Leveraging the nonuniform PSO network model as a benchmark for performance evaluation in community detection and link prediction

Muscoloni Alessandro and Cannistraci Carlo Vittorio

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, Tatzenberg 47/49, D-01307 Dresden, Germany
2 Brain Bio-Inspired Computing (BBC) Lab, IRCCS Centro Neurolesi ‘Bonino Pulejo’, Messina, Italy

E-mail: kalokagathos.agon@gmail.com

Keywords: network models, hyperbolic geometry, network benchmark, community detection, link prediction

Abstract

Advances in network geometry pointed out that structural properties observed in networks derived from real complex systems can emerge in the hyperbolic space (HS). The nonuniform popularity-similarity-optimization (nPSO) is a generative model recently introduced in order to grow random geometric graphs in the HS, reproducing networks that have realistic features such as high clustering, small-worldness, scale-freeness and rich-clubness, with the additional possibility to control the community organization. Generative models allowing to tune the structural properties of ‘realistic’ synthetic networks are fundamental, because they offer a ground truth to investigate how predictive algorithms react to controlled topological variations. Here, we discuss how to leverage the nPSO model as a synthetic benchmark to compare the performance of methods for community detection and link prediction; and we prove that the nPSO offers a reliable and realistic testing framework which can complement other existing benchmarks not based on latent geometry. Furthermore, we confirm that network embedding information can improve community detection, whereas boosting link prediction in HS still needs further investigations. Indeed, we find that the presence of communities in nPSO significantly modifies the performance of link predictors and is fundamental for the reproducibility of results observed on real networks. The nPSO can trigger valuable insights to understand the intrinsic rules of link-growth and self-organization that connect topology to geometry and that are encoded in link prediction algorithms differentiating their performance.

1. Introduction

Several generative models have been proposed in the past years as synthetic benchmarks for community detection [1]. One of the first benchmark has been developed by Girvan and Newman [2], where all the nodes have the same degree and all the communities identical size, and has been later extended by Danon et al [3] in order to generate communities with different size. A generative model able to reproduce structural properties closer to the ones observed in real networks has been proposed by Lancichinetti–Fortunato–Radicchi (LFR) [4], which is characterized by a power-law distribution of degree and community size. This model turned out to be a particular version of the degree-corrected stochastic block model (SBM) [1, 5]. The LFR benchmark has been later extended to directed and weighted networks with overlapping communities [6], meaning that communities can have also nodes in common [1]. A further benchmark for overlapping communities has been introduced by Ball et al [7], however, having all the nodes the same expected degree, it is less realistic and flexible than the LFR [1]. Interestingly, all these well-known models of networks with communities are not generated according to a latent geometry, which instead will be the fundamental theme of this study.

The investigation of hidden geometrical spaces behind complex network topologies has been a fervid topic in recent years and, currently, the hyperbolic space (HS) seems to be one of the most appropriate in order to...
2.1. Comparison of community detection algorithms on nPSO networks

2. Results and discussion

2.1. Comparison of community detection algorithms on nPSO networks

The community detection algorithms Louvain [21], Infomap [22], Walktrap [23] and Label propagation [24] are four state-of-the-art approaches that have been shown to provide high performances on synthetic benchmarks [25–27]. Louvain [21] is a model-free and unsupervised heuristic method based on modularity optimization; Infomap [22] finds the community structure by minimizing the expected description length of a random walker trajectory using the Huffman coding process; Walktrap [23] is based on an agglomerative method for hierarchical clustering, where the similarities between the nodes or groups are obtained using random walks; Label propagation [24] is an iterative algorithm in which each node label is updated with the one owned by the majority of the neighbours until reaching a consensus (see Methods for details). Louvain and Infomap have been also recently tested on small-size and large-size real networks, resulting overall among the best performing on recovering ground truth communities associated to metadata [28]. In this study, we compare these four community detection approaches across synthetic networks generated using diverse nPSO parameter combinations: \( N = [100, 500, 1000] \) (network size), \( m = [10, 12, 14] \) (half of average node degree), \( T = [0.1, 0.3, 0.5] \) (temperature, inversely related to clustering), \( \gamma = 3 \) (power-law degree distribution exponent) and angular coordinates sampled according to a Gaussian mixture distribution with equal proportions and components \( C = [4, 8] \). The values chosen for the parameter \( m \) are centred around the average \( m \) (which is equal to 12 rounded) computed on the dataset of small-size real networks. The values chosen for \( N \) and \( T \) are intended to cover the range of network size and clustering coefficient observed in the dataset of small-size real networks. Since the average \( \gamma \) estimated on the dataset of small-size real networks is higher than the typical range \( 2 < \gamma < 3 \), we choose \( \gamma = 3 \).

Figure 1 reports the mean normalized mutual information (NMI) [29] performance (NMI is a measure to assess the performance of community detection, see Methods for details) and related standard errors (10 network repetitions considered for each parameter combination) of the community detection algorithms applied to nPSO networks with 4 communities. The results indicate that overall Louvain appears as the strongest approach, with an almost perfect detection over different values of network size, average node degree and temperature. Infomap highlights problems in correctly detecting the communities when there are too many inter-community links, as can be seen for \( N = 100 \) and increasing temperature. The higher temperature in fact leads to a higher number of links between nodes that are geometrically far in the disk, which increases the mixing between the communities. The performance is more stable for bigger networks, although the performance is slightly worse than Louvain. Walktrap results as robust as Louvain to the increase of network temperature, but the NMI is slightly lower for \( N = 100 \) and \( N = 1000 \). As last, Label propagation, which is the fastest approach, but the one with lowest accuracy, performs worse than the other methods and presents the same problem as Infomap for \( N = 100 \). This issue has been already pointed out in the study of Yang et al [25], in which it is shown that for a
high mixing of the communities Louvain and Walktrap are more robust, whereas Infomap and Label propagation tend to drop in performance. Hence, the nPSO model here proposed seems to provide a good benchmark to test community detection algorithms on networks generated using a latent geometry model which is based on the HS.

Figure 2 reports the NMI performance on networks with 8 communities. Focusing firstly on the performance on bigger networks ($N = [500–1000]$), it can be noticed that Louvain and Infomap swap their behaviour, with Infomap going close to the perfect community detection and Louvain slightly decreasing its performance. Walktrap, instead, remains quite robust and slightly improves for $N = 1000$. Label propagation still remains the most unstable, although surpassing Louvain for very low temperature. In the study of Yang et al [25] it is shown that, when the mixing of the communities is not high, Louvain can slightly underestimate the number of communities for networks of increasing size, which might be the reason of its reduced performance in large networks ($N = [500–1000]$) with respect to the other approaches. Focusing now on the small-size networks ($N = 100$), suppl. figure 1 (available online at stacks.iop.org/NJP/20/063022/mmedia) highlights that the methods preserve the same ranking with respect to the case with 4 communities, but they all decrease their performance. The reason is that, being the network small and keeping the average node degree constant, the increase of the communities leads to more inter-community links, making the community structure less detectable.

2.2. Community detection on nPSO networks using network embedding information

Recently, Muscoloni et al [16] introduced coalescent embedding, a model-free topological-based machine learning class of algorithms that exploits nonlinear unsupervised dimensionality reduction to infer the node coordinates in the HS. The study also demonstrates that, exploiting the geometrical embedding information in order to weight the adjacency matrix in input to the community detection algorithms, the performance of the respective unweighted variants can be improved [16]. Since the evaluation tests in Muscoloni et al [16] were executed on real network datasets, it remains the doubt that the performance evaluation was not objective and possibly biased by the restricted metadata available on the community annotation. This is a perfect example for...
clarifying the utility of the nPSO model. We would like to have a dataset of realistic networks for which we know the ground truth community organization and that we can use as a benchmark to test whether the community detection algorithms benefit from being applied according to the geometrical embedding information in the HS. Therefore, we repeat the same tests made by Muscoloni et al.\cite{16} on real networks, but here we use nPSO synthetic networks as benchmark. More precisely, each network is embedded in the HS using the coalescent embedding techniques and the inverse hyperbolic distances (HD) between the nodes are used to weight the observed links of the input network for the Louvain, Infomap, Walktrap and Label propagation algorithms. In a second variant (working only for the Louvain method) the non-observed links are also weighted, using the inverse hyperbolic shortest paths (HSP) (see Methods for details). The results of these tests are in tables 1 and suppl. tables 1–2, which report the mean NMI for nPSO networks with 4 and 8 communities, the mean NMI over all the networks and the mean ranking, comparing the weighted variants and the unweighted ones.

Table 1 displays for the Louvain algorithm the mean NMI and the mean ranking over all the networks, which are indicators of the general performance on many parameter combinations of the nPSO. Almost all the HD-based methods increase their performance with respect to the unweighted Louvain, going from a mean NMI of 0.89 up to 0.93. A few HSP-based methods are also able to reach similar performances, whereas most of the others exhibit a decrease in NMI. Figure 3 provides a more detailed comparison between the unweighted Louvain and the top-ranked of the weighted variants (RA-LE-HD) for nPSO networks with 8 communities, showing the NMI performance in all the network configurations. The improvement is consistent over all the parameter combinations and becomes more pronounced for bigger networks, remarkably bringing the NMI close to a perfect detection.

Table 2 displays the mean NMI and the mean ranking results for the Infomap algorithm and shows that all the HD-based methods increase the mean NMI with respect to the unweighted Infomap, going from a value of 0.74 up to 0.87, but not all of them obtain a higher ranking. The reason is explained in figure 4, which provides a more detailed comparison between the unweighted Infomap and the top-ranked of the weighted variants (EBC-ncISO-HD) for nPSO networks with 8 communities. As already discussed in the previous section, Infomap highlights problems in correctly detecting the communities when there are too many inter-community links, as
can be seen for $N = 100$ and increasing temperature. In this scenario, the increase in NMI given by the HD-weighting is remarkable. For bigger networks, instead, the detection using the unweighted Infomap is almost perfect and the HD-based variants tend to obtain the same performance, except for a slight decrease at $N = 1000$ and $T = 0.1$.

Suppl. table 1 displays the results for the Walktrap algorithm and shows that all the HD-based methods obtain a higher ranking, although the mean NMI has only a small increase with respect to the unweighted Walktrap, going from a value of 0.89 up to 0.91.

At last, suppl. table 2 displays the results for the Label propagation algorithm and shows that only some of the HD-based methods increase the mean NMI with respect to the unweighted Label propagation. However, the

| RA-LE-HD         | 0.95 | 0.91 | 0.93 | 11.5 |
| RA-LE-HSP        | 0.96 | 0.91 | 0.93 | 11.7 |
| RA-ncISO-HD      | 0.96 | 0.90 | 0.93 | 11.8 |
| RA-ISO-HD        | 0.96 | 0.89 | 0.93 | 11.9 |
| RA-LE-HD         | 0.95 | 0.91 | 0.93 | 11.9 |
| RA-LE-HSP        | 0.96 | 0.91 | 0.93 | 12.0 |
| RA-ISO-HD        | 0.95 | 0.89 | 0.92 | 13.8 |
| RA-ncISO-HD      | 0.95 | 0.89 | 0.92 | 13.9 |
| RA-ncISO-HSP     | 0.95 | 0.89 | 0.92 | 15.9 |
| RA-ncISO-EA-HD   | 0.95 | 0.87 | 0.91 | 16.1 |
| RA-ncISO-EA-HD   | 0.95 | 0.86 | 0.91 | 16.6 |
| RA-ncISO-EA-HD   | 0.95 | 0.85 | 0.91 | 16.6 |
| RA-ncISO-EA-HD   | 0.95 | 0.84 | 0.92 | 16.8 |
| RA-ncISO-EA-HD   | 0.95 | 0.83 | 0.90 | 17.4 |
| RA-LE-EA-HD      | 0.94 | 0.86 | 0.90 | 17.7 |
| RA-ncMCE-EA-HD   | 0.95 | 0.85 | 0.90 | 18.0 |
| RA-LE-EA-HD      | 0.95 | 0.85 | 0.90 | 18.2 |
| RA-LE-EA-HD      | 0.93 | 0.86 | 0.90 | 18.3 |
| RA-ncMCE-HD      | 0.92 | 0.88 | 0.90 | 18.5 |
| RA-ncMCE-HD      | 0.94 | 0.86 | 0.90 | 18.7 |
| RA-ncMCE-HD      | 0.94 | 0.85 | 0.90 | 18.9 |
| RA-ncMCE-HD      | 0.92 | 0.87 | 0.90 | 19.1 |
| mean NMI         | 0.95 | 0.89 | 0.89 | 19.2 |
| mean ranking     | 11.5 | 11.7 | 11.8 | 11.9 |
| RA-ncISO-HSP     | 0.94 | 0.85 | 0.89 | 20.7 |
| RA-ncMCE-EA-HD   | 0.93 | 0.83 | 0.88 | 21.3 |
| RA-ncMCE-HD      | 0.93 | 0.83 | 0.89 | 21.5 |
| RA-ncMCE-HD      | 0.93 | 0.82 | 0.88 | 22.2 |
| RA-ncMCE-HD      | 0.90 | 0.84 | 0.87 | 24.4 |
| RA-ncISO-EA-HSP  | 0.90 | 0.81 | 0.86 | 27.4 |
| RA-ncISO-EA-HSP  | 0.90 | 0.81 | 0.86 | 27.9 |
| RA-ncISO-EA-HSP  | 0.89 | 0.81 | 0.86 | 28.0 |
| RA-ncISO-EA-HSP  | 0.89 | 0.80 | 0.85 | 28.2 |
| RA-ncISO-EA-HSP  | 0.89 | 0.80 | 0.85 | 29.2 |
| RA-ncISO-EA-HSP  | 0.88 | 0.81 | 0.85 | 29.4 |
| RA-ncISO-EA-HSP  | 0.87 | 0.80 | 0.84 | 29.4 |
| RA-ncISO-EA-HSP  | 0.87 | 0.80 | 0.84 | 29.7 |
| RA-ncISO-EA-HSP  | 0.89 | 0.79 | 0.84 | 29.8 |
| RA-ncISO-EA-HSP  | 0.87 | 0.79 | 0.83 | 30.4 |
| RA-ncISO-EA-HSP  | 0.87 | 0.79 | 0.83 | 31.0 |
| RA-ncISO-EA-HSP  | 0.88 | 0.78 | 0.83 | 31.1 |
| RA-ncISO-EA-HSP  | 0.83 | 0.76 | 0.79 | 34.6 |
Table 2. Community detection on nPSO networks using network embedding and Infomap. Synthetic networks have been generated using the nPSO model with parameters $N = [100, 500, 1000], m = [10, 12, 14], T = [0.1, 0.3, 0.5], \gamma = 3$ and angular coordinates sampled according to a Gaussian mixture distribution with equal proportions and components $C = [4, 8]$. For each combination of parameters, 10 networks have been generated. Each network has been embedded in the hyperbolic space using coalescent embedding techniques and the embedding coordinates are used to weight the input matrix for the Infomap algorithm: observed links are weighted using the hyperbolic distances (HD) between the nodes (see Methods for details). As a reference, the Infomap algorithm has been run giving in input also the unweighted adjacency matrix. The communities detected have been compared to the annotated ones computing the normalized mutual information (NMI). The table reports the mean NMI for nPSO networks with 4 and 8 communities, the mean NMI over all the networks and the mean ranking. The best coalescent embedding technique and the unweighted variant are highlighted. The acronym RA refers to the repulsion-attraction pre-weighting of the coalescent embedding techniques, and not to the link prediction method resource-allocation.

| Method            | nPSO $C = 4$ | nPSO $C = 8$ | mean NMI | mean ranking |
|-------------------|--------------|--------------|----------|--------------|
| EBC-ncISO-HD      | 0.89         | 0.80         | 0.84     | 8.9          |
| EBC-ISO-HD        | 0.88         | 0.79         | 0.84     | 9.1          |
| EBC-LE-HD         | 0.91         | 0.81         | 0.86     | 9.6          |
| RA-LE-HD          | 0.92         | 0.81         | 0.87     | 10.2         |
| EBC-ISO-EA-HD     | 0.81         | 0.73         | 0.77     | 10.5         |
| EBC-ncISO-EA-HD   | 0.80         | 0.73         | 0.77     | 10.6         |
| unweighted        | **0.77**     | **0.71**     | **0.74** | **10.7**     |
| RA-ISO-EA-HD      | 0.81         | 0.73         | 0.77     | 10.7         |
| RA-ISO-HD         | 0.89         | 0.80         | 0.84     | 10.8         |
| RA-ncISO-HD       | 0.89         | 0.79         | 0.84     | 11.0         |
| EBC-LE-EA-HD      | 0.80         | 0.73         | 0.77     | 11.0         |
| RA-ncISO-EA-HD    | 0.80         | 0.73         | 0.77     | 11.0         |
| RA-LE-EA-HD       | 0.80         | 0.73         | 0.76     | 11.4         |
| EBC-MCE-EA-HD     | 0.79         | 0.73         | 0.76     | 11.4         |
| EBC-ncMCE-EA-HD   | 0.80         | 0.73         | 0.76     | 11.4         |
| EBC-MCE-HD        | 0.83         | 0.77         | 0.80     | 11.5         |
| RA-ncMCE-EA-HD    | 0.80         | 0.73         | 0.76     | 11.7         |
| RA-MCE-EA-HD      | 0.79         | 0.73         | 0.76     | 12.0         |
| EBC-ncMCE-HD      | 0.88         | 0.83         | 0.86     | 12.0         |
| RA-ncMCE-HD       | 0.87         | 0.81         | 0.84     | 12.8         |
| RA-MCE-HD         | 0.83         | 0.76         | 0.80     | 12.9         |

increase can be outstanding, going from a NMI of 0.56 up to 0.70. Indeed, Label propagation is the least accurate among the methods and therefore the one with the largest margin for improvement.

Overall, this investigation on nPSO networks confirms that the performance of community detection algorithms can be improved using network embedding information, in agreement with what has been previously demonstrated on real networks by Muscoloni et al [16].

2.3. Comparison of link prediction algorithms on PSO and nPSO networks

We investigated if the nPSO networks could represent a realistic framework also for testing link prediction algorithms. We compared the performance of state-of-the-art approaches [30, 31] (CRA [32–34], RA [35], SPM [36], SBM [37]) across both PSO and nPSO networks generated using diverse parameter combinations. Cannistraci-resource-allocation (CRA) is a mechanistic model which implements a local-topology-based parameter-free deterministic rule for topological link prediction motivated by the local-community-paradigm (LCP) [32–34]; the standard resource-allocation (RA) is instead motivated by the RA process; structural perturbation method (SPM) is a global and model-free approach that relies on a theory similar to the first-order perturbation in quantum mechanics [36]; SBM is a global approach based on general idea of a block model, where the nodes are partitioned into groups and the probability that two nodes are connected depends only on the groups to which they belong [37] (see Methods for details). The aim of this section is to understand whether using synthetic networks (generated by PSO or nPSO) is possible to replicate the same and diverse link predictors’ performance obtained on real complex networks. In case this is not possible, we would like to check whether the community organization present in the nPSO plays an important role to replicate the results on real networks and to facilitate the performance of some algorithms in comparison to the case of networks where communities are not present, such as the ones generated by classical PSO.

In the evaluation framework adopted, 10% of the links are randomly removed from the network and used to test the precision of the link predictors in recovering them. Every method is executed on the network deprived of the randomly removed links and provides a ranking of candidate links. The precision is computed as the proportion of links correctly recovered considering a number of top-ranked candidate links equivalent to the 10% removed. The percentage of removed links is kept at 10% in order to ensure that the original topological structure of the network is minimally perturbed and the topological properties are conserved. This removal and re-prediction procedure is repeated multiple times (100 repetitions for all the methods, only 10 repetitions for...
SBM due to the high computational time) and the average precision is considered to evaluate the performance of each link predictor.

In the PSO networks, as reported in figure 5, the three methods CRA, SPM and RA obtain a comparable precision, with RA performing slightly better (in particular for N = 100) than CRA, which in turn offers a small improvement with respect to SPM for low temperature T = 0.1. SBM is slightly inferior to the other predictors for N = 100, and it decreases performance for larger networks. The fact that RA performs slightly better than CRA in wiring-prediction of synthetic networks generated by a uniform model without communities is expected, because RA does not account for local-community-organization in the network, and therefore it should adhere better than CRA to the community-free structure of the PSO. In the PSO model a node connects to the other nodes in an ‘isotropic’ manner, it means that there is not any connection preference on the left or right side of the angular coordinates. Whereas, in the nPSO model a node connects to the other nodes in an ‘anisotropic’ manner, it means that there is a preference for nodes on the left or right side of the angular coordinates in relation to the direction of localization of the community to which that node belongs. In practice, RA is a weighted version of common-neighbours similarity that penalizes each common neighbour for its degree ‘isotropically’ (in the sense that: taking a common neighbour the penalization is the same for each link that contributes to determine its degree). Instead, CRA penalizes each common neighbour for its degree ‘anisotropically’ (in the sense that: taking a common neighbour the penalization is not effective for links that—although contribute to determine its degree—are connected to other common-neighbours and create a local-community). Hence, if our rationale is correct, we expect that CRA should clearly outperform RA not only on nPSO but also on real networks, and this improvement should emerge especially for growing network size, because a large number of nodes favours the generation of nonuniform topology.

When the link prediction techniques are tested on the nPSO networks, as shown in figures 6–7, it can be noticed that the introduction of the communities leads to a different scenario. For N = 100 (small-size
networks) the performance of CRA, SPM and RA remains overall comparable, with SBM slightly inferior. For increasing network size, and particularly for low temperature, all the methods reduce their precision. However, the decrease is smaller for SBM and CRA, leading to a different ranking, where CRA overcomes the other link prediction approaches regardless of the number of communities adopted. In order to check whether these results resemble a real scenario, we tested the link prediction algorithms also on real networks (SBM only on small-size networks due to the high computational time complexity). Suppl. tables 3–5 summarize the performance of the approaches for small-size and large-size real networks, both in predicting randomly removed links and in predicting links with time information. For small-size networks SPM obtains the highest mean precision-ranking, followed by CRA, SBM and, as last, RA. This result does not approximate well what has been seen in the artificial networks, and the main reason can be that the networks present different characteristics. In particular, most of the small-size real networks do not have a marked power-lawness, as reported in suppl. table 6, whereas the PSO and nPSO networks, at least in their basic implementation, are designed to follow a power-law degree distribution. We emphasize that in the other study in which we theoretically introduce and discuss the nPSO model [20], it has been numerically proven that changing or removing the community structure in the nPSO networks (while keeping fixed the other model parameters) does not significantly affect the main structural features of the network, like clustering coefficient, characteristic path length, power-law exponent, assortativity and LCP-correlation [20]. Looking at the large-size networks, which tend to be scale-free (see suppl. table 7), CRA obtains the best mean precision-ranking, reproducing the results reported for increasing network size on the nPSO networks, which, therefore, seems to offer a more realistic framework with respect to the original PSO model. SBM performance is in general comparable to the one of the other algorithms only on small-size networks (N = 100), whereas for increasing network size SBM considerably loses performance showing a clear problem of inference.
2.4. Link prediction using network embedding information

While the previous section focused on link prediction approaches exploiting only topological information, we are now going to discuss whether adopting also information gained by the embedding in the HS could be beneficial.

The plots in figures 5–7, previously commented, report as a reference also the performance that is obtained if the links are predicted ranking them by the HD or the HSP, computed using the original coordinates of the nodes and without considering the links randomly removed. It can be noticed that in the PSO model the HD performance is slightly lower than CRA for \( T = 0.1 \) and higher than CRA for \( T = [0.3, 0.5] \), whereas in the nPSO model, with the introduction of the communities, the HD performance consistently decreases, since most of the non-existing links within each community are at low HD and therefore are top-ranked. The HSP, instead, provide a quite low precision and always lower than the other methods, which could be expected because the links in the PSO and nPSO networks are established only depending on the HD.

Since the real networks tend to present a community structure, these results obtained on the nPSO suggest that embedding a network in the HS and using the ranked HD for predicting the links will not generally lead to high values of precision.

In order to prove this we applied the coalescent embedding techniques [16], a topological-based machine learning class of algorithms that provides a fast and efficient hyperbolic embedding, and the other main hyperbolic embedding methods: HyperMap-CN [38] and link prediction with community structure (LPCS) [39]. We employ here the same procedure of link prediction evaluation adopted above. The small-size real networks are embedded in the HS after the 10% random link removal. Given the geometrical coordinates of the embedded nodes, we adopted both the HD-ranking and HSP-ranking to re-predict the randomly removed links, the results are reported in suppl. table 8. The maximum average precision offered by the techniques is 0.17, which confirms the expectations independently from the mapping method used. However, if the HSP-ranking is adopted, we notice a general increase of performance with respect to the HD-ranking. From an applied standpoint, this result suggests that on real networks combining both the geometrical and the topological
information (using the HSP) might help to improve the prediction, and actually this is a confirmation of a result already presented by Cannistraci et al [40] on prediction of protein interactions by network embedding. At last, we underline that the small difference in performance between HD and HSP detected in the real networks is better resembled by the nPSO model. In fact, using the PSO model, the difference in performance between HD and HSP is large, and this is markedly in disagreement with the results obtained in real networks.

As discussed for the nPSO networks, we argue that most of the non-existing links within the communities would be characterized by a low HD or HSP, and therefore they tend to be top-ranked, reducing the likelihood to predict the randomly removed (missing) links, which are not necessarily inside the community. In order to provide an explanatory example for this point, we focused on three small-size real networks for which community metadata are available, and we analysed the top-ranked links provided by the hyperbolic embedding techniques after the random link removal. suppl. table 9 reports for each method and each network the percentage of intra-community candidate links predicted in the top-ranking and the percentage of these which corresponds to correct predictions. The results support our argument showing that on average ∼80% of the top-ranked candidate links are intra-community, and that among these only ∼10% represent correct predictions.

To conclude, we would like to discuss the fact that in the generative procedure of the PSO and nPSO models the links are deterministically established between the closest nodes only for the case \( T = 0 \), whereas for \( T > 0 \) with a probability dependent on the HD. Since real networks are hardly characterized by a temperature zero, we argue that the usage of the HD-ranking might be not the best solution for link prediction. Ideally, the connection probability as a function of the HD should be empirically estimated from the length gained by the links when the network is embedded in the hyperbolic disk. Assuming that this well approximates the original connection probability function of the unknown generative model, the HD between non-connected nodes can be converted in connection probabilities according to the probability distribution function inferred from the existing links. Alternatively, as second option, the HD between the nodes could be converted in connection probabilities according to the mathematical formula of the model, after having fitted the temperature parameter of the network. Therefore, instead to adopt a greedy procedure that ranks (smallest values come first) the candidate
links according to their distances in the HS, a correct procedure should sample the candidate links respecting the empirically estimated connection probability of the networked system. However, since the main focus of the article is not link prediction, we procrastinate this investigation to future studies.

3. Conclusion

In this work, the nPSO model [20] has been used as a benchmark for testing four state-of-the-art community detection approaches, and the evaluations through several parameter combinations highlighted two main points: firstly, the communities are always detected with high accuracy by at least one method; secondly, performance limitations arisen in particular conditions for some community detection methods are in agreement with findings produced in previous studies. These points represent the first main result of our study and suggest that the model can be adopted as a valid framework for community detection studies. It is not necessarily intended to replace the previous benchmarks, but potentially to complement them, in particular the LFR model [4] that seems the most realistic among the ones not based on latent geometry.

Furthermore, making tests on nPSO networks, we confirmed that community detection algorithms can be boosted if their input is not a simple unweighted adjacency matrix but a pre-weighted one, where the strength of each link is the inverse of the respective distance in the HS. While this has been previously demonstrated on real networks [16], a confirmation on networks with ground truth community structure (such as nPSO) was necessary because the metadata on the community structure of real networks remains always questionable. Therefore, this represents the second main result of this study.

At last, we tested state-of-the-art link prediction algorithms both in real and artificial networks, and we showed that the ranking of the methods according to their performance on the nPSO model is closer to the one in the real networks with respect to the PSO model.
In addition, we highlighted that embedding a network in the HS and adopting the HD-ranking for suggesting the links more likely to appear will not lead generally to an efficient prediction, pointing out that the usage of a weighted sampling of candidate links according to empirically estimated connection probabilities needs to be investigated.

To conclude, we propose the nPSO model as a valid framework able to generate realistic networks that can be adopted, among the many possibilities, as a reliable benchmark for testing algorithms designed for community detection and link prediction. Although several benchmarks for community detection have been proposed in previous studies, at the best of our knowledge the nPSO model is the first one able to produce realistic networks with a tailored community structure based on a latent geometrical space, and the first able to reproduce results for link prediction evaluation reasonably close to the ones obtained on real networks.

4. Methods

4.1. Community detection

4.1.1. Louvain

The Louvain algorithm [21] 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.

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 used the \( R \) function `multilevel.community`, an implementation of the method available in the \texttt{igraph} package [41].

In this study, the embedding of the network in the HS has been exploited in order to weight the input adjacency matrix. Given the hyperbolic coordinates, the observed links have been weighted using the formula

\[
A_{ij}^{\text{HD}} = \frac{1}{1 + \text{HD}_{ij}},
\]

where \( \text{HD}_{ij} \) is the HD between nodes \( i \) and \( j \). For the Louvain algorithm a further variant has been tested in which also the non-observed links have been weighted using the formula

\[
A_{ij}^{\text{HSP}} = \frac{1}{1 + \text{HSP}_{ij}},
\]

where \( \text{HSP}_{ij} \) is the HSP between nodes \( i \) and \( j \), computed as the sum of the HD over the shortest path.

4.1.2. Infomap

The Infomap algorithm [22] finds the community structure by minimizing the expected description length of a random walker trajectory using the Huffman coding process [42].

It uses the hierarchical map equation (a further development of the map equation, to detect community structures on more than one level) in the form

\[
L(M) = q_r H(Q) + \sum_{i=1}^{m} L(M^i).
\]

\( L(M) \) is the lower bound of the code length to specify a network path of a partitioning \( M \), \( q_r H(Q) \) is the Shannon information at the coarsest level of the partitioning, \( L(M^i) = q^i_r H(Q^i) + \sum_{j=1}^{m} L(M^{i,j}) \) is the lower bound of the code length to specify a network path of a partitioning \( M \) at sublevel \( i \) and \( L(M^{i,j}) \) is the lower bound of the code length at the finest modular level with submap \( M^{i,j} \).

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.

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 used the C implementation released by the authors at http://mapequation.org/code.html.

In this study, the embedding of the network in the HS has been exploited in order to weight the input adjacency matrix. Given the hyperbolic coordinates, the observed links have been weighted using the HD as described for Louvain.

4.1.3. Walktrap

The Walktrap algorithm \cite{23} is based on an agglomerative method for hierarchical clustering: the nodes are iteratively grouped into communities exploiting the similarities between them. The nodes similarities are obtained using random walks and are based on the idea that random walks tend to get trapped into densely connected subgraphs corresponding to communities.

The agglomerative method uses heuristics to choose which communities to merge and implements an efficient way to update the distances between communities. At the end of the procedure a hierarchy of communities is obtained and each level offers a possible partition. The algorithm chooses as final result the partition that maximizes the modularity.

We used the R function \texttt{walktrap.community}, an implementation of the method available in the \textit{igraph} package \cite{41}.

4.1.4. Label propagation

The Label propagation algorithm \cite{24} initializes each node with a unique label and iteratively updates each node label with the one owned by the majority of the neighbours, with ties broken uniformly at random. The update is performed in an asynchronous way and the order of the nodes at each iteration is chosen randomly. As the labels propagate through the network, densely connected groups of nodes quickly reach a consensus on a unique label. The iterative process stops when every node has the same label as the majority its neighbours, ties included. At the end of the procedure the nodes having the same label are grouped together to form a community. Since the aim is not the optimization of an objective function and the propagation process contains randomness, there are more possible partitions that satisfy the stop criterion and therefore the solution is not unique.

We used the R function \texttt{label.propagation.community}, an implementation of the method available in the \textit{igraph} package \cite{41}.

4.1.5. Normalized mutual information

The evaluation of the community detection has been performed using the NMI as in \cite{43}. 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:

\[
\text{NMI} = \frac{I(X, Y)}{H(X)H(Y)}.
\]

If we consider a partition of the nodes in communities as a distribution (probability of one node falling into one community), we can compute the matching between the annotation obtained by the community detection algorithm and the ground truth communities of a network as follows:

\[
H(C_D) = \sum_{h=1}^{n_h} \frac{m_h^D}{N} \log \left( \frac{m_h^D}{N} \right)
\]

\[
H(C_T) = \sum_{l=1}^{n_l} \frac{m_l^T}{N} \log \left( \frac{m_l^T}{N} \right)
\]
\[
I(C_D, C_T) = \sum_h \sum_i \frac{n_{h,i}}{N} \log \left( \frac{n_{h,i}}{n_h n_i} \right)
\]

\[
\text{NMI}(C_D, C_T) = \frac{I(C_D, C_T)}{\sqrt{H(C_D)H(C_T)}}
\]

where:
- \(N\) — number of nodes;
- \(n^D, n^T\) — number of communities detected by the algorithm (D) or ground truth (T);
- \(n_{h,l}\) — number of nodes assigned to the \(h\)th community by the algorithm and to the \(l\)th community according to the ground truth annotation.

We used the MATLAB implementation available at http://commdetect.weebly.com/. As suggested in the code, when \(\sum \frac{n_{h,i}}{n^T} \leq 100\), the NMI should be adjusted in order to correct for chance [44].

4.2. Link prediction methods

4.2.1. Cannistraci-resource-allocation (CRA)

CRA is a local-based, parameter-free and model-based deterministic rule for topological link prediction in both monopartite [32] and bipartite networks [33, 34]. It is based on the LCP which is a bioinspired theory recently proposed in order to model local-topology-dependent link-growth in a class of real complex networks characterized by the development of diverse, overlapping and hierarchically organized local-communities [32].

Being a local-community-based method, it assigns to every candidate interaction a likelihood score looking only at the neighbours nodes and their cross-interactions. In particular, the paradigmatic shift introduced by the LCP is to consider not only the common-neighbours of the interacting nodes but also the links between those common-neighbours, which in practice form all together a local-community.

For each candidate interaction between nodes \(i\) and \(j\), the score is assigned according to the following equation [32]:

\[
\text{CRA}(i, j) = \sum_{k \in \Phi(i) \cap \Phi(j)} \frac{|\varphi(k)|}{|\Phi(k)|},
\]

where:
- \(k\): common neighbour of nodes \(i\) and \(j\)
- \(\Phi(i)\): set of neighbours of node \(i\)
- \(|\Phi(k)|\): cardinality of set \(\Phi(k)\), equivalent to the degree of \(k\)
- \(\varphi(k)\): subset of neighbours of \(k\) that are also common-neighbours of \(i\) and \(j\)
- \(|\varphi(k)|\): equivalent to the local-community degree of \(k\) (see figure 2).

The higher the CRA score, the higher the likelihood that the interaction exists, therefore the candidate interactions are ranked by decreasing CRA scores and the obtained ranking is the link prediction result. The method has been implemented in MATLAB. The code is available at: https://sites.google.com/site/carlovittoriocannistraci/.

4.2.2. Resource-allocation (RA)

RA is a local-based, parameter-free and model-based deterministic rule for topological link prediction [35], motivated by the RA process taking place in networks. Considering a pair of nodes that are not directly connected, one node can send some resource to the other one through their common-neighbours, which play the role of transmitters. It assumes the simplest case where every transmitter equally distributes a unit of resource between its neighbours. For each candidate interaction between nodes \(i\) and \(j\), the score is assigned according to the following equation [35]:

\[
\text{RA}(i, j) = \sum_{k \in \Phi(i) \cap \Phi(j)} \frac{1}{|\Phi(k)|},
\]

where:
- \(k\): common neighbour of nodes \(i\) and \(j\)
- \(\Phi(i)\): set of neighbours of node \(i\)
- \(|\Phi(k)|\): cardinality of set \(\Phi(k)\), equivalent to the degree of \(k\).

The higher the RA score, the higher the likelihood that the interaction exists, therefore the candidate interactions are ranked by decreasing RA scores and the obtained ranking is the link prediction result. The method has been implemented in MATLAB.
4.2.3. Structural perturbation method (SPM)

SPM relies on a theory similar to the first-order perturbation in quantum mechanics [36]. Unlike CH, it is a global approach, meaning that it exploits the information of the complete adjacency matrix in order to compute the likelihood score to assign to every candidate interaction. A high-level description of the procedure is the following:

1. Randomly remove a subset of the edges $\Delta E$ (usually 10%) from the network adjacency matrix $x$, obtaining a reduced adjacency matrix $x^R$.
2. Compute the eigenvalues and eigenvectors of $x^R$.
3. Considering $\Delta E$ as a perturbation of $x^R$, construct the perturbed matrix $\tilde{x}$ via a first-order approximation that allows the eigenvalues to change while keeping fixed the eigenvectors.
4. Repeat steps 1–3 for 10 independent iterations and take the average of the perturbed matrices $\tilde{x}$.

The idea behind the method is that a missing part of the network is predictable if it does not significantly change the structural features of the observable part, represented by the eigenvectors of the matrix. If this is the case, the perturbed matrices should be good approximations of the original network [36]. The entries of the average perturbed matrix represent the scores for the candidate links. The higher the score the greater the likelihood that the interaction exists, therefore the candidate interactions are ranked by decreasing scores and the obtained ranking represents the link prediction result.

The MATLAB implementation of the method has been provided by the authors.

4.2.4. Stochastic block model (SBM)

The framework based on SBM considered in this study has been introduced by Guimerà et al [37] in order to identify both missing and spurious interactions in noisy network observations. The general idea of a block model is that the nodes are partitioned into groups and the probability that two nodes are connected depends only on the groups to which they belong. Assuming that there is no prior knowledge about which partition is more suitable for the observed network, the mathematical formula for obtaining the reliability of an individual link between nodes $i$ and $j$ is [37]:

$$R_{ij} = \frac{1}{Z} \sum_{p \in P} \left( \frac{l_{\alpha i} + 1}{r_{\alpha i}} + 2 \right) \exp[-H(p)],$$

where the sum is over every partition $p$ in the space $P$ of all the possible partitions of the network into groups, $\sigma_i$ is the group of node $i$ in partition $p$, $l_{\alpha i}$ is the number of links between groups $\alpha$ and $\beta$, $r_{\alpha i}$ is the maximum number of possible links between groups $\alpha$ and $\beta$. The function $H(p)$ is:

$$H(p) = \sum_{\alpha \leq \beta} \left[ \ln(r_{\alpha \beta} + 1) + \ln\left(l_{\alpha i}\right) \right]$$

And the normalization factor is:

$$Z = \sum_{p \in P} \exp[-H(p)].$$

However, since the exploration of all the possible partitions is too computationally expensive, the Metropolis algorithm, which is based on a stochastic procedure, is exploited in order to sample only a subset of partitions that are relevant for the estimation of the link reliability [37]. The higher the reliability the greater the likelihood that a non-observed interaction actually exists, therefore the candidate interactions are ranked by decreasing scores and the obtained ranking represents the link prediction result. The C code of the method has been released by the authors and can be download from the website http://seeslab.info/downloads/network-c-libraries-rgraph/.

4.2.5. Coalescent embedding

The expression coalescent embedding refers to a topological-based machine learning class of algorithms that exploits nonlinear unsupervised dimensionality reduction to infer the nodes’ angular coordinates in the HS [16]. The techniques are able to perform a fast and accurate mapping of a network in the 2D hyperbolic disk, the 3D hyperbolic sphere, and potentially also in higher dimensions.

The first step of the algorithm for a 2D embedding consists in weighting the network in order to suggest geometrical distances between connected nodes, since it has been shown that improves the mapping accuracy [16]. If the network is unweighted, the topological-based pre-weighting rules repulsion-attraction (RA) or edge-
betweenness-centrality (EBC) can be applied. The rules are devised to suggest geometrical distances between the connected nodes, using either local (RA) or global (EBC) topological information [16].

Given the weighted network, the second step consists in performing the nonlinear dimensionality reduction. Two different kinds of machine learning approaches can be used, manifold-based (LE, ISO, ncISO) or minimum-curvilinearity-based (MCE, ncMCE). The details about which dimensions of the embedding should be considered are provided in the original publication [16].

In order to assign the angular coordinates in the 2D embedding space, either a circular adjustment or an equidistant angular (EA) adjustment can be considered. The circular adjustment for the manifold-based approaches consists in exploiting directly the polar coordinates of the 2D reduced space, whereas for the MCE-based in rearranging the node points on the circumference following the same ordering of the 1D reduced space and proportionally preserving the distances. Using the EA adjustment, instead, the nodes are equidistantly arranged on the circumference, which might help to correct for short-range angular noise present in the embedding [16].

The radial coordinates are assigned according to a mathematical formula which takes into account both the position of the nodes in their ranking by degree and the scale-freeness of the node degree distribution [45]. The exponent γ of the power-law degree distribution has been fitted using the MATLAB script pflit.m, a procedure described by Clauset et al [46] and released at http://santafe.edu/~aaronc/powerlaws/.

In the link prediction application, the HD are computed between the nodes in the HS and the candidate interactions are ranked by increasing HD, the obtained ranking is the link prediction result.

In a second variant that combines both geometrical and topological information, the network is weighted using the HD and the HSP are computed as sum of the HD over the shortest path between each pair of nodes. The candidate interactions are ranked by increasing HSP and the obtained ranking is the link prediction result. The method has been implemented in MATLAB.

The code is available at https://github.com/biomedical-cybernetics/coalescent_embedding.

4.2.6. HyperMap-CN
HyperMap [45] is a method to map a network into the HS based on maximum likelihood estimation (MLE). For sake of clarity, the first algorithm for MLE-based network embedding in the HS is not HyperMap, but to the best of our knowledge is the algorithm proposed by Boguñá et al in [11]. HyperMap is basically an extension of that method applied to the PSO model [12]. It replays the hyperbolic growth of the network and at each time step i it finds the coordinates of the added node i by maximizing the likelihood that the network was produced by the E-PSO model [45]. According to the MLE procedure, the nodes are added in decreasing order of degree. The radial coordinates depend on the time step i and on the exponent γ of the power-law degree distribution. The angular coordinates, instead, are assigned by maximizing a likelihood function \( L_{i,1} \), with the aim of mapping connected nodes at a low HD and disconnected nodes at a high HD. The maximization is done by numerically trying different angular coordinates in steps of \( 2\pi / N \) and choosing the one that leads to the biggest \( L_{i,1} \).

HyperMap-CN [38] is a further development of HyperMap, where the inference of the angular coordinates is not performed anymore maximizing the likelihood \( L_{i,1} \), based on the connections and disconnections of the nodes, but using another local likelihood \( L_{i,\text{CN}} \), based on the number of common-neighbours between each node i and the previous nodes \( j < i \) at final time. Here the hybrid model has been used, a variant of the method in which the likelihood \( L_{i,\text{CN}} \) is only adopted for the high degree nodes and \( L_{i,1} \) for the others, yielding a shorter running time. Furthermore, a speed-up heuristic and corrections steps can be applied. The speed-up can be achieved by getting an initial estimate of the angular coordinate of a node i only considering the previous nodes \( j < i \) that are i’s neighbours, the maximum likelihood estimation is then performed only looking at an interval around this initial estimate. Correction steps can be used at predefined times \( \tau \) each existing node \( j < i \) is visited and with the knowledge of the rest of the coordinates the angle of \( j \) is updated to the value that maximizes the likelihood \( L_{i,1} \). The C++ implementation of the method has been released by the authors at the website https://bitbucket.org/dk-lab/2015_code_hypermap. In our simulations, neither correction steps nor speed-up heuristic have been used. The input parameter \( \gamma \) has been fitted as described for the coalescent embedding method. The temperature has been set to a default value \( T = 0.1 \).

4.2.7. LPCS
LPCS [47] is a hyperbolic embedding technique that consists of the following steps: (1) detect the hierarchical organization of communities. (2) Order the top-level communities starting from the one that has the largest number of nodes and using the community intimacy index, which takes into account the proportion of edges within and between communities. (3) Recursively order the lower level communities based on the order of the higher-level communities, until reaching the bottom level in the hierarchy. (4) Assign to every bottom level community an angular range of size proportional to the nodes in the community, in order to cover the complete circle with non-overlapping angular ranges. Sample the angular coordinates of the nodes uniformly at random.
within the angular range of the related bottom level community. (5) Assign the radial coordinates as described for the MCA method.

The LPCS code firstly takes advantage of the R function `multilevel.community` for detecting the hierarchy of communities, an implementation of the Louvain method \([21]\) available in the `igraph` package \([41]\), while the following embedding steps have been implemented in MATLAB.

### 4.3. Networks datasets

#### 4.3.1. PSO model

The PSO model \([12]\) is a generative network model recently introduced in order to describe how random geometric graphs grow in the HS. In this model the networks evolve optimizing a trade-off between node popularity, abstracted by the radial coordinate, and similarity, represented by the angular coordinate distance, and they exhibit many common structural and dynamical characteristics of real networks.

The model has five input parameters:

- \(N > 0\), number of nodes in the network;
- \(m > 0\), equal to half of the average node degree;
- \(T \geq 0\), network temperature, which controls the network clustering: the network clustering is maximized at \(T = 0\), it decreases almost linearly for \(T = [0, 1)\) and it becomes asymptotically zero if \(T > 1\);
- \(\beta \in (0, 1]\), popularity fading parameter, or alternatively \(\gamma \geq 2\), exponent of the power-law degree distribution, due to the relationship \(\gamma = 1 + 1/\beta\);
- \(\zeta = \sqrt{-K} > 0\), where \(K\) is the curvature of the hyperbolic plane. Since changing \(\zeta\) rescales the node radial coordinates and this does not affect the topological properties of network \([12]\), in the rest of the article we will consider \(K = -1\).

Building a network in the hyperbolic disk requires the following steps:

1. Initially the network is empty.
2. At time \(i = 1, 2, \ldots, N\) a new node \(i\) appears with radial coordinate \(r_i = 2 \ln(i)\) and angular coordinate \(\theta_i\) uniformly sampled in \([0, 2\pi]\); all the existing nodes \(j < i\) increase their radial coordinates according to \(r_j = r_j + (1 - \beta) r_i\) in order to simulate popularity fading.
3. If \(T = 0\), the new node connects to the \(m\) hyperbolically closest nodes; if \(T > 0\), the new node picks a randomly chosen existing node \(j < i\) and, given that it is not already connected to it, it connects to it with probability

\[
p(i, j) = \frac{1}{1 + \exp\left(\frac{h_{ij} - R_i}{2T}\right)}
\]

repeating the procedure until it becomes connected to \(m\) nodes.

Note that

\[
R_i = r_i - 2 \ln\left[\frac{2T(1 - e^{-(1-\beta)\ln(i)})}{\sin(T\pi) m (1 - \beta)}\right]
\]

is the current radius of the hyperbolic disk, and

\[
h_{ij} = \arccosh(\cosh r_i \cosh r_j - \sinh r_i \sinh r_j \cos \theta_{ij})
\]

is the HD between node \(i\) and node \(j\), where

\[
\theta_{ij} = \pi - |\pi - |\theta_i - \theta_j||
\]

is the angle between these nodes.

4. The growing process stops when \(N\) nodes have been introduced.

#### 4.3.2. nPSO model

The nPSO model \([20]\) is a variation of the PSO model introduced in order to confer to the generated networks an adequate community structure, which is lacking in the original model. Since the connection probabilities are inversely proportional to the HD, a uniform distribution of the nodes over the hyperbolic disk does not create
agglomerates of nodes that are concentrated on angular sectors and that are more densely connected between each other than with the rest of the network. A nonuniform distribution, instead, allows to do it by generating heterogeneity in the angular node arrangement. Given the parameters of the PSO model $(N, m, T, \gamma)$ and a nonuniform probability distribution defined in $[0, 2\pi]$, the procedure to generate a network is the same described in the section for the uniform case, with the only difference that the angular coordinates of the nodes are not sampled uniformly but according to the given nonuniform probability distribution.

In this study, without loss of generality, we will concentrate on Gaussian mixture distributions, with communities that emerge in correspondence of the different components.

A Gaussian mixture distribution is characterized by the following parameters [48]:

- $C > 0$, which is the number of components, each one representative of a community;
- $\mu_1, \ldots, \mu_C \in [0, 2\pi]$, which are the means of the components, representing the central locations of the communities in the angular space;
- $\sigma_1, \ldots, \sigma_C > 0$, which are the standard deviations of the components, determining how much the communities are spread in the angular space; a low value leads to isolated communities, a high value makes the adjacent communities to overlap;
- $\rho_1, \ldots, \rho_C (\sum_i \rho_i = 1)$, which are the mixing proportions of the components, determining the relative sizes of the communities.

Note that, although the means of the components are located in $[0, 2\pi]$, the sampling of the angular coordinate $\theta$ can fall out of this range. In this case, it has to be shifted within the original range using the modulo operator: $\theta = \mod(\theta, 2\pi)$.

Although the parameters of the Gaussian mixture distribution allow for the investigation of disparate scenarios, as a first case of study we focused on the most straightforward setting. For a given number of components $C$, we considered their means equidistantly arranged over the angular space, the same standard deviation and equal mixing proportions:

- $\mu_i = \frac{2\pi}{C} \times (i - 1) i = 1 \ldots C$
- $\sigma_1 = \sigma_2 = \ldots = \sigma_C = \sigma$
- $\rho_1 = \rho_2 = \ldots = \rho_C = \frac{1}{C}$

In particular, in our simulations we fixed the standard deviation to $1/6$ of the distance between two adjacent means $\left(\sigma = \frac{\frac{1}{6} + \frac{2\pi}{C}}{C}\right)$, which allowed for a reasonable isolation of the communities independently from their number. The community memberships are assigned considering for each node the component whose mean is at the lowest angular distance.

4.3.3. Real networks
The real networks have been transformed into undirected and unweighted, self-loops have been removed and the largest connected component has been considered.

Mouse neural: in-vivo single neuron connectome that reports mouse primary visual cortex (layers 1, 2/3 and upper 4) synaptic connections between neurons [49].

Karate: social network of a university karate club collected by Wayne Zachary in 1977. Each node represents a member of the club and each edge represents a tie between two members of the club [50]. Community metadata are available, the communities are formed by a split of the club into two parts, each following one trainer.

Dolphins: a social network of bottlenose dolphins. The nodes are the bottlenose dolphins (genus Tursiops) of a bottlenose dolphin community living off Doubtful Sound, a fjord in New Zealand. An edge indicates a frequent association. The dolphins were observed between 1994 and 2001 [51].

Macaque neural: a macaque cortical connectome, assembled in previous studies in order to merge partial information obtained from disparate literature and database sources [52].

Polbooks: nodes represent books about US politics sold by the online bookseller Amazon.com. Edges represent frequent co-purchasing of books by the same buyers, as indicated by the ‘customers who bought this book also bought these other books’ feature on Amazon. The network was compiled by V Krebs and is unpublished, but can found at http://personal.umich.edu/~mejn/netdata/. Community metadata are
available, the three communities are given by the political orientation of the books as either conservative, neutral or liberal.

ACM2009_contacts: network of face-to-face contacts (active for at least 20 s) of the attendees of the ACM Conference on Hypertext and Hypermedia 2009 [53].

Football: network of American football games between Division IA colleges during regular season Fall 2000 [2].

Physicians innovation: the network captures innovation spread among physicians in the towns in Illinois, Peoria, Bloomington, Quincy and Galesburg. The data was collected in 1966. A node represents a physician and an edge between two physicians shows that the left physician told that the right physician is his friend or that he turns to the right physician if he needs advice or is interested in a discussion [54].

Manufacturing email: email communication network between employees of a mid-sized manufacturing company [55].

Little rock food web: food web of Little Rock Lake, Wisconsin in the United States of America. Nodes are autotrophs, herbivores, carnivores and decomposers; links represent food sources [56].

Jazz: collaboration network between Jazz musicians. Each node is a Jazz musician and an edge denotes that two musicians have played together in a band. The data was collected in 2003 [57].

Residence hall friends: friendship network between residents living at a residence hall located on the Australian National University campus [58].

Haggle contacts: contacts between people measured by carried wireless devices. A node represents a person and an edge between two persons shows that there was a contact between them [59].

Worm nervous: a C Elegans connectome representing synaptic interactions between neurons [60].

Netsci: a co-authorship network of scientists working on networks science [61].

Infectious contacts: network of face-to-face contacts (active for at least 20 s) of people during the exhibition INFECTIOUS: STAY AWAY in 2009 at the Science Gallery in Dublin [53].

Flightmap: a network of flights between American and Canadian cities [62].

Email: email communication network at the University Rovira i Virgili in Tarragona in the south of Catalonia in Spain. Nodes are users and each edge represents that at least one email was sent [63].

Polblog: a network of front-page hyperlinks between blogs in the context of the 2004 US election. A node represents a blog and an edge represents a hyperlink between two blogs [64]. Community metadata are available, the two communities represent the political opinions of the blogs (conservative or liberal).

Odlis: Online Dictionary of Library and Information Science (ODLIS): ODLIS is designed to be a hypertext reference resource for library and information science professionals, university students and faculty, and users of all types of libraries. Version December 2000 [65].

Advogato: a trust network of the online community platform Advogato for developers of free software launched in 1999. Nodes are users of Advogato and the edges represent trust relationships [66].

Arxiv astroph: collaboration graph of authors of scientific papers from the arXiv’s Astrophysics (astro-ph) section. An edge between two authors represents a common publication [67].

Thesaurus: this is the Edinburgh Associative Thesaurus. Nodes are english words, and a directed link from A to B denotes that the word B was given as a response to the stimulus word A in user experiments [68].

Arxiv hep-th: this is the network of publications in the arXiv’s High Energy Physics—Theory (hep-th) section. The links that connect the publications are citations [67].

Facebook: a network of a small subset of posts to user’s walls on Facebook. The nodes of the network are Facebook users, and each edge represents one post, linking the users writing a post to the users whose wall the post is written on [69].

ARK200909-ARK201012: Autonomous systems (AS) Internet topologies extracted from the data collected by the Archipelago active measurement infrastructure (ARK) developed by CAIDA, from September 2009 up to December 2010. The connections in the topology are not physical but logical, representing AS relationships [70].

Most of the networks in the dataset can be downloaded from the Koblenz Network Collection at http://konect.uni-koblenz.de.

Hardware and software

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

We thank Alexander Mestiashvili and the BIOTEC System Administrators for their IT support, Gloria Marchesi for the administrative assistance and the Centre for Information Services and High Performance Computing (ZIH) of the TU Dresden.

Author contributions

CVC conceived the study and designed the numerical experiments, AM implemented the code and performed the computational analysis. Both the authors analysed and interpreted the results. AM wrote the draft of the article according to CVC suggestions and CVC corrected and improved it to arrive to the final draft. CVC designed the figures and AM realized them. CVC planned, directed and supervised the study.

Competing interests

The authors declare no competing financial interests.

Funding

Work in the CVC laboratory 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. We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the TU Dresden.

ORCID iDs

Muscoloni Alessandro https://orcid.org/0000-0002-9238-3357
Cannistraci Carlo Vittorio https://orcid.org/0000-0003-0100-8410

References

[1] Fortunato S and Hric D 2016 Community detection in networks: a user guide Phys. Rep. 659 1–44
[2] Girvan M and Newman M E J 2002 Community structure in social and biological networks Proc. Natl Acad. Sci. 99 7821–6
[3] Danon L, Diaz-Guilera A and Arenas A 2006 Effect of size heterogeneity on community identification in complex networks J. Stat. Mech. 2006 P11010
[4] Lancichinetti A, Fortunato S and Radicchi F 2008 Benchmark graphs for testing community detection algorithms Phys. Rev. E 78 046110
[5] Karrer B and Newman M E J 2011 Stochastic blockmodels and community structure in networks Phys. Rev. E 83 016107
[6] Lancichinetti A and Fortunato S 2009 Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities Phys. Rev. E 80 036110
[7] Ball B, Karrer B and Newman M E J 2011 Efficient and principled method for detecting communities in networks Phys. Rev. E 84 036103
[8] Serrano MÁ, Krioukov D and Boguñá M 2008 Self-similarity of complex networks and hidden metric spaces Phys. Rev. Lett. 100 1–4
[9] Krioukov D, Papadopoulos F, Vahdat A and Boguñá M 2009 Curvature and temperature of complex networks Phys. Rev. E 80 035101
[10] Krioukov D, Papadopoulos F, Kitsak M, Vahdat A and Boguñá M 2010 Hyperbolic geometry of complex networks Phys. Rev. E 82 06108
[11] Boguñá M, Papadopoulos F and Krioukov D 2010 Sustaining the internet with hyperbolic mapping Nat. Commun. 1 1–8
[12] Papadopoulos F, Kitsak M, Serrano M A, Boguñá M and Krioukov D 2012 Popularity versus similarity in growing networks Nature 489 353–40
[13] Kleinberg J-K, Boguñá M, Serrano M Á and Papadopoulos F 2016 Hidden geometric correlations in real multiplex networks Nat. Phys. 12 1076–81
[14] Bianconi G and Rahmede C 2017 Emergent hyperbolic network geometry Sci. Rep. 7 41971
[15] Allard A, Serrano M Á, García-Pérez G and Boguñá M 2016 The hidden geometry of weighted complex networks Nat. Commun. 8 14103
[16] Muscoloni A, Thomas J M, Ciucci S, Bianconi G and Cannistraci C V 2017 Machine learning meets complex networks via coalescent embedding in the hyperbolic space Nat. Commun. 8 1615
[17] Muscoloni A and Cannistraci C V 2018 Minimum curvilinear automata with similarity attachment for network embedding and link prediction in the hyperbolic space arXiv:1802.01183 [physics.soc-ph]
[18] Cacciola A et al 2017 Coalescent embedding in the hyperbolic space unsupervisedly discloses the hidden geometry of the brain arXiv:1705.04192 [q-bio.NC]
[19] Muscoloni A and Cannistraci C V 2017 Local-ring network automata and the impact of hyperbolic geometry in complex network link-prediction arXiv:1707.09496 [physics.soc-ph]
[20] Musoloni A and Cannistraci C V 2018 A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities New J. Phys. 20 020502
[21] Blondel V D, Guillaume J-L, Lambiotte R and Lefebvre E 2008 Fast unfolding of communities in large networks J. Stat. Mech. 2008 P10008
[22] Rosvall M and Bergstrom C T 2011 Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems PLoS One 6 e18209
[23] Pons P and Latapy M 2006 Computing communities in large networks using random walks J. Graph Algorithms Appl. 10 191–218
[24] Raghavan U N, Albert R and Kumara S 2007 Near linear time algorithm to detect community structures in large-scale networks Phys. Rev. E 76 036106
[25] Yang Z, Algesheimer R and Tessone C J 2016 A comparative analysis of community detection algorithms on artificial networks Sci. Rep. 6 30750
[26] Orman G K and Labatut V 2009 A comparison of community detection algorithms on artificial networks Discovery Science (vol 5808) pp 242–56
[27] Lanchinetti A and Fortunato S 2009 Community detection algorithms: a comparative analysis Phys. Rev. E 80 066117
[28] Hric D, Darst R K and Fortunato S 2014 Community detection in networks: structural functionalities versus ground truth Phys. Rev. E 90 062805
[29] Danon L, Díaz-Guilera A, Duch J and Arenas A 2005 Comparing community structure identification J. Stat. Mech. 1–10
[30] Pech R, Hao D, Pan L, Cheng H and Zhou T 2017 Link prediction via matrix completion Europhys. Lett. 117 38002
[31] Wang W, Cai F, Jiao P and Pan L 2016 A perturbation-based framework for link prediction via non-negative matrix factorization Sci. Rep. 6 38938
[32] Cannistraci C V, Alanis-Lobato G and Ravasi T 2013 From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks Sci. Rep. 3 1–13
[33] Danimelli S, Thomas J M, Durán C and Cannistraci C V 2015 Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks New J. Phys. 17 113037
[34] Durán C, Danimelli S, Thomas J M, Haupt V J, Schroeder M and Cannistraci C V 2017 Pioneering topological methods for network-based drug-target prediction by exploiting a brain-network self-organization theory Brief. Bioinform. 8 3–62
[35] Zhou T, Lu L and Zhang Y-C 2009 Predicting missing links via local information Eur. Phys. J. B 71 623–30
[36] Lü L, Pan L, Zhou T, Zhang Y-C and Stanley H E 2013 Toward link predictability of complex networks Proc. Natl Acad. Sci. 112 2325–30
[37] Guimerà R and Sales-Pardo M 2009 Missing and spurious interactions and the reconstruction of complex networks Proc. Natl Acad. Sci. USA 106 22073–8
[38] Papadopoulos F, Áldecoa R and Krioukov D 2013 Network geometry inference using common neighbors Phys. Rev. E 92 228007
[39] Wang Z, Li Q, Jin F, Xiong W and Wu Y 2016 Hyperbolic mapping of complex networks based on community information Physica A 455 1–29
[40] Cannistraci C V, Alanis-Lobato G and Ravasi T 2013 Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding Bioinformatics 29 199–209
[41] Csardi G and Nepusz T 2006 The igraph software package for complex network research InterJ. Complex Syst. 1695
[42] Huffman D A 1952 A method for the construction of minimum-redundancy codes Proc. IRE 40 1098–101
[43] Danon L, Díaz-Guilera A, Duch J, Díaz-Guilera A and Arenas A 2005 Comparing community structure identification J. Stat. Mech. 2005 P09008
[44] Vinh N X, Epps J and Bailey J 2010 Information theoretic measures for clustering comparison: variants, properties, normalization and correction for chance J. Mach. Learn. Res. 11 2837–54
[45] Papadopoulos F, Psomas C and Krioukov D 2015 Network mapping by replaying hyperbolic growth IEEE/ACM Trans. Netw. 23 198–211
[46] Clauset A, Rohilla Shalizi C and Newman M E J 2009 Power-law distributions in empirical data SIAM Rev. 51 661–703
[47] Wang Z, Wu Y, Li Q, Jin F and Xiong W 2016 Link prediction based on hyperbolic mapping with community structure for complex networks Physica A 450 69–73
[48] McLachlan G and Peel D 2000 Finite Mixture Models (New York: Wiley)
[49] Bock D D et al 2011 Network anatomy and in vivo physiology of visual cortical neurons Nature 471 177–82
[50] Zachary W W 1977 An information flow model for conflict and fission in small groups J. Anthropol. Res. 33 452–73
[51] Lusseau D, Schneider K, Boisseau O J, Haase P, Slooten E and Dawson S M 2007 The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: can geographic isolation explain this unique trait? Behav. Ecol. Sociobiol. 54 396–405
[52] Kötter R 2004 Online retrieval, processing, and visualization of primate connectivity data from the CoCoMac database Neuroinformatics 2 127–44
[53] Isella L, Stehle J, Barrat A, Cattuto C, Pinton J-F and Van den Broeck W 2011 What’s in a crowd? Analysis of face-to-face behavioral networks J. Theor. Biol. 271 166–80
[54] Coleman J, Katz E and Menzel H 1957 The diffusion of an innovation among physicians Sociometry 20 253–70
[55] Michalski R, Palus S and Kazienko P 2011 Matching organizational structure and social network extracted from email communication Bus. Inf. Syst. 87 197–206
[56] Martinez N D 1991 Artifacts or attributes? Effects of resolution on the Little Rock Lake food web Ecol. Monogr. 61 367–92
[57] Geiser P M and Danon L 2003 Community structure in jazz Adv. Complex Syst. 6 365–73
[58] Freeman L C, Webster C M and Kirke D M 1998 Exploring social structure using dynamic three-dimensional color images Soc. Netw. 20 109–18
[59] Chaintreau A, Hui P, Crowcroft J, Dietz C, Gass R and Scott J 2007 Impact of human mobility on opportunistic forwarding algorithms IEEE Trans. Mob. Comput. 6 606–20
[60] Watts D J and Strogatz S H 1998 Collective dynamics of ‘small-world’ networks Nature 393 440–2
[61] Newman M E J 2006 Finding community structure in networks using the eigenvectors of matrices Phys. Rev. E 74 036104
[62] Frey B J and Dueck D 2007 Clustering by passing messages between data points Science 315 972–6
[63] 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 68 1–4
[64] Adams I A and Glance N 2005 The political blogosphere and the 2004 US election: divided they blog LinkKDD 2005 pp 36–43
[65] Reitz J M 2002 Online Dictionary for Library and Information Science
[66] Massa P, Salvetti M and Tomasoni D 2009 Bowling alone and trust decline in social network sites 8th IEEE Int. Symp. on Dependable, Autonomic and Secure Computing, DASC 2009 pp 658–63

[67] Leskovec J, Kleinberg J and Faloutsos C 2007 Graph evolution: densification and shrinking diameters ACM Trans. Knowl. Discovery Data 1 2

[68] Kiss G R, Armstrong C, Milroy R and Piper J 1973 An associative thesaurus of English and its computer analysis The Computer and Literary Studies ed A J Aitkin et al (Edinburgh: Edinburgh University Press)

[69] Viswanath B, Mislove A, Cha M and Gummadi K P 2009 On the evolution of user interaction in facebook Proc. 2nd ACM Work. Online Soc. networks—WOSN ’09 p 57

[70] Claffy K, Hyun Y, Keys K, Fomenkov M and Krioukov D 2009 Internet mapping; from art to science CATCH 2009 pp 205–11