Journal of Statistical Physics
https://doi.org/10.1007/s10955-018-2084-z

Soft Communities in Similarity Space

Guillermo García-Pérez[1,2] · **M. Ángeles Serrano**[1,2,3] · **Marián Boguñá**[1,2]

Received: 28 November 2017 / Accepted: 13 June 2018

Abstract
The $\mathbb{S}^1$ model has been central in the development of the field of network geometry. It places nodes in a similarity space and connects them with a likelihood depending on an effective distance which combines similarity and popularity dimensions, with popularity directly related to the degrees of the nodes. The $\mathbb{S}^1$ model has been mainly studied in its homogeneous regime, in which angular coordinates are independently and uniformly scattered on the circle. We now investigate if the model can generate networks with targeted topological features and soft communities, that is, inhomogeneous angular distributions. To that end, hidden degrees must depend on angular coordinates, and we propose a method to estimate them. We conclude that the model can be topologically invariant with respect to the soft-community structure. Our results expand the scope of the model beyond the independent hidden variables limit and can have an important impact in the embedding of real-world networks.

Keywords Complex networks · Hidden metric spaces · Similarity space · Communities

1 Introduction

Complex networks have been widely studied in the last twenty years in many different contexts, from biology to the social sciences and technology [1,2]. This transversality comes from the fact that the topology of networks is characterised by universal features. For instance, most networks are scale-free, meaning that their degrees are power-law distributed, a phenomenon that was explained in early times of network theory by the preferential attachment mech-

✉ M. Ángeles Serrano
marian.serrano@ub.edu

Guillermo García-Pérez
guille.garcia@ub.edu

Marián Boguñá
marian.boguna@ub.edu

1 Departament de Física de la Matèria Condensada, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain

2 Universitat de Barcelona Institute of Complex Systems (UBICS), Universitat de Barcelona, Barcelona, Spain

3 ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

Published online: 19 June 2018

anism: as the network grows, new nodes connect to highly connected—or popular—nodes with higher probability [3].

However, preferential attachment alone cannot explain the high level of clustering—the fraction of existing triangles—observed in real systems. To explain clustering, the concept of similarity was introduced [4]. The basic idea is that nodes connect not only because they are popular, but also because they are similar. Thus, if node A connects to nodes B and C because they are similar to A, B and C should also be similar and therefore have a high probability of being connected. This *transitivity of similarity* suggests to encode similarities between nodes as distances in metric spaces, since the triangle inequality is one of their defining properties: if the distance d_{BC} in the underlying metric space measures the dissimilarity between B and C, it must be bounded by $d_{BC} \leq d_{AB} + d_{AC}$, therefore inducing the observed transitive connections.

The $\mathbb{S}^1$ model was proposed based on these ideas [4]. In this model, N nodes are randomly scattered into a circle of radius $R = N/2\pi$ to keep the density constant. Every node i is also assigned a hidden degree κ_i from a distribution (for instance, a power-law $P(\kappa) \sim \kappa^{-\gamma}$), and every pair of nodes i and j is connected with probability

$$p_{ij} = \frac{1}{1 + \left(\frac{d_{ij}}{\mu \kappa_i \kappa_j}\right)^{\beta}}, \tag{1}$$

where d_{ij} is the distance along the circle; μ and β are two global parameters controlling the average degree and the clustering coefficient, respectively. Notice that this connection probability takes the form of a gravity law, as it increases with the product of hidden degrees (popularities) and decreases with the distance between them (dissimilarity). Despite its apparent simplicity, this model generates networks that reproduce very well real networks; they are scale-free, small-world and have high levels of clustering. The reasons for this capability to reproduce observed features of real-world networks can be summarized as follows. On the one hand, it can be shown [4] that, for networks generated according to Eq. (1), the degrees k_i are proportional to the hidden degrees κ_i, so the model is versatile enough to generate networks with different degree distributions, including power laws[1]. However, the key ingredient that provides the model with control over the level of clustering is the underlying geometry; the dependence on the distances d_{ij} along the circle allows nodes to establish connections with other nodes with low expected degree when they are close in the space. Since, as previously discussed, the triangle inequality guarantees the transitivity of this closeness, triangles emerge naturally. Moreover, parameter β determines the level of randomness, and therefore clustering, in the model—as $\beta \to \infty$, the connection probability becomes $p_{ij} = 1$ if $\frac{d_{ij}}{\mu \kappa_i \kappa_j} \leq 1$ and 0 otherwise, whereas for low values of β, the connection probability for $\frac{d_{ij}}{\mu \kappa_i \kappa_j} > 1$ becomes non-zero, thus allowing long-range connections, that is, among distant nodes, to appear.

The power of the $\mathbb{S}^1$ model goes beyond generating realistic networks. The coordinates in the similarity space of the nodes of a real network can be inferred by finding the coordinates that maximise the likelihood for the real network to be generated by the model [6–8]. This embedding process yields a map of the network strikingly meaningful. For instance, it allows

[1] The original model defined in [4] is in fact more general, allowing for any connection probability p_{ij} as long as it depends on the argument $d_{ij}/(\kappa_i \kappa_j)^{1/D}$, where the space is the D-dimensional sphere and d_{ij} the geodesic distance on the sphere. The particular functional form in Eq. (1) allows us to interpret the network as a set of non-interacting fermions (the links) embedded in the hyperbolic plane, with the hyperbolic length of a link playing the role of its energy and β playing the role of the inverse of the bath temperature [5].

to navigate the network efficiently by projecting degrees and angles to coordinates in hyperbolic space [6]. Moreover, the maps open the path to a completely new way of analysing complex networks. For example, in Refs. [6–8] it was shown that the angular coordinates of nodes in real-world networks are not uniformly distributed. Instead, they are scattered in a heterogeneous manner, forming *soft communities*.

Soft communities are defined as angular regions more densely populated than others. That is, the soft-community structure of a network is the partition of nodes into groups such that the average angular distances between nodes belonging to the same soft community are smaller than the corresponding expected average in the bare $\mathbb{S}^1$ model. This, in turn, implies that the average angular distance between nodes belonging to different groups must be larger than expected in the homogeneous version of the model and, as a result, soft communities are separated by large angular gaps in the underlying metric space. Interestingly, the topological community structure of the network, as defined for instance in [9–11], is highly correlated with such soft communities. Indeed, partitioning the network using the Critical Gap Method (CGM) algorithm, a geometric method based on the largest gaps between consecutive nodes along the circle as community boundaries [8], gives a modularity comparable to that of other community detection methods currently available in the literature, but with higher resolution.

In this paper, we address the question of whether the $\mathbb{S}^1$ model can generate networks with given target topological features and soft communities, that is, *inhomogeneous* angular distributions. As we show in the following sections, considering heterogeneous angular distributions requires some level of adjustment over the hidden degrees. We propose a general algorithm to find the corrected hidden-degree distribution yielding the desired degree distribution for *any* angular distribution of nodes. Even though the method is general enough to be applied on different angular distributions, in this work we illustrate our results on a particular angular distribution: the angular distribution of the Geometric Preferential Attachment (GPA) model [12]. This model is a generalised version of the growing geometric model Popularity vs. Similarity Optimization [13] in which soft communities, as they named these denser angular regions, emerge from the growth dynamics of the network without altering topological properties like the degree distribution or the clustering spectrum of the resulting network. The main reason why we choose to use this angular distribution in the present paper is that the angular distribution from Ref. [12] is not an arbitrary choice, but it rather emerges from a preferential attachment process in similarity space that seems to be a plausible explanation for the nature of communities in real systems. However, we would like to emphasize that our model is not a generalization of the GPA model. Our algorithm can be used with other angular distributions—for instance, a predefined angular density, as in Ref. [14].

2 Results

In the bare $\mathbb{S}^1$ model, hidden degrees and similarity coordinates are typically assumed to be uncorrelated, so every node's hidden variables are withdrawn independently from some joint distribution $\rho(\kappa, \theta)$ that factorises [4,15]. In contrast, the GPA is a growing model in which the angular coordinates and hidden degrees of different nodes are correlated. In the GPA growth process, the degree of a node is determined by its age—the older the node is, the higher its degree. Moreover, when a new node t is added to the system, the probability for it to be placed at polar coordinate θ_t depends on the number of nodes $s < t$ at angular distance $\Delta\theta_{st} < 2/(s^{\frac{1}{\gamma-1}} t^{\frac{\gamma-2}{\gamma-1}})$, where γ is the exponent of the power-law degree distribution. This

implies a very particular dependence between similarity coordinates and degrees: the angular coordinate of a node must depend on the angular coordinates of all nodes with higher degree.

Following a similar strategy to design a version of the $\mathbb{S}^1$ model able to generate soft communities, we must consider the impact of the degrees of the nodes in their implicit ordering in the circle. A potential way of action would proceed by first assigning a hidden degree κ_i from a power-law distribution $P(\kappa) \sim \kappa^{-\gamma}$ to every node, ordering the nodes according to their hidden degrees, and reproducing the angular preferential attachment from the GPA model with that particular ordering. At the end of the process, we would obtain a set of N nodes with hidden degrees power-law distributed with exponent γ and the same angular distribution as the GPA model for that value of γ. However, the problem is not that simple and if we then connected every pair of nodes with the probabilities given by Eq. (1), degrees and hidden degrees would not be proportional, which is a desirable property, as it allows us to control the degree distribution. The reason for such deviation from the usual behaviour of the model is that a homogeneous angular distribution is required for the proportionality between hidden and observed degrees [15], which is not fulfilled here by construction. The solution is that hidden degrees must depend on the spatial distribution of nodes as well. In the following subsection, we address this issue. We explore the inhomogeneous similarity regime of the $\mathbb{S}^1$ model and show that it is capable of generating networks with power-law degree distributions, high clustering and soft communities.

2.1 Geometric Preferential Attachment in the $\mathbb{S}^1$ model

From the previous discussion, we see that hidden degrees and angles should be considerably entangled in the modelling of geometric networks with soft communities. In this context, the soft-communities $\mathbb{S}^1$ model requires the following steps:

1. *Assigning angular coordinates*: Angular coordinates are assigned according to the Geometric Preferential Attachment. First, a label $i = 1, \ldots, N$ is assigned to every node. Then, for every i from 1 to N:
   (a) Sample i candidate angular positions $\phi_l, l = 1, \ldots, i$ for node i from $U(0, 2\pi)$.
   (b) For every candidate position, define its attractiveness $A(\phi_l)$ as the number of nodes—belonging to the set of s nodes with an already defined angular position, that is, with $s < i$—at angular distance $\Delta\theta_{ls} < 2/(s^{\frac{1}{\gamma-1}} i^{\frac{\gamma-2}{\gamma-1}})$, where γ is the exponent of the target power-law degree distribution.
   (c) Assign to node i the angular coordinate of position l, i. e. set $\theta_i = \phi_l$, with probability
   $$\Pi(\phi_l) = \frac{A(\phi_l) + \Lambda}{\sum_{n=1}^{i} (A(\phi_n) + \Lambda)}. \tag{2}$$
   The initial attractiveness $\Lambda \geq 0$ is a parameter that sets the strength of the geometric preferential attachment. For very high values of Λ, all candidate angles become equally likely, so the resulting angular distribution is homogeneous as in the standard $\mathbb{S}^1$ model (see Fig. 1).

   This process generates a distribution of nodes in the circle analogous to the angular distribution of the GPA model. However, notice that, in the GPA model, connections are established *at the same time* as positions are decided, whereas in the former steps, no connections have yet been made.
2. *Assigning hidden degrees*: Once every node has a defined angular position, we need to determine its hidden degree such that the resulting observed degrees, that is, after the

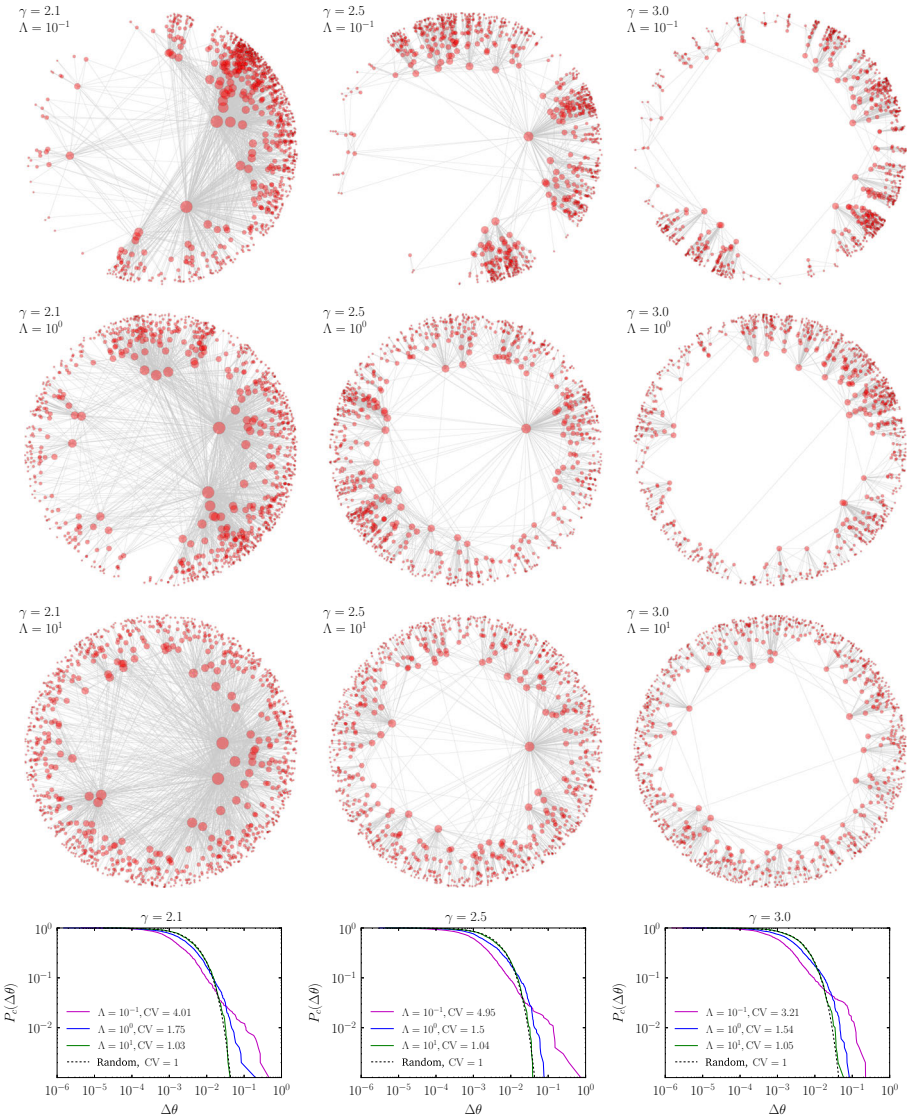

Fig. 1 Geometric layout of the networks generated by the soft-communities $\mathbb{S}^1$ model with Geometric Preferential Attachment. In all cases, $N = 1000$ and $\beta = 2.5$. Every column corresponds to a value of γ and every row to a value of Λ. As in Ref. [12], soft communities emerge for low values of the initial attractiveness Λ. In order to clarify the figure, every node's target degree is represented as a radial coordinate $r_i = R - 2\ln k_i^{\text{tar}}/k_N^{\text{tar}}$, where k_N^{tar} is the smallest target degree and $R = 2\ln\left(N/(\pi\mu(k_N^{\text{tar}})^2)\right)$. When using the hidden degrees instead of the target degrees, this mapping constitutes the isomorphism between the $\mathbb{S}^1$ model and the $\mathbb{H}^2$ model in hyperbolic space [6–8,15]. The bottom row shows the distributions of gaps for the three values of Λ considered. For high values of the initial attractiveness (green curves), the distribution is almost exponential, as in the case of a uniform angular distribution (dashed line), whereas it becomes fat-tailed for low values of Λ. The legend also displays the coefficient of variation, defined as $\text{CV} = \sigma/\mu$, where σ is the standard deviation and μ is the mean of the distribution of gaps. This quantity is equal to 1 for an exponential distribution and increases for increasing angular heterogeneity (Color figure online)

connections have been actually established, are power-law distributed with exponent γ. As mentioned earlier in this paper, we must take into account that the spatial distribution is inhomogeneous (especially for low values of Λ). Indeed, if two nodes with the same hidden degree κ have local geometric neighbourhoods with different densities, they will have different degrees simply because the distances d_{ij} in the connection probability, Eq. (1), are smaller on average in the higher-density neighbourhood. We propose the following method to correct every hidden degree κ_i accordingly:

(a) Generate a set of N target degrees k^{tar} from a power-law distribution with exponent γ. Order the target degrees such that $k_1^{\text{tar}} > k_2^{\text{tar}} > \cdots > k_N^{\text{tar}}$.
(b) Assign to every node i a hidden degree κ_i, where i corresponds to the labelling in step 1, initially set to $\kappa_i = k_i^{\text{tar}}$.
(c) Repeat N times:
   i. Choose some node i randomly.
   ii. Compute the expected degree $\bar{k}_i$ of node i as

$$\bar{k}_i = \sum_{j \neq i} \frac{1}{1 + \left(\frac{d_{ij}}{\mu \kappa_i \kappa_j}\right)^{\beta}}. \tag{3}$$

   iii. Correct the value of κ_i so that the expected degree $\bar{k}_i$ matches the target degree k_i^{tar}. We propose to reset $|\kappa_i + \left(k_i^{\text{tar}} - \bar{k}_i\right)\delta| \to \kappa_i$, where δ is a random variable withdrawn from the uniform distribution $U(0, 0.1)$. The random fluctuations of δ prevents the system from getting stuck at local optima. Other numerical methods could be used with the same end.
(d) Compute all relative deviations

$$\epsilon_i = \frac{|k_i^{\text{tar}} - \bar{k}_i|}{k_i^{\text{tar}}}. \tag{4}$$

If $\max\{\epsilon_i\}_i < \eta$, where η is a tolerance which we set to $\eta = 10^{-2}$, continue to step 3. Otherwise, go back to step 2c.

3. *Generating the network with the* $\mathbb{S}^1$ *model*: In the last step, we simply connected every pair of nodes with the probabilities in Eq. (1). Since step 2 assigns a hidden degree to every node such that its expected degree matches its target degree, the resulting observed degrees in the network must be similar to the target degrees as well.

Figure 1 shows the networks generated by the model for different values of γ and Λ. As in Ref. [12], the angular distribution has an evident soft-community structure for low values of Λ, whereas for high values of the initial attractiveness, the angular density resembles that of the homogeneous $\mathbb{S}^1$ model. Despite the considerable differences in the similarity distances between nodes for different values of Λ, the displayed networks are extremely similar from a topological perspective (see Fig. 2), with almost undistinguishable degree distributions and clustering spectra. Notice that step 2 in the algorithm above, corresponding to the assignment of degrees, is not specifically designed for the GPA angular distribution[2]. In principle, it should be valid for other distributions as well.

[2] The ordering of the target degrees might not be necessary in a more general situation where, for instance, hidden degrees are not correlated with angles.

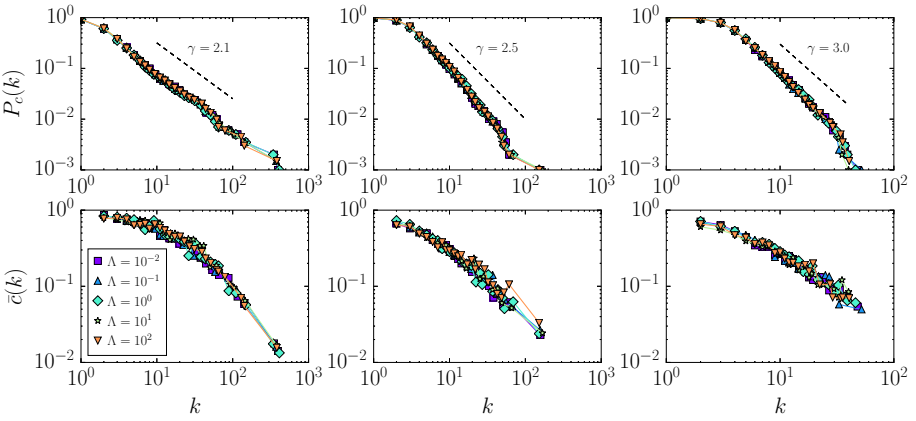

Fig. 2 Topological properties of the networks generated by the $\mathbb{S}^1$ model with Geometric Preferential Attachment with $N = 1000$ and $\beta = 2.5$. Every color corresponds to a different value of the initial attractiveness Λ. The top row shows the complementary cumulative degree distribution $P_c(k) = \int_k^\infty P(k)\mathrm{d}k$, which behaves as $P_c(k) \sim k^{1-\gamma}$ for $P(k) \sim k^{-\gamma}$. Black dashed lines indicate such scaling. In the bottom row, the mean local clustering spectra $\bar{c}(k)$ are drawn. To avoid fluctuations in the target degrees, all networks with the same exponent γ have been generated with the same target degree sequence $\{k_i^{\mathrm{tar}}\}$. Despite their angular distributions being completely different (Fig. 1), their topologies are extremely similar (Color figure online)

3 Discussion

There is abundant evidence of the geometric origin of many properties of complex networks, not only regarding their topology [6–8,16,17], but also their weighted organisation [18]. The field of network geometry has therefore attracted much attention recently, and the $\mathbb{S}^1$ model is one of its cornerstones. On the one hand, it provides an intuitive and plausible explanation for clustering in real networks by introducing the concept of similarity space. On the other hand, it allows to build geometric maps of real networks by embedding them. These maps are remarkably meaningful, to the extent of predicting symmetries in real systems [17]. In addition, they are very useful; they can be used to navigate the network efficiently [6], to detect communities [6,8] or even to construct smaller-scale replicas of real networks for efficiently testing dynamics on real networks [17].

So far, the $\mathbb{S}^1$ model has only been studied under several simplifying premises, like power-law degree distributions or independent hidden variables. Yet, it has been able to explain many observed phenomena in complex networks. However, it can be exploited beyond these assumptions, since the correlation between hidden degrees and angles might clarify many more topological features of real-world networks. This work opens the path towards such line of study by showing that the model *does not require* those simplifying assumptions. As a result, we can now generate complex networks with soft communities for any given angular distribution; although in this paper we use as particular example the GPA angular distributions, our method is valid for other angular distributions as well.

Moreover, the results presented in this paper might also have an impact on the embedding of real networks. Typically, the likelihood maximisation procedure only seeks the best angular coordinates, whereas hidden degrees are considered to be a function of degree only and known from the start [6,7]. This hypothesis is a direct consequence of the aforementioned simplifying assumptions usually contemplated in the $\mathbb{S}^1$ model. Nevertheless, as we have shown in this work, an inhomogeneous angular distribution requires correcting hidden degrees in such a way that they depend on the hidden variables of all other nodes. This is a very important

result, since it suggests that inferring hidden degrees via likelihood maximisation as well as angles might noticeably improve the quality of embeddings of real-world networks with community structure.

Acknowledgements We acknowledge support from a James S. McDonnell Foundation Scholar Award in Complex Systems; the ICREA Academia prize, funded by the Generalitat de Catalunya; Ministerio de Economía y Competitividad of Spain Projects No. FIS2013-47282-C2-1-P and no. FIS2016-76830-C2-2-P (AEI/FEDER, UE).

References

1. Newman, M.E.J.: Networks: An Introduction. Oxford University Press, Oxford (2010)
2. Dorogovtsev, S.N., Mendes, J.F.F.: Accelerated Growth of Networks. Handbook of Graphs and Networks. Wiley-VCH, Berlin (2003)
3. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**, 509 (1999)
4. Serrano, M.Á., Krioukov, D., Boguñá, M.: Self-similarity of complex networks and hidden metric spaces. Phys. Rev. Lett. **100**, 078701 (2008)
5. Krioukov, D., Papadopoulos, F., Vahdat, A., Boguñá, M.: Hyperbolic geometry of complex networks. Phys. Rev. E **80**, 035101 (2009)
6. Boguñá, M., Papadopoulos, F., Krioukov, D.: Sustaining the internet with hyperbolic mapping. Nat. Commun. **1**, 62 (2010)
7. Serrano, M.Á., Boguñá, M., Sagués, F.: Uncovering the hidden geometry behind metabolic networks. Mol. BioSyst. **8**, 843 (2012)
8. García-Pérez, G., Boguñá, M., Allard, A., Serrano, M.Á.: The hidden hyperbolic geometry of international trade: World Trade Atlas 1870–2013. Sci. Rep. **6**, 33441 (2016)
9. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004)
10. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proc. Natl. Acad. Sci. USA **101**, 2658 (2004)
11. Arenas, A., Fernández, A., Gómez, S.: Analysis of the structure of complex networks at different resolution levels. New J. Phys. **10**(5), 053039 (2008)
12. Zuev, K., Boguñá, M., Bianconi, G., Krioukov, D.: Emergence of soft communities from geometric preferential attachment. Sci. Rep. **5**, 9421 (2015)
13. Papadopoulos, F., Kitsak, M., Serrano, M.Á., Boguñá, M., Krioukov, D.: Popularity versus similarity in growing networks. Nature **489**(7417), 537 (2012)
14. Muscoloni, A., Cannistraci, C.V.: A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities. New J. Phys. **20**, 052002 (2018)
15. Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., Boguñá, M.: Hyperbolic geometry of complex networks. Phys. Rev. E **82**, 036106 (2010)
16. Gulyás, A., Bíró, J.J., Kőrösi, A., Rétvári, G., Krioukov, D.: Navigable networks as Nash equilibria of navigation games. Nat. Commun. **6**, 7651 (2015)
17. García-Pérez, G., Boguñá, M., Serrano, M.A.: Multiscale unfolding of real networks by geometric renormalization. Nat. Phys. **14**, 583–589 (2018)
18. Allard, A., Serrano, M.Á., García-Pérez, G., Boguñá, M.: The geometric nature of weights in real complex networks. Nat. Commun. **8**, 14103 (2017)