We set a link if there was an agreement (4) or strong agreement (5). Directions were ignored.

The Opsahl_10 and Opsahl_11 networks come from the research team of a manufacturing company and nodes represent employees. The annotated communities indicate the company locations (Paris, Frankfurt, Warsaw and Geneva). For Opsahl_10 the researchers were asked to indicate the extent to which their co-workers provide them with information they use to accomplish their work. The answers were on the following scale: 0—I do not know this person / I never met this person; 1—Very infrequently; 2—Infrequently; 3—Somewhat frequently; 4—Frequently; 5—Very frequently. We set an undirected link when there was at least a weight of 4.

For Opsahl_11 the employees were asked about their awareness of each other's knowledge ("I understand this person's knowledge and skills. This does not necessarily mean that I have these skills and am knowledgeable in these domains, but I understand what skills this person has and domains they are knowledgeable in."). The weighting was on the scale: 0—I do not know this person / I have never met this person; 1—Strongly disagree; 2—Disagree; 3—Somewhat disagree; 4—Somewhat agree; 5—Agree; 6—Strongly agree. We set a link when there was at least a 4, ignoring directions.

The Polbooks network represents frequent co-purchases of books concerning US politics on amazon.com. Ground-truth communities are given by the political orientation of

the books as either conservative, neutral or liberal. The network is unpublished but can be downloaded at http://www-personal.umich.edu/~mejn/netdata/, as well as with the Karate, Football and Polblogs networks.

The Football network (Girvan and Newman 2002) presents games between division IA colleges during regular season fall 2000. Ground-truth communities are the conferences that each team belongs to.

The Polblogs (Adamic and Glance 2005) network consists of links between blogs about the politics in the 2004 US presidential election. The ground-truth communities represent the political opinions of the blogs (right/conservative and left/liberal). We let notice that most of this Methods section is equivalent to an analogous Methods section present in other studies of the authors (Cannistraci and Muscoloni 2018; Muscoloni et al. 2017).

Synthetic networks generated by the nPSO model

The Popularity-Similarity-Optimization (PSO) model (Papadopoulos et al. 2012b) is a generative network model recently introduced in order to describe how random geometric graphs grow in the hyperbolic space. In this model the networks evolve optimizing a trade-off between node popularity, abstracted by the radial coordinate, and similarity, represented by the angular distance. The PSO model can reproduce many structural properties of real networks: clustering, small-worldness (concurrent low characteristic path length and high clustering), node degree heterogeneity with power-law degree distribution and rich-clubness. However, being the nodes uniformly distributed over the angular coordinate, the model lacks a non-trivial community structure.

The nonuniform PSO (nPSO) model (Muscoloni and Cannistraci 2018b, c) is a variation of the PSO model that exploits a nonuniform distribution of nodes over the angular coordinate in order to generate networks characterized by communities, with the possibility to tune their number, size and mixing property. We adopted a Gaussian mixture distribution of angular coordinates, with communities that emerge in correspondence of the different Gaussians, and the parameter setting suggested in the original study (Muscoloni and Cannistraci 2018b, c). Given the number of components C, they have means equidistantly arranged over the angular space, $\mu_i = \frac{2\pi}{C} \cdot (i-1)$, the same standard deviation fixed to 1/6 of the distance between two adjacent means, $\sigma_i = \frac{1}{6} \cdot \frac{2\pi}{C}$, and equal mixing proportions, $\rho_i = \frac{1}{C} (i = 1 \dots C)$. The community memberships are assigned considering for each node the component whose mean is the closest in the angular space. The other parameters of the model are the number of nodes N, half of the average node degree m, the network temperature T (inversely related to the clustering) and the exponent γ of the power-law degree distribution. Given the parameters (N, m, T, γ, C), for details on the generative procedure please refer to the original study (Muscoloni and Cannistraci 2018b, c).

Results

The first investigation of this study has been carried out on real datasets. In Table 2 we report the comparison of MCL in its original form, the three LGI-MCL variants (EBC, RA and ER) and the state of the art methods for community detection Infomap and Louvain. In addition, we made two in-silico experiments to test the robustness of the techniques in case of noise injection in the real topologies. In the first case we perturbed the

Table 2 Community detection on real networks

	Infomap	Louvain	LGI-MCL ER	LGI-MCL RA	LGI-MCL EBC	MCL
karate	0.55	0.46	**0.83**	**0.83**	0.73	0.73
opsahl 8	**0.69**	0.55	0.59	0.55	0.55	0.55
opsahl 9	**0.47**	0.41	0.39	0.40	0.40	0.43
opsahl 10	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
opsahl 11	**1.00**	0.96	0.96	0.75	0.75	0.68
polbooks	0.52	0.50	**0.57**	**0.57**	**0.57**	**0.57**
football	0.92	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**
polblogs	0.52	**0.64**	0.00	0.00	0.00	0.00
Mean NMI	**0.71**	0.68	0.66	0.63	0.62	0.61
Mean ranking	**3.06**	3.69	3.19	3.56	3.81	3.69

The table reports the Normalized Mutual Information (NMI) computed between the ground truth communities and the ones detected by every community detection algorithm for 8 real networks. NMI = 1 indicates a perfect match between the two partitions of the nodes. The methods are ranked by mean NMI over the dataset. The best result for each network as well as the best mean results are marked in bold

network structure by random deletion of 10% of the links. We repeated this procedure

Table 3 Community detection on real networks perturbed with random removal of links

	Infomap	Louvain	LGI-MCL ER	LGI-MCL RA	LGI-MCL EBC	MCL
karate	0.54	0.49	0.72	0.73	0.72	**0.74**
opsahl 8	0.55	0.51	**0.56**	**0.56**	**0.56**	**0.56**
opsahl 9	**0.49**	0.42	0.38	0.39	0.39	0.41
opsahl 10	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
opsahl 11	**0.96**	**0.96**	0.90	0.82	0.79	0.63
polbooks	0.50	0.49	**0.57**	**0.57**	**0.57**	**0.57**
football	**0.92**	0.90	**0.92**	**0.92**	**0.92**	**0.92**
polblogs	0.51	**0.63**	0.00	0.00	0.00	0.00
Mean NMI	**0.68**	**0.68**	0.63	0.62	0.62	0.60
Mean ranking	**3.25**	4.00	3.56	3.31	3.63	**3.25**

For each real network, 100 perturbed networks have been generated removing at random the 10% of links. The table reports the Normalized Mutual Information (NMI) computed between the ground-truth communities and the ones detected by every community detection algorithm for the 8 real networks, averaged over the 100 repetitions. NMI = 1 indicates a perfect match between the two partitions of the nodes. The methods are ranked by mean NMI over the dataset. The best result for each network as well as the best mean results are marked in bold

for 100 realizations, and the average results are reported in Table 3. This experiment simulates the behaviour of our algorithms in case of partial (10%) missing topological information. In the second case we perturbed the network structure by random addition of 10% of the links. We repeated this procedure for 100 realizations, and the average results are reported in Table 4. This experiment simulates the behaviour of our algorithms in case of partial (10%) addition of wrong topological information.

As a first key result, LGI-MCL outperforms the original MCL in all the three scenarios. Remarkably, LGI-MCL ER displays a higher mean NMI than the other LGI-MCL variants in the original topologies and in the random removal experiment, whereas they equally perform in the random addition framework. Furthermore, LGI-MCL ER reaches a mean NMI close to the state of the art method Louvain and a better mean ranking, highlighting the importance of merging the RA and EBC measures in a unique combined similarity. Lastly, Infomap attains overall the best result in the original topologies

Table 4 Community detection on real networks perturbed with random addition of links

	Louvain	LGI-MCL RA	LGI-MCL ER	LGI-MCL EBC	MCL	Infomap
karate	0.45	**0.76**	0.75	0.70	0.68	0.53
opsahl 8	0.51	0.53	0.54	0.54	**0.55**	**0.55**
opsahl 9	**0.42**	0.39	0.38	0.40	0.41	0.00
opsahl 10	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
opsahl 11	**0.96**	0.73	0.76	0.69	0.53	0.00
polbooks	0.49	**0.57**	**0.57**	**0.57**	**0.57**	0.50
football	0.90	**0.93**	**0.93**	0.92	0.92	0.92
polblogs	**0.41**	0.08	0.07	0.19	0.20	0.31
Mean NMI	**0.64**	0.62	0.62	0.62	0.61	0.47
Mean ranking	3.81	**3.19**	3.25	3.44	**3.19**	4.13

For each real network, 100 perturbed networks have been generated adding at random the 10% of links. The table reports the Normalized Mutual Information (NMI) computed between the ground-truth communities and the ones detected by every community detection algorithm for the 8 real networks, averaged over the 100 repetitions. NMI = 1 indicates a perfect match between the two partitions of the nodes. The methods are ranked by mean NMI over the dataset. The best result for each network as well as the best mean results are marked in bold

and in case of missing information, however it turns out to be the most unstable when spurious links are added, since in two cases (Opsahl_9, Opsahl_11) it detects the whole network as a unique community (NMI = 0).

In order to provide additional and more detailed results regarding the behavior of the clustering methods, we performed comparative tests on artificial networks produced by the nonuniform Popularity-Similarity-Optimization (nPSO) model (Muscoloni and Cannistraci 2018b, c). Indeed, the nPSO is an efficient generative model recently proposed to grow realistic complex networks, which not only are clustered, small-word, scale-free and rich-club, but also present communities whose number and size can be a priory defined. These artificial networks with known community structure offer the ground-truth to build a valid benchmark to test the performance of algorithms for community detection. The results of wide-range simulations (Figs. 1, 2, 3 and 4 and Additional file 1: Suppl. Figures 1–9)—where synthetic networks were obtained by tuning several parameter combinations of the nPSO model—highlight similarities with respect to the results on real networks. First, LGI-MCL, compared to MCL, improves significantly the community detection performance for small-size networks ($N = 100$) and high clustering ($T = 0.1$), regardless of γ changes. Instead, for middle-size networks ($N = 500$), this is mainly true when there are more communities (larger C), higher average degree (m) and $\gamma = 2$. The ranking of performance of the LGI-MCL variants, from the highest to lowest, is generally LGI-MCL ER, LGI-MCL RA and LGI-MCL EBC (Fig. 1), similarly to the real networks. Second, the performance of MCL increases and stabilizes with increasing network size (N) at $\gamma = 3$, independently from changes in temperature (T) and number of communities (C), achieving performances close to the state of the art algorithms Louvain and Infomap (Fig. 2). In this parameter setting it can be noticed that Infomap attains a slightly higher NMI than Louvain in several cases, but, on the other side, it drastically drops to NMI = 0 when the network is too dense (low N and high m), as already pointed out by the experiments of random link addition on real topologies. Third, MCL presents problems to correctly detect the communities in networks of middle ($N = 500$) and large ($N = 1000$) size at $\gamma = 2$, but improves and stabilizes the

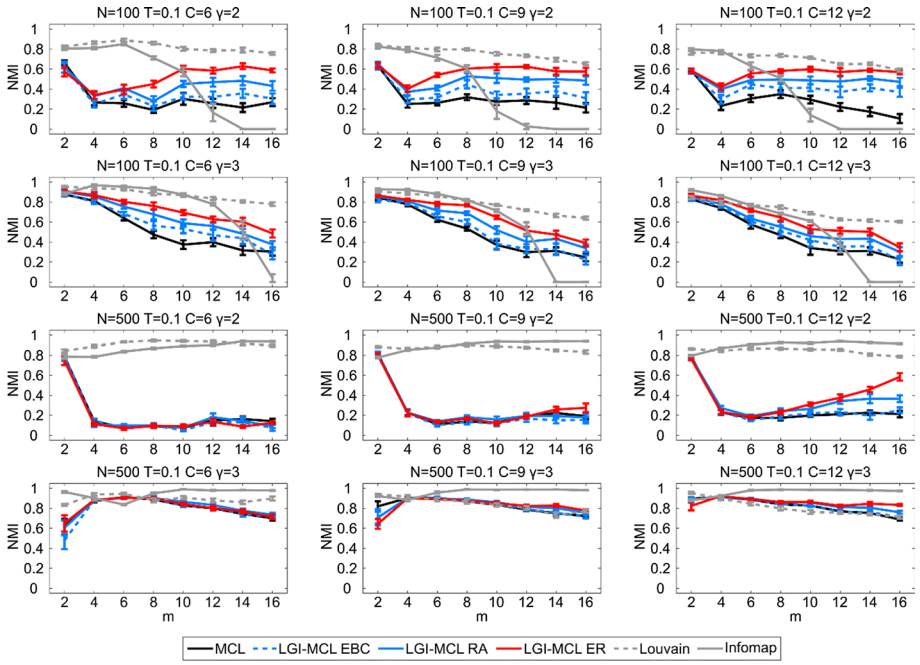

Fig. 1 Community detection on nPSO networks (1st setting: *T* fixed; *N*, *γ*, *m* and *C* changing). Synthetic networks have been generated using the nPSO model with parameters $N = [100, 500]$ (network size) $\gamma = [2, 3]$ (power-law degree distribution exponent), $m = [2, 4, 6, 8, 10, 12, 14, 16]$ (half of average degree), $T = 0.1$ (temperature, inversely related to the clustering coefficient) and $C = [6, 9, 12]$ (number of communities). For each combination of parameters, 10 networks have been generated. For each network the community detection methods have been executed and the communities detected have been compared to the annotated ones computing the Normalized Mutual Information (NMI). The plots report for each parameter combination the mean NMI and standard error over the random repetitions and show that LGI-MCL, compared to MCL, significantly improves the performance for small *N* and low *T*, regardless of *γ* changes. Instead, for middle-size networks, this is mainly true for large *C*, large *m* and low *γ*

performance for increasing γ (Fig. 3). An exception to this situation is found at very low average degree (mostly $m = 2$) (Fig. 4), where there is a peak of performance for middle ($N = 500$) and large size ($N = 1000$) networks.

Discussion

The eight considered real networks represent a benchmark with ground-truth annotation generally adopted to test algorithms for non-overlapping community detection on real network topologies. However, the results we obtain suggest that this benchmark, collecting networks of different size (from tenths to thousands of nodes), seems enough complete and diversified to adequately investigate the performance of each method here suggested. In fact, LGI-MCL should offer better results than pure MCL, because the similarity pre-weighting is derived from dissimilarity measures that approximate a network geometry. This theoretical expectation is confirmed not only on the original real networks, but also when their topology is perturbed by noise simulated by random deletion of links (missing topological information) or random addition of links (spurious topological information), where the three LGI-MCL variants achieve a greater mean NMI than the unweighted MCL, corroborating our rationale on how to design similarity measures that favour the stochastic simulation procedure of MCL. On the other

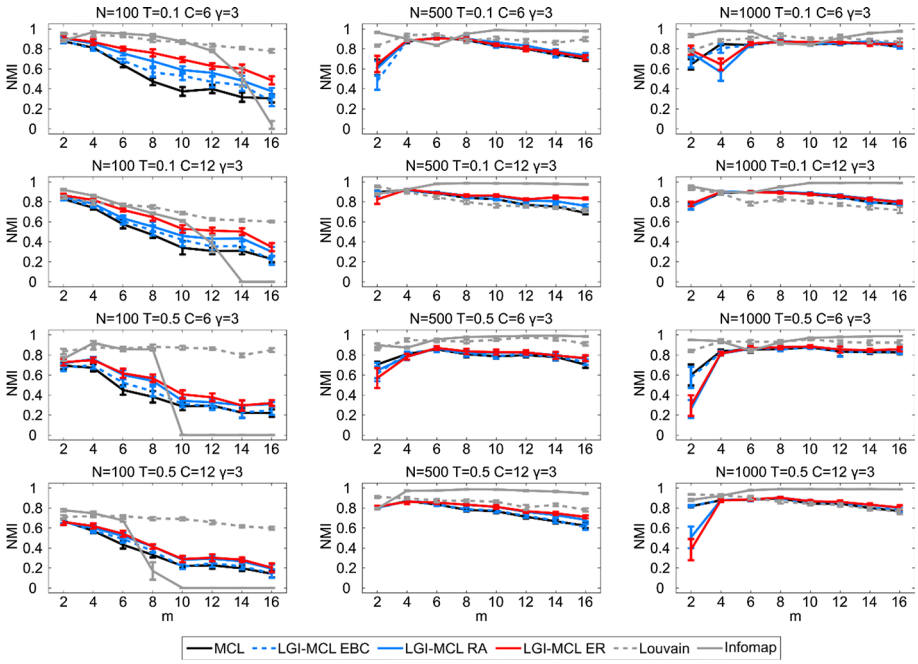

Fig. 2 Community detection on nPSO networks (2nd setting: γ fixed; N, m, T and C changing). Synthetic networks have been generated using the nPSO model with parameters $N = [100, 500, 1000]$ (network size) $\gamma = 3$ (power-law degree distribution exponent), $m = [2, 4, 6, 8, 10, 12, 14, 16]$ (half of average degree), $T = [0.1, 0.5]$ (temperature, inversely related to the clustering coefficient) and $C = [6, 12]$ (number of communities). For each combination of parameters, 10 networks have been generated. For each network the community detection methods have been executed and the communities detected have been compared to the annotated ones computing the Normalized Mutual Information (NMI). The plots report for each parameter combination the mean NMI and standard error over the random repetitions and show that the MCL performance increases and stabilizes with larger network size at $\gamma = 3$, independently from changes in T and C

hand, when considering the synthetic networks as ground-truth benchmark, LGI-MCL clearly improves the performance compared to MCL in certain scenarios, mostly for small ($N = 100$) and medium ($N = 500$) size networks, whereas for large size networks ($N = 1000$) the improvement is often missing or less notorious.

Despite the improvements that LGI measures can bring to MCL, the method is still affected by certain types of network topologies, as demonstrated for example in Fig. 3 where at low γ the MCL performance is dramatically reduced and far from the state of the art. This can be explained because with lower γ there is a stronger presence of hubs, central nodes with a large degree acting as bridges between different regions of the network, which increases the likelihood for a random walk to move from one cluster to another one, and therefore makes more difficult for MCL to correctly infer the boundaries of the clusters. Similarly, the peak of MCL performance at low average degree (Fig. 4) can be explained because the network topology is very sparse and therefore it is less likely for a random walk to reach a hub and later move to another cluster. One goal of our wide experiments was indeed to point out the topological configurations affecting

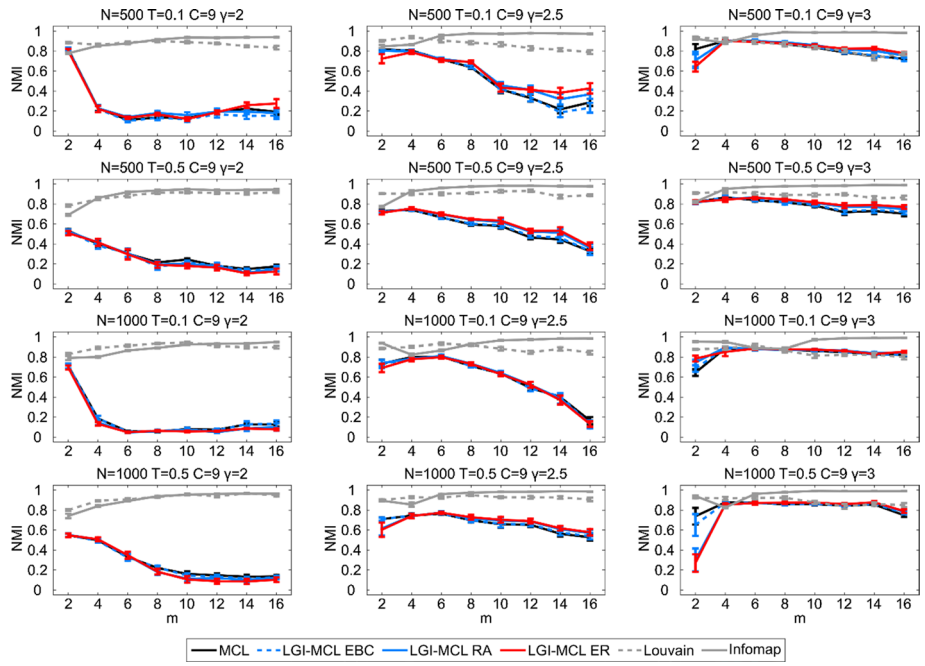

Fig. 3 Community detection on nPSO networks (3rd setting: C fixed; N, γ, m and T changing). Synthetic networks have been generated using the nPSO model with parameters $N = [500, 1000]$ (network size) $\gamma = [2, 2.5, 3]$ (power-law degree distribution exponent), $m = [2, 4, 6, 8, 10, 12, 14, 16]$ (half of average degree), $T = [0.1, 0.5]$ (temperature, inversely related to the clustering coefficient) and $C = 9$ (number of communities). For each combination of parameters, 10 networks have been generated. For each network the community detection methods have been executed and the communities detected have been compared to the annotated ones computing the Normalized Mutual Information (NMI). The plots report for each parameter combination the mean NMI and standard error over the random repetitions and show that MCL improves its performance with the increase of γ for middle and large size networks

the MCL inference, so that further studies might investigate how to improve the performance in presence of these structural patterns and make the method more robust.

In conclusion, this article introduces a rationale on how to design similarity measures for MCL, which starting from the pure topology try to approximate the hidden geometry of the manifold that generates the network topology. Since the hidden geometry of many real complex networks is hyperbolic and tree-like (Papadopoulos et al. 2012b; Muscoloni and Cannistraci 2018a), its congruous approximation can favour the stochastic simulation procedure of MCL for community detection. The empirical and numerical results provided in this study support the rationale, and the derived similarity measures EBC, RA and ER seem to boost MCL both in real and synthetic networks. Network geometry was already shown to facilitate greedy routing (Cannistraci and Muscoloni 2018; Athanassopoulos et al. 2010) and affinity propagation (Cannistraci and Muscoloni 2018), and to the best of our knowledge this is the first time that is applied to better guide random-walk (stochastic flow) based simulations. These results provide a further confirmation that network geometry can be adopted to make information flow processes more efficient, and therefore pave the

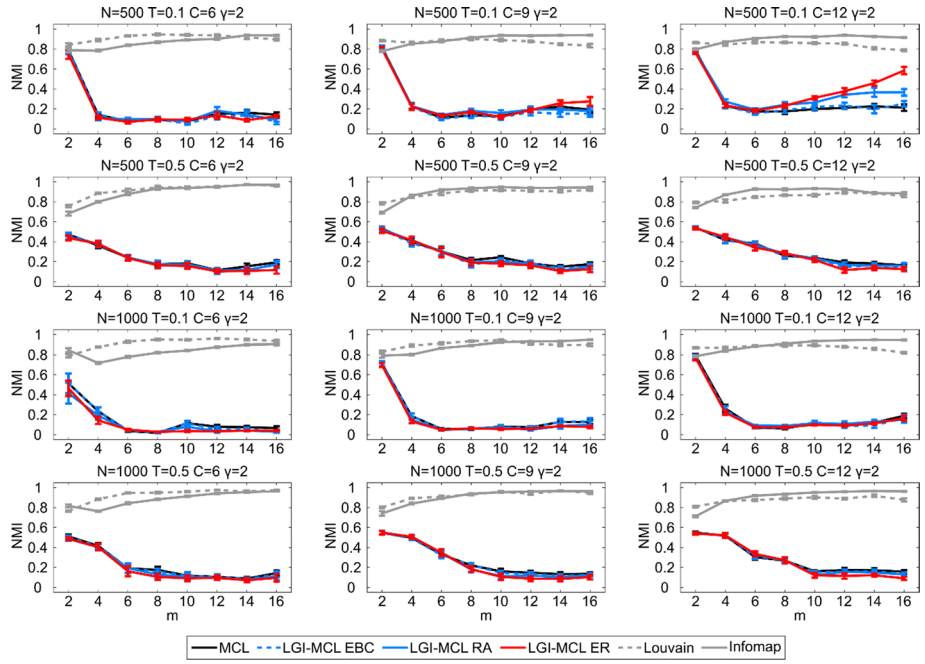

Fig. 4 Community detection on nPSO networks (4th setting: γ fixed; N, m, T and C changing). Synthetic networks have been generated using the nPSO model with parameters $N = [500, 1000]$ (network size) $\gamma = 2$ (power-law degree distribution exponent), $m = [2, 4, 6, 8, 10, 12, 14, 16]$ (half of average degree), $T = [0.1, 0.5]$ (temperature, inversely related to the clustering coefficient) and $C = [6, 9, 12]$ (number of communities). For each combination of parameters, 10 networks have been generated. For each network the community detection methods have been executed and the communities detected have been compared to the annotated ones computing the Normalized Mutual Information (NMI). The plots report for each parameter combination the mean NMI and standard error over the random repetitions and show that, at low γ, the MCL performance is close to state of the art methods for low m, whereas it decreases for higher m

way for the generalized understanding of the impact of network geometry on algorithms based on network navigability.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s41109-021-00370-x.

Additional file 1. Community detection performance on nPSO networks. Synthetic networks have been generated using the nPSO model with different network sizes, power-law degree distribution exponent, half of average degree, temperature, and number of communities.

Acknowledgements
We thank the BIOTEC System Administrators for their IT support and the Centre for Information Services and High Performance Computing (ZIH) of the TUD. We thank Gloria Marchesi for the administrative assistance.

Authors' contributions
CVC invented the latent geometry inspired graph similarities and designed the numerical experiments. CD and AM implemented the main function for community detection and performed the analysis on real networks, on perturbed real networks and on synthetic networks. All the authors analyzed and interpreted the results. CD and AM wrote the draft of the article and CVC corrected it. CVC designed the figures and tables and CD and AM realized them. CVC planned, directed and supervised the study. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. The work was supported by the independent research group leader starting grant of the Technische Universität Dresden. AM was partially supported by the funding provided by the Free State of Saxony in accordance with the Saxon Scholarship Program Regulation, awarded by the Studentenwerk Dresden based on the recommendation of the board of the Graduate Academy of TU Dresden. CD is supported by the Research Grant – Doctoral Programs in Germany (DAAD), Promotion program Nr: 57299294.

Availability of data and material
The networks analyzed during the current study are available and can be accessed from the links shared in the ***Methods*** section, under the ***Real networks datasets*** subsection. The MATLAB scripts used in this manuscript can be accessed from the Github repository: https://github.com/biomedical-cybernetics/LGI-MCL.

Declarations

Competing interests
The authors declare no competing financial interests.

Hardware and software
MATLAB code has been used for all the simulations, carried out in a workstation under Windows 8.1 Pro with 512 GB of RAM and 2 Intel(R) Xenon(R) CPU E5-2687 W v3 processors with 3.10 GHz.

Author details
[1]Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany. [2]Center for Complex Network Intelligence (CCNI) at Tsinghua Laboratory of Brain and Intelligence (THBI), Department of Biomedical Engineering, Tsinghua University, 160 Chengfu Rd., SanCaiTang Building, Haidian District, Beijing 100084, China.

Received: 13 November 2018 Accepted: 18 March 2021
Published online: 09 April 2021

References

Adamic LA, Glance N (2005) The political blogosphere and the 2004 U.S. Election: divided they blog. LinkKDD 2005, pp 36–43
Athanassopoulos S, Kaklamanis C, Laftsidis I, Papaioannou E (2010) An experimental study of greedy routing algorithms. In: Proceedings of International Conference on High Performance Computing & Simulation, pp 150–156. https://doi.org/10.1109/HPCS.2010.5547143
Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008:10008. https://doi.org/10.1088/1742-5468/2008/10/P10008
Boguñá M, Krioukov D, Claffy KC (2008) Navigability of complex networks. Nat Phys 5:74–80. https://doi.org/10.1038/nphys1130
Cannistraci CV, Muscoloni A (2018) Latent geometry inspired graph dissimilarities enhance affinity propagation community detection in complex networks. 20:063022. ArXiv: 180404566
Cannistraci CV, Ravasi T, Montevecchi FM, Ideker T, Alessio M (2010) Nonlinear dimension reduction and clustering by Minimum Curvilinearity unfold neuropathic pain and tissue embryological classes. Bioinformatics 26:i531–i539
Cannistraci CV, Alanis-Lobato G, Ravasi T (2013) Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. Bioinformatics 29:199–209. https://doi.org/10.1093/bioinformatics/btt208
Clauset A, Rohilla Shalizi C, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Rev 51:661–703. https://doi.org/10.1214/13-AOAS710
Cross R, Parker A (2004) The hidden power of social networks. Harvard Business School Press, Brighton
Csárdi G, Nepusz T (2006) The igraph software package for complex network research. Int J Complex Syst. p 1695
Danon L, Diaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. J Stat Mech Theory Exp P09008:1–10
Fortunato S, Hric D (2016) Community detection in networks: a user guide. Phys Rep 659:1–44. https://doi.org/10.1016/j.physrep.2016.09.002
Girvan M, Newman MEJ (2002) Community structure in social and biological networks. PNAS 99:7821–7826. https://doi.org/10.1073/pnas.122653799
Hric D, Darst RK, Fortunato S (2014) Community detection in networks: structural communities versus ground truth. Phys Rev E Stat Nonlinear Soft Matter Phys. https://doi.org/10.1103/PhysRevE.90.062805
Jalili M, Perc M (2017) Information cascades in complex networks. J Complex Netw 5:665–693. https://doi.org/10.1093/comnet/cnx019
Jia G, Cai Z, Musolesi M, Wang Y, Tennant DA, Weber RJM, Heath JK, He S (2012) Community detection in social and biological networks using differential evolution, pp 71–85. https://doi.org/10.1007/978-3-642-34413-8_6
Krioukov D, Papadopoulos F, Kitsak M, Vahdat A, Boguñá M (2010) Hyperbolic geometry of complex networks. Phys Rev E Stat Nonlinear Soft Matter Phys 82:036106. https://doi.org/10.1103/PhysRevE.82.036106
Lancichinetti A, Fortunato S (2009) Community detection algorithms: a comparative analysis. Phys Rev E 80:056117. https://doi.org/10.1103/PhysRevE.80.056117
Muscoloni A, Cannistraci CV (2018a) Minimum curvilinear automata with similarity attachment for network embedding and link prediction in the hyperbolic space. ArXiv: 180201183
Muscoloni A, Cannistraci CV (2018b) A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities. New J Phys 20:052002. https://doi.org/10.1088/1367-2630/aac06f
Muscoloni A, Cannistraci CV (2018c) Leveraging the nonuniform PSO network model as a benchmark for performance evaluation in community detection and link prediction. New J Phys 20:063022
Muscoloni A, Cannistraci CV (2019) Navigability evaluation of complex networks by greedy routing efficiency. Proc Natl Acad Sci 116:1468–1469. https://doi.org/10.1073/pnas.1817880116

Muscoloni A, Thomas JM, Ciucci S, Bianconi G, Cannistraci CV (2017) Machine learning meets complex networks via coalescent embedding in the hyperbolic space. Nat Commun 8:1615
Orman GK, Labatut V (2009) A comparison of community detection algorithms on artificial networks. In: Discovery science, pp 242–256
Papadopoulos S, Kompatsiaris Y, Vakali A, Spyridonos P (2012a) Community detection in social media performance and application considerations. Data Min Knowl Discov 24:515–554. https://doi.org/10.1007/s10618-011-0224-z
Papadopoulos F, Kitsak M, Serrano MA, Boguñá M, Krioukov D (2012b) Popularity versus similarity in growing networks. Nature 489:537–540. https://doi.org/10.1038/nature11459
Rosvall M, Bergstrom CT (2011) Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. PLoS ONE 6:e18209. https://doi.org/10.1371/journal.pone.0018209
Satuluri V, Parthasarathy S (2009) Scalable graph clustering using stochastic flows. In: Proceedings of 15th ACM SIGKDD international conference on knowledge discovery data mining—KDD '09, p 737. https://doi.org/10.1145/1557019.1557101
Serrano MÁ, Krioukov D, Boguñá M (2008) Self-similarity of complex networks and hidden metric spaces. Phys Rev Lett 100:1–4. https://doi.org/10.1103/PhysRevLett.100.078701
van Dongen S (2000) Graph clustering by flow simulation. Graph Stimul by flow Clust. https://doi.org/10.1016/j.cosrev.2007.05.001
Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J Mach Learn Res 11:2837–2854
Vlasblom J, Wodak SJ (2009) Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. BMC Bioinform 10:99. https://doi.org/10.1186/1471-2105-10-99
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442. https://doi.org/10.1038/30918
Xie J, Szymanski BBK (2013) Labelrank: a stabilized label propagation algorithm for community detection in networks. Netw Sci Work: NSW 2013:138–143. https://doi.org/10.1109/NSW.2013.6609210
Yang Z, Algesheimer R, Tessone CJ (2016) A comparative analysis of community detection algorithms on artificial networks. Sci Rep 6:30750. https://doi.org/10.1038/srep30750
Zachary WW (1977) An information flow model for conflict and fission in small groups. J Anthropol Res 33:452–473. https://doi.org/10.2307/3629752

Publisher's Note