SCIENTIFIC REPORTS
natureresearch

OPEN

Interplay between *k*-core and community structure in complex networks

Irene Malvestio[1], Alessio Cardillo[2,1,3] & Naoki Masuda[4,5,1]

The organisation of a network in a maximal set of nodes having at least *k* neighbours within the set, known as *k*-core decomposition, has been used for studying various phenomena. It has been shown that nodes in the innermost *k*-shells play a crucial role in contagion processes, emergence of consensus, and resilience of the system. It is known that the *k*-core decomposition of many empirical networks cannot be explained by the degree of each node alone, or equivalently, random graph models that preserve the degree of each node (i.e., configuration model). Here we study the *k* -core decomposition of some empirical networks as well as that of some randomised counterparts, and examine the extent to which the *k*-shell structure of the networks can be accounted for by the community structure. We find that preserving the community structure in the randomisation process is crucial for generating networks whose *k*-core decomposition is close to the empirical one. We also highlight the existence, in some networks, of a concentration of the nodes in the innermost *k*-shells into a small number of communities.

Whenever a system can be abstracted as a set of units (*nodes*) interacting in pairs (*edges*), we can describe it as a network (also called a graph). Network analysis has proven to be a valuable framework to aid us to understand a plethora of phenomena taking place in many complex systems. Examples include cascades and collective behaviour in socio-technical systems, the emergence of cognitive functions in neural systems, the stability of chemical/biological systems, and the shape of spatially embedded systems, to cite a few[1–3].

One of the advantages of the network representation is the possibility to probe the system in a coarse-grained manner, going beyond dyadic interactions by identifying high-order structures of the network[4,5]. Examples include tightly connected groups of nodes, i.e., communities[6], multiscale coarse-grained structures[7], core-periphery structure[8–10], nested assembly of nodes[11], rich clubs[12,13], and the *k*-core[14,15].

The *k*-core decomposition of a network is the maximal set of nodes that have at least *k* neighbours within the set[14,15]. The algorithm to extract the *k*-core consists in recursively removing the nodes having less than *k* connections. A *k*-shell is defined as the set of nodes belonging to the k^{th} core but not to the $(k+1)^{\text{th}}$ core[15]. The *k*-core decomposition has proven to be useful in a variety of domains such as identifying and ranking the most influential spreaders in networks, identifying keywords used for classifying documents, and in assessing the robustness of mutualistic ecosystem and protein networks[16,17].

Models to generate random networks with specific features should help us to understand how the mechanisms governing the establishment of edges account for properties of empirical networks. Despite the vast range of applications of the *k*-core decomposition, to the best of our knowledge, there have been only few attempts to build models to generate networks with a given *k*-core structure. One indirect attempt to generate networks with a given *k*-core decomposition is the so-called BRITE model[18]. Originally, this model sought to replicate the features (including the *k*-core) of the Internet network at the Autonomous System (AS) level by mixing the mechanism of growth with preferential attachment[19,20] and that of adding edges between already existing nodes. Another model aimed at generating networks with a *k*-core structure akin to an empirical one by leveraging the information stored in the so-called core fingerprint[21]. The core fingerprint corresponds to knowing the number of nodes in

[1]Department of Engineering Mathematics, University of Bristol, Bristol BS8 1UB, UK. [2]Department of Computer Science and Mathematics, University Rovira i Virgili, 43007 Tarragona, Spain. [3]GOTHAM Lab – Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, 50018 Zaragoza, Spain. [4]Department of Mathematics, University at Buffalo, Buffalo, NY 14260-2900, United States. [5]Computational and Data-Enabled Science and Engineering Program, University at Buffalo, State University of New York, Buffalo, NY 14260-5030, USA. ✉email: alessio.cardillo@urv.cat; naokimas@buffalo.edu

https://doi.org/10.1038/s41598-020-71426-8

each k-shell, the number of intra-shell edges (i.e., those connecting nodes belonging to the same k-shell), and the number of inter-shell edges (i.e., those connecting nodes belonging to different k-shells) of a given network. Moreover, the authors qualitatively compared the Internet AS networks and synthetic networks preserving the core fingerprint of the original networks using several indicators[21]. More recently, models based on modified versions of the so-called configuration model have been proven to be effective in generating networks with k-core structure akin to that of empirical networks[22,23]. In a nutshell, in these models the edge stubs attached to each node are divided into two groups: red and blue. Red stubs can create any edges regardless of the k-shell structure. Blue stubs only form edges connecting nodes belonging to distinct k-shells. Among the possible pairs of stubs' colours, only the blue-blue pair is forbidden.

As mentioned above, another type of mesoscale structure is communities. Although there is not a univocal definition of what a community is, in general the community refers to a group of nodes that are more tightly connected between each other than with the other nodes of the network[6]. Communities are also defined by the concept of stochastic equivalence, i.e., nodes in the same group/community interact, on average, with nodes in other groups in the same way[24]. Methods based on different definitions of communities may return different partitions of the node set. However, there is often some consistency between those partitions, which indicates the presence of groups of nodes acting like the building blocks of communities[25]. The presence of communities is an important large-scale characteristic of many empirical networks because a system's different functions tend to be located in different communities (e.g., in functional brain networks[26] and protein-protein interaction networks[27]). Moreover, it has been proven that communities play a role in the resilience of the system[28] and the presence of triangles[29], as well as in the emergence of collective behaviour including synchronisation[30], the emergence of cooperation[31,32], spreading of a pandemic[33], and the attainment of consensus[34,35].

Although k-core and communities are two ways of decomposing the same network, there may be overlaps or intricate relationships between them. In the present paper, we study the relation between the k-core decomposition and the community structures of several empirical and synthetic networks. In particular, we leverage the work of Alvarez-Hamelin et al.[36] and confirm that the nodes' degrees (i.e., their number of edges) alone are not capable of reproducing the network's k-shell structure. We find that one has to include information about the community structure to obtain networks whose k-core decomposition looks sufficiently close to the empirical one. We also highlight the existence of a concentration-like phenomenon of the innermost k-shells into a small number of communities, which is stronger in some data sets than others.

Results

Degree-based reconstruction of the *k*-core.

As stated above, various studies on networks leverage the k-core decomposition to extract insightful information from networks. However, less studies have asked which mechanisms are sufficient for explaining generation of networks having empirically observed patterns of k-core decomposition. More specifically, Alvarez-Hamelin et al. found that networks generated using the configuration model[37] having a Poisson or power-law degree distribution do not display a k-core structure similar to the one displayed by the AS network[36]. Using the results of Alvarez-Hamelin et al. as a starting point, given an empirical network G with N nodes, we check whether its k-core decomposition can be reproduced solely from the degree of each node i (i.e., the number of edges that node i has), denoted by k_i. We generated random networks by a standard configuration model preserving the degree of each node of G, which we denote by `deg` (see "Methods" section for details).

We have analysed several empirical networks encompassing social, technological, linguistic, and transportation systems whose main properties are summarised in Table 1. In Fig. 1, we show the survival function of the probability distributions of the k-shell index, $P_{\geq}(k_{\mathrm{s}})$ (i.e., fraction of nodes whose k-shell index is larger than or equal to k_{s}), for a selection of data sets, compared across the original networks and their synthetic counterparts (see Supplementary Fig. S1 in SM for the other data sets). Figure 1 indicates that the degree of each node is not sufficient for reproducing the k-core structure of the original networks because $P_{\geq}(k_{\mathrm{s}})$ for `deg` considerably deviates from that for the original networks. This result is consistent with the previous results[36]. In fact, we find that fixing the degree of each node is sufficient to recover the k-core profile in some networks. For these networks the empirical and `deg` networks are not too different in terms of $P_{\geq}(k_{\mathrm{s}})$ (e.g., Facebook 2 and Cookpad networks). We point out two main differences in $P_{\geq}(k_{\mathrm{s}})$ between the empirical and `deg` networks. First, for most data sets, the largest k_{s} value, which is denoted by D and called the degeneracy, is considerably smaller for the networks generated by `deg` than the original networks. Second, the $P_{\geq}(k_{\mathrm{s}})$ of some empirical networks have plateaus and abrupt drops in $k_{\mathrm{s}} \leq D$. The plateaus imply that some of the k-shells are completely or almost empty, whereas the abrupt drops indicate that some k-shells are more densely populated than those adjacent to them. In contrast, $P_{\geq}(k_{\mathrm{s}})$ for the `deg` networks does not have a notable plateau or drop in $k_{\mathrm{s}} \leq D$. Therefore, in the `deg` networks, all the k-shells up to $k_{\mathrm{s}} = D$ are populated, and there is no k-shell that is substantially more populated than its adjacent k-shells.

A more quantitative comparison of distribution $P_{\geq}(k_{\mathrm{s}})$ between the empirical and `deg` networks may be done by, for example, the Kolmogorov-Smirnov (KS) test[38]. However, because a majority of the nodes usually belongs to outer k-shells, (i.e., set of nodes with small k_{s} values) and Fig. 1 shows that the strongest discrepancies between the two distributions tend to occur at large k_{s} values, the KS test fails to grasp the differences at large k_{s} values that we are mostly interested in. Therefore, we compare the k-core decomposition of the empirical and `deg` networks using four indicators, i.e., the relative difference in the average k-shell index, $\Delta\langle k_{\mathrm{s}}\rangle$, the relative difference in the network's degeneracy, ΔD, the Jaccard score, J, and Kendall's, τ_K of the nodes belonging to the top 10% (i.e., innermost k-shells) of the $P_{\geq}(k_{\mathrm{s}})$ distribution. The average of each indicator over all the data sets for the networks obtained with the `deg` shuffling method is equal to $\langle\Delta\langle k_{\mathrm{s}}\rangle\rangle = 0.052 \pm 0.056$, $\langle\Delta D\rangle = 0.302 \pm 0.288$, $\langle J\rangle = 0.563 \pm 0.194$, and $\langle\tau_K\rangle = 0.763 \pm 0.176$. The value of $\langle\Delta\langle k_{\mathrm{s}}\rangle\rangle$ indicates that $\langle k_{\mathrm{s}}\rangle$ is only $\approx 5\%$ different

Data set	N	L	$\langle k \rangle$	$k_{\max}$	$\langle k_s \rangle$	D	N_c^{Lvn}	Q^{Lvn}	N_c^{SBM}	Q^{SBM}	References
Facebook 1	4039	88234	43.691	1045	26.880	115	16	0.835	62	0.551	[55,74]
Facebook 2	6386	217662	68.168	930	35.712	56	19	0.419	198	0.158	[56–58,75]
Facebook 3	2235	90954	81.391	467	44.508	63	8	0.436	87	0.139	[56–58,76]
Facebook 4	11247	351358	62.480	415	32.413	63	10	0.438	274	0.193	[56–58,77]
Facebook 5	27737	1034802	74.615	2555	38.681	81	18	0.470	547	0.172	[56–58,78]
Twitter	81306	1342296	33.018	3383	17.762	96	73	0.808	510	0.511	[55,79]
Web-blogs	1490	16715	22.436	351	12.154	36	275	0.426	17	0.076	[60,80]
Emails	1005	16064	31.968	345	17.063	34	26	0.410	33	0.232	[81–83]
Cond. Matter	23133	93439	8.078	279	4.900	25	619	0.730	203	0.633	[83,84]
Comp. Science	317080	1049866	6.622	343	4.215	113	209	0.822	676	0.726	[59,85,86]
Global airline	3376	19179	11.362	248	6.123	31	26	0.665	40	0.311	[61]
Words	146005	656999	9.000	1008	5.289	31	378	0.759	548	0.583	[59,87,88]
Cookpad Greece	32235	745178	46.234	8196	23.709	158	40	0.166	76	0.020	–
Cookpad Spain	122158	1749751	28.647	12637	14.547	162	262	0.270	90	0.035	–
Cookpad UK	13758	47525	6.909	1880	3.558	33	199	0.350	8	0.114	–

Table 1. Main properties of the data sets used in the present study. N: number of nodes, L: number of edges, $\langle k \rangle$: average degree, $k_{\max}$: maximum degree, $\langle ks \rangle$: average value of the k-shell index, D: maximum value of the k -shell index, N_c^{Lvn}, Q^{Lvn}: number of communities determined by the Louvain method and the corresponding modularity, respectively, N_c^{SBM}, Q^{SBM}: number of communities determined by the SBM and the corresponding modularity, respectively.

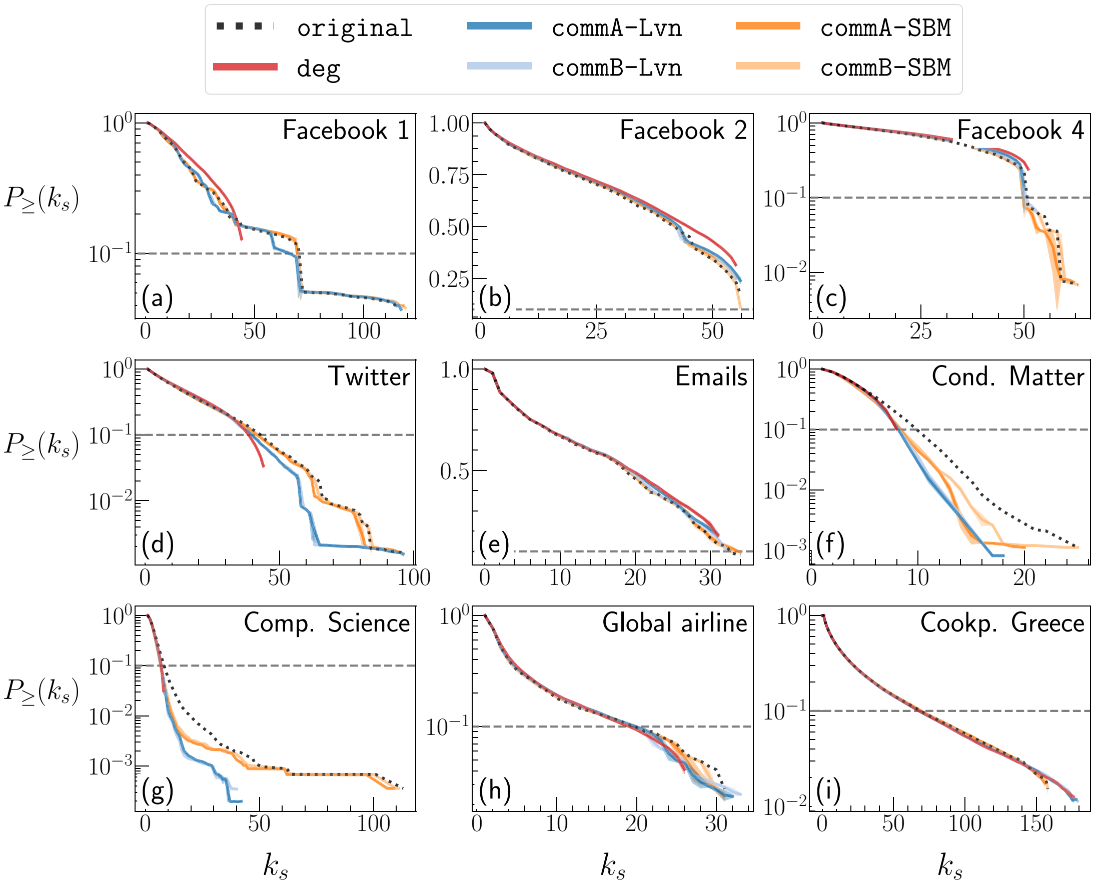

Figure 1. Survival function of the probability distributions of the k-shell index, i.e., $P_{\geq}(k_s)$ as a function of k_s for the original network (dotted line) and shuffled networks (solid line). Each panel corresponds to a data set, i.e., Facebook 1 (panel **a**), Facebook 2 (**b**), Facebook 4 (**c**), Twitter (**d**), Emails (**e**), Cond. Matter (**f**), Comp. Science (**g**), Global airline (**h**), and Cookpad Greece (**i**). The horizontal dashed lines indicate that $P_{\geq}(k_s) = 0.1$. Results are averaged over 10 different runs of each shuffling method, and the shaded areas (when visible) represent the standard deviations.

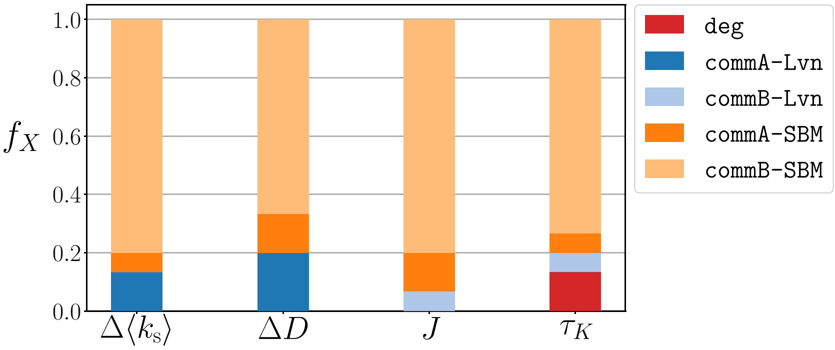

Figure 2. Performances of different shuffling methods in terms of four indicators. We report the fraction of data sets for which a given combination of the shuffling method and the community detection method yields an indicator's value closest to that for the original network. Each bar refers to an indicator, i.e., average k-shell's difference, $\Delta\langle k_s\rangle$, degeneracy's difference, ΔD, Jaccard score, J, and Kendall's tau, τ_K.

between the original and deg networks on average. However, their degeneracy differs by $\approx$ 30% on average. The $\langle J\rangle$ and $\langle\tau_K\rangle$ values inform us that innermost k-shells of the original networks and those of the deg networks tend to share approximately half of the nodes, albeit their ranking seems to be fairly preserved. Supplementary Table S1 reports the values of each indicator.

Community-aware reconstruction of the *k*-core.

We have seen that the degree distribution by itself does not reproduce main features of the k-shell index distribution. An alternative feature that may explain the k-shell index distribution is the community structure. For this reason, we generated synthetic networks that preserve both the degree of each node and the community structure, $\mathcal{C} = \{C_1, \ldots, C_{N_c}\}$, where N_c is the number of communities of the original network. To account for the multiple definitions of what a community is, we identified the communities of each network using two methods: the Louvain method[39], denoted by Lvn, and the degree-corrected stochastic block model[40], denoted by SBM. In combination with each of the two community detection methods, we considered two rewiring methods preserving $\mathcal{C}$ and the degree of each node, denoted by commA and commB. Method commA preserves the exact number of inter- and intra-community edges at the level of single communities. Method commB preserves the number of inter- and intra-community edges for each node.

Figure 1 indicates that preserving the community structure in addition to the degree of each node improves the similarity in $P_{\geq}(k_s)$ between the empirical and synthetic networks, especially at large k_s values, which correspond to inner k-shells. In particular, commA and commB generate networks whose D value tends to be closer to the empirical value than deg does. Furthermore, $P_{\geq}(k_s)$ for commA and commB tends to have plateaus and abrupt drops at $k_s \leq D$ similarly to the empirical networks. Overall, synthetic networks preserving the SBM community structure have a k-core decomposition more akin to the empirical one than those preserving the Lvn community structure. This observation is quantitatively supported by the values of the four indices reported in Supplementary Table S1.

To obtain an overview of the performances of different network randomisation methods, in Fig. 2 we show the fraction of data sets, f_X, for which a certain shuffling method generates a k-core decomposition that is the most similar to that of the empirical network according to each indicator. The figure indicates that commB-SBM (i.e., the commB shuffling method that preserves the community structure determined by SBM) performs the best in mimicking the k-shell index features for approximately 65–80% of the data sets, depending on the indicator. Detailed results for the performance of each method for each empirical network are shown in Supplementary Fig. S2 and Supplementary Table S1.

One issue of Lvn is that it cannot discover small communities[6,41]. One way to mitigate this limitation is to introduce in Lvn a resolution parameter, $r \in (0, 1]$, regulating the resolution scale. It is possible to detect small communities when r is small, whereas the original Lvn corresponds to $r = 1$[42]. We denote by LvnR the Louvain method with $r < 1$, i.e., with a resolution higher than that used by Lvn. In Sect. 2 of SM we report whether preserving the communities found using LvnR instead of Lvn improves our ability to reproduce the k-core decomposition of the original network. We found that LvnR performs better than Lvn (Supplementary Figs. S3 and S4) but worse than SBM in general (Supplementary Fig. S5).

Imposing the simultaneous conservation of each node's degree and community structure may result in synthetic networks that are not substantially different from the original ones. To exclude this possibility, we computed the Jaccard score, $J(\mathcal{L}, \mathcal{L}')$, (see Eq. (3)) for the sets of edges, $\mathcal{L}$ and $\mathcal{L}'$, of the original and shuffled networks, respectively. The values of J approximately fall between 0.01 and 0.5, confirming that the set of edges – hence, the networks – are considerably different.

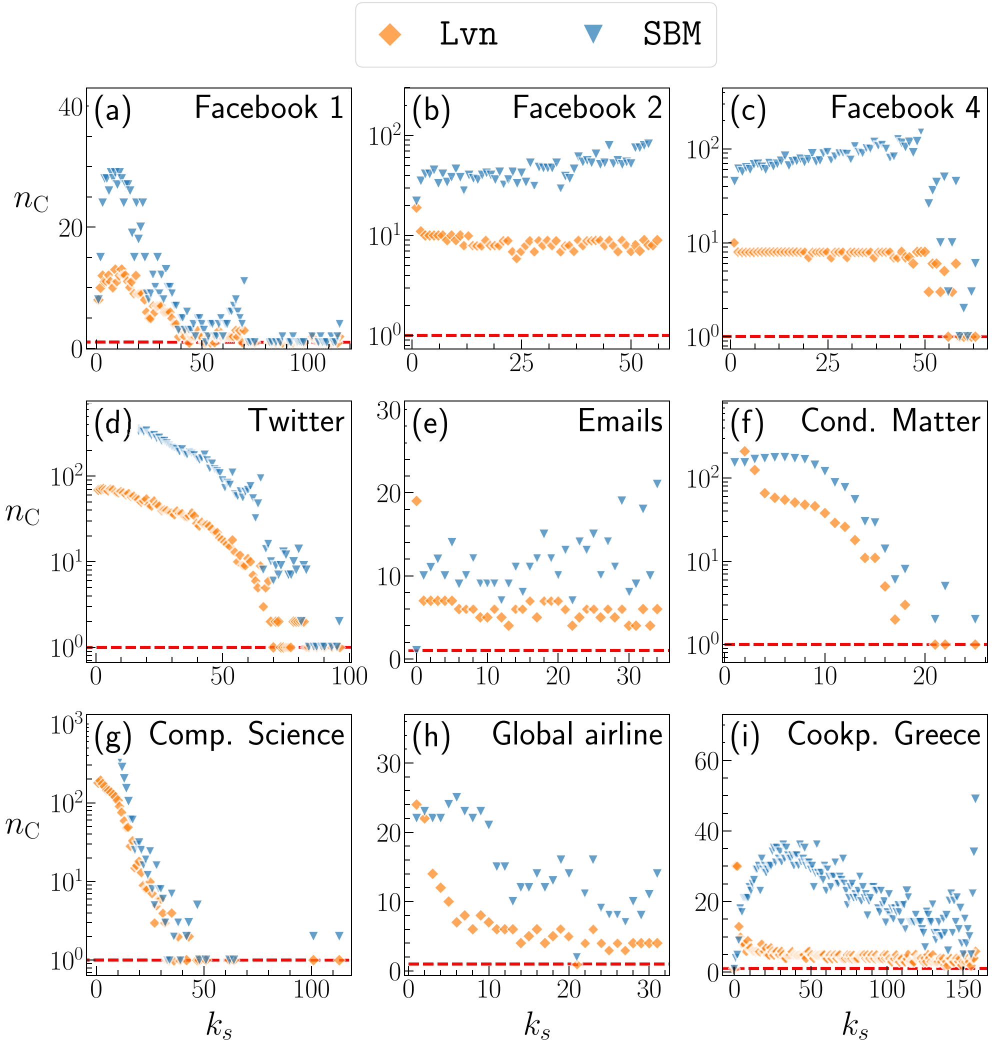

Figure 3. Number of different communities, $n_C(k_s)$, that the set of nodes of a given k-shell value, k_s, overlaps. The horizontal dashed line is a guide to the eyes showing $n_C(k_s) = 1$. Each panel accounts for a different data set (see the caption of Fig. 1 for the details). For each data set, we show the results corresponding to the community structure obtained using either `Lvn` or `SBM`.

The results presented so far suggest that preserving the community structure improves the preservation of the k-core decomposition of the original network. Therefore, the mere presence of a community structure may be enough to preserve the main features of the k-core decomposition of the original networks. To test this possibility, we applied the k-core decomposition to networks with communities generated using the LFR model[43] (see Sect. 3 of SM). The plots of $P_{\geq}(k_s)$ shown in Supplementary Figs. S7–S10 indicate that the presence of a community structure alone is not sufficient for producing major features of the k-core structure of the empirical networks. Specifically, the $P_{\geq}(k_s)$ of the networks generated by the LFR model is always smooth and shows neither plateaus nor abrupt drops as k_s increases. Moreover, with the LFR, k_s is narrowly distributed, i.e., $\max(k_s) - \min(k_s) \approx 10$. These differences between the k-core structure of the LFR model and that of empirical networks are not sensitive to the value of the mixing parameter, μ, of the LFR model, which controls how distinct the communities are. It should also be noted that for the LFR model, as for the empirical network, the `commB-SBM` generates networks that are the most similar to the original LFR networks among the different shuffling methods in terms of $P_{\geq}(k_s)$.

Overlap between communities and *k*-core. Preserving the community structure in addition to the node's degree can lead to preservation of features of the k-core structure possibly because nodes with high values of k_s form a k-core which tend to belong to the same community. To examine this possibility, we show the number of communities to which the nodes of a given k-shell belong, $n_C(k_s)$, in Fig. 3 (see Supplementary Fig. S11 for the other data sets). Although each data set shows a distinct pattern, for many data sets, inner k-shells (i.e., nodes with large k_s values) are concentrated into one or a few communities. The concentration effect is particularly noticeable for some data sets, e.g., Facebook 1 and Twitter. To check whether the number of communities per k-shell is merely a byproduct of the random combinatorial effect owing to the number of communities, the distribution of the community size, and the distribution of k_s, we computed a random assignments of the nodes to communities and then calculated $n_C(k_s)$ for each k_s value (see Sect. 4 and Supplementary Fig. S12 of the SM). We have found that the nodes in each k-shell are almost always more concentrated into a smaller number of communities than what is expected by the random assignment of the nodes to communities for all the data sets and community detection methods, with the only exception of `SBM` for Cookpad's data sets. This finding is in agreement with the previous result that nodes with high k_s tend to belong to the same community, which has been observed in networks embedded into hyperbolic spaces[44,45]. In particular, we observe a strong concentration of the k-shells into a few communities for the Facebook 1, Twitter, Cond. Matter, Comp. Science, and Words networks, which are those showing a more pronounced difference in the values of D between the original and `deg` networks.

Discussion

The information encoded in the degree of each node is not sufficient for generating networks with a k-core structure that is similar to those of empirical networks[36]. This gap of knowledge calls for the design of generative models of networks beyond the configuration model. Such models are expected to be useful to generate benchmark networks and to understand the mechanisms behind the emergence of the k-core. To the best of our knowledge, few models are available to generate networks with a given k-core decomposition[21–23].

In the present study, we investigated how much the combination of the nodes' degrees and community structure accounts for k-core structure of empirical networks. Given a network G, we randomly shuffled G's edges to generate its synthetic counterparts preserving each node's degree and/or community structure of G. We found that randomised networks preserving the community structure obtained through a stochastic block model showed a k-shell index distribution that was reasonably similar to the distribution for the original networks. The success of the stochastic block model in mimicking the features of the k-core decomposition might be due to its ability to approximate the mesoscale structures of networks with a good accuracy[24,46], including communities. We also sought to understand more the relationship between k-core and communities by studying networks generated by the LFR model which enables us to control the extent to which the communities are distinguished from each other. However, regardless of whether or not different communities are relatively distinguished from each other in a network, the k-shell index distribution of LFR networks does not show the same features as those observed in the empirical networks. Finally, we have investigated the overlap between communities and k-shells and found that, in some empirical networks, the nodes in inner k-shells are concentrated into a small number of communities much more so than a randomised counterpart. This result is in agreement with the observations made for networks embedded in hyperbolic spaces[44,45]. Up to our numerical efforts, the concentration is observed if and only if the empirical network and its `deg` counterpart are substantially different in terms of their k-core decomposition. The concentration suggests that inner k-shells may perform specific functions in such networks, corresponding to the functions of the communities they belong to as observed in, for instance, functional brain networks[26] and protein-protein interaction networks[27].

The "community aware" rewiring mechanisms introduced in this paper can be used for assessing whether or not a given property of a network is a direct expression of its community structure. One example of such an approach is given in[47], where the authors have improved the robustness against attacks on a network while keeping its community structure. In that case, the method only preserves the communities and alters the connectivity pattern by increasing the density of intra-community edges as well as changing the edges between communities. It may be interesting, instead, to check whether the robustness of the network can be improved even when one also preserves the degree of the nodes using our community-aware rewiring mechanisms.

One viable extension of our work is to the case of k-peak graph decomposition method[48]. In Ref.[48], the authors argue that for networks with communities, the k-core decomposition should be performed locally rather than globally, thus returning the k-peak decomposition of each of the system's regions. The rationale behind this approach is to avoid that, if the network contains regions with different densities of edges, the standard k-core decomposition would fail to recognise local core nodes in sparser regions. Studying the evolution of the k-peak decomposition in response to the rewiring of the connections may unveil salient features of complex systems. Another possible direction of research is to concatenate the information encoded into the k-shell index, k_s, with the one provided by the so-called onion decomposition (OD)[49]. The OD is an extension of the k-core decomposition where a node is labelled with both its k_s and its layer index. The layer of a node i represents the iteration number with which node i is removed in the recursive pruning process of the k-core decomposition. The OD provides a further characterisation of the structure of the network than the k-core, revealing, for example, how tree-like the network is.

Summing up, in this work we have analysed the interplay between the k-core decomposition and community structure of networks. Understanding such a relationship is useful not only owing to the broad range of applications of k-core decomposition, but also to inform the design of models capable of generating networks with both a community structure and k-core's features beyond those explainable by the degree distribution. Such models may stand on, for instance, the stochastic block model[40], the enhanced configuration model based on maximum entropy[50], or the hierarchical extension of the LFR model[51]. Alternatively, models based on microscopic growth mechanisms such as triadic closure[52,53] or modified preferential attachment[54] may deserve further investigation.

Methods

Data. We have considered networks corresponding to systems of different types: from social to technological, from semantic to transportation. Table 1 summarises main properties of such networks. Except for Cookpad networks, all the data sets are publicly available and have been retrieved from the Stanford Large Network data set Collection[55] (Facebook 1, Twitter, Emails, and Cond. Matter), the Network Repository[56–58] (Facebook 2, 3, 4, and 5), the Koblenz Network Collection (KONECT)[59] (Comp. Science, and Words), Mark E. J. Newman's personal network data repository[60] (Web-blogs), and the OpenFlights data repository[61] (Global airline). In the following text, we provide a brief description of each data set.

Facebook and Twitter These networks describe social relationships. Nodes are people. Edges represent their friendship relations.

Web-blogs This network is composed of the hyperlinks (edges) between weblogs on US politics (nodes) recorded in 2005.

Emails This is a network of email data from a large European research institution. Nodes are people. Edges connect pairs of individuals who have exchanged at least one e-mail.

Cond. Matter and Comp. Science The former network is the co-authorship network of the authors of preprint manuscripts submitted to the Condensed Matter Physics `arXiv` e-print archive from January 1993 to April 2003. The latter network is similarly defined using manuscripts appearing in the DBLP computer science bibliography, using a comprehensive list of research papers in computer science. The submission time of the papers of the DBLP collection is unavailable. A node is an author. An edge represents the existence of at least one manuscript co-authored by two authors.

Global airline In this network nodes are airports across the globe. An edge indicates direct commercial flights between two airports.

Words This network accounts for the lexical relationships among words extracted from the WordNet data set. Nodes are English words. Edges are relationships (synonymy, antonymy, meronymy, etc.) between pairs of words.

Cookpad These networks are extracted from the Cookpad online recipe sharing platform[62]. Users can post and browse recipes, as well as interact with other users through recipes in multiple ways including liking, sharing, and posting a comment. The platform is present in many countries (e.g., Japan, Indonesia, United Kingdom, and Italy). Here, we consider the data collected from September to November of 2018 in Greece, Spain, and the United Kingdom, separately for each country. In the three networks, nodes are users. An edge between a pair of users exists if one or more of the following types of events takes place: like or follow a user, viewing, bookmarking, commenting, or making a cooksnap of another user's recipe.

All the networks considered in this work are treated as undirected and unweighted, even when the original data contains more information. Finally, we also consider synthetic networks, generated using the LFR (Lancichinetti–Fortunato–Radicchi) model[43] (see Sect. 3 of SM for details).

Network shuffling. Given a network, G, with N nodes and L edges, we generate a randomised counterpart, G', that has the same nodes and the same number of edges by shuffling the edges of G. We consider three shuffling methods denoted by `deg`, `commA`, and `commB`; each shuffling method preserves different properties of G. The shuffling consists in selecting uniformly at random two edges (a, b) and (c, d), and replacing them with, e.g., (a, c) and (b, d), if the swapping of the edges is accepted. An attempt to swap edges is accepted, in which case we call the swapping effective, if and only if it respects the rule of the specific shuffling method and the swapping does not generate self-loops or multiple edges. We continued the shuffling until we carried out $2L$ effective swaps, such that an edge was swapped four times on average.

In the following text, we provide the details of each shuffling method. Assume that network G partitions into communities such that the set of the communities is $\mathcal{C} = \{C_1, \ldots, C_{N_c}\}$, where N_c is the number of communities. Furthermore, let $g(i) \in \mathcal{C},\ i = 1, \ldots, N$, be the community to which the ith node belongs and k_i be the degree of node i. We have:

Degree-preserving shuffling (`deg`) This method preserves degree k_i of each node i and is equivalent to the configuration model[37].

Community-preserving shuffling of type A (`commA`) On top of the degree of each node, this method preserves the total number of edges within each community and between each pair of communities. In attempts to swap edges, we replace two randomly selected edges (a, b) and (c, d) by (a, c) and (b, d) if and only if an end node of edge (a, b) and an end node of edge (c, d) belong to the same community (i.e., if $g(b) = g(c)$ or $g(a) = g(d)$).

Community-preserving shuffling of type B (`commB`) Like `commA`, this method preserves the degree of each node and the number of edges within each community and between each pair of communities. In contrast with `commA`, the `commB` method preserves the numbers of edges within and across communities for each node, and not only for each community or pairs of communities. Given two selected edges (a, b) and (c, d), we replace them with (a, c) and (b, d) if and only if the two new edges connect the same community pairs as before the swapping (i.e., $g(b) = g(c)$ and $g(a) = g(d)$).

Comparison of the *k*-core decomposition. To assess the similarity between the k-core decomposition of the original network, G, and of its shuffled counterpart, G', we used four indicators: the average k-shell index, $\langle k_s \rangle$, the network's degeneracy, D, the Jaccard score, J, and the generalised Kendall's tau, τ_K. The indicator $\langle k_s \rangle$ explicitly depends on all the nodes in the network, whereas D, J and τ_K only depend on the nodes belonging to the innermost k-shell(s). We use the latter three indicators because, although a majority of nodes tends to belong to outer k-shells, it is a difference in the tails of the k_s distributions that often affect functions of networks such as the impact of influencers in contagion processes[63]. The four indicators are defined as follows.The average of the k-shell index, $\langle k_s \rangle$, is equal to

$$\langle k_s \rangle = \frac{1}{N} \sum_{i=1}^{N} k_s(i) \,, \tag{1}$$

where $k_s(i)$ is the k-shell index of node i. The degeneracy, D, of a network G is given by[64]

$$D = \max_{i \in G} \{ k_s(i) \} \,. \tag{2}$$

Rather than using these raw indicators, to compare across the different data sets, we compute their relative difference between the empirical network and its shuffled counterpart given by $\Delta X = |X_G - X_{G'}|/X_G$, where $X \in \{\langle k_s \rangle, D\}$.

To compute J and τ_K, we need to define a criterion to select nodes belonging to the innermost k-shells. We decided to confine the comparison to the nodes whose k_s falls within the top 10% among the N nodes. The horizontal lines in Fig. 1 indicate the threshold values of $k_s^\star$ such that $P_\geq(k_s^\star) = 0.1$. In the same manner, we define $k_s^{\star\prime}$ such that $P_\geq(k_s^{\star\prime}) = 0.1$ in network G'. To calculate J and τ_K, we use the nodes belonging to k-shells with $k_s \geq k_s^\star$ in G and the nodes belonging to k-shells with $k_s \geq k_s^{\star\prime}$ in G' without duplication of the nodes. There are several remarks. First, it may hold that $k_s^\star \neq k_s^{\star\prime}$. Second, the value of $k_s^{\star\prime}$ varies from one combination of a run of shuffling and community detection to another. Third, as in the case of the Facebook 2 data set, $k_s^{\star\prime}$ sometimes does not even exist. In such a case, we set $k_s^{\star\prime} = D$ and select all the nodes belonging to the innermost k-shell although they constitute more than 10% of the nodes in the network. Fourth, additional tests using different threshold percentages, 5% and 20%, instead of 10%, did not qualitatively change the results. Fifth, while the Jaccard score simply compares the nodes belonging to two sets, the generalised Kendall's tau, τ_K compares ranked sets. In our case, the node's rank is equivalent to the k_s value.

Given two sets $\mathcal{A}$ and $\mathcal{B}$, the Jaccard score quantifies their overlap and is given by

$$J(\mathcal{A}, \mathcal{B}) = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|} \,. \tag{3}$$

The Jaccard score ranges between 0 and 1. A value of 1 indicates the complete overlap between the two sets (i.e., the sets are the same), whereas a value of 0 indicates that the sets are completely different.

The generalised Kendall's tau, τ_K, measures the consistency between two rankings by assigning penalties to pairs of elements on which the two rankings disagree[65,66]. Given two sets $\mathcal{A}$ and $\mathcal{B}$ having m_A and m_B elements, respectively, consider their associated ranking functions $\mathcal{X}$ and $\mathcal{Y}$. We denote with (z_1, z_2) an arbitrary pair of elements of $\mathcal{A} \cup \mathcal{B}$. We assign a penalty $K_{z_1,z_2}(\mathcal{X}, \mathcal{Y}) = 1$ to (z_1, z_2) if (a) the rankings of the two elements within each set are different (i.e., $\mathcal{X}(z_1) \gtreqless \mathcal{X}(z_2)$ and $\mathcal{Y}(z_1) \lesseqgtr \mathcal{Y}(z_2)$), (b) the element with the higher rank in one set is missing in the other set, i.e., $\mathcal{X}(z_1) > \mathcal{X}(z_2)$ and $z_1 \notin \mathcal{B}$ (or $\mathcal{X}(z_2) > \mathcal{X}(z_1)$ and $z_2 \notin \mathcal{B}$), or (c) both elements belong to one set each, which is not the same set, i.e., $z_1 \notin \mathcal{B}$ and $z_2 \notin \mathcal{A}$ (and vice-versa). In all the other cases $K_{z_1,z_2}(\mathcal{X}, \mathcal{Y}) = 0$, such that we do not penalise the (z_1, z_2) pair. Finally, we sum the penalties over all the possible pairs of elements and normalise it, thus obtaining the generalised Kendall's tau:

$$\tau_K(\mathcal{X}, \mathcal{Y}) = 1 - \frac{1}{m_A m_B} \sum_{z_1, z_2 \in \mathcal{A} \cup \mathcal{B}} K_{z_1,z_2}(\mathcal{X}, \mathcal{Y}). \tag{4}$$

Index τ_K ranges between 0 and 1. If $\tau_K = 1$, the two rankings are completely coherent. If $\tau_K = 0$, the two sets $\mathcal{A}$ and $\mathcal{B}$ have no pair of elements on which rankings $\mathcal{X}$ and $\mathcal{Y}$ are coherent. The above formulation of the Kendall's tau is the so-called optimistic approach[65]. This means that we do not penalise the case in which a pairs of elements is present in one set and not in the other set.

Community detection methods. We considered two methods for community detection. The first is the Louvain method (`Lvn`)[39], which is a heuristic greedy multiscale method that approximately maximises the modularity function. Given a network with N nodes distributed among N_c communities, the modularity, Q, reads

$$Q = \frac{1}{2L} \sum_{i,j=1}^{N} \left[a_{i,j} - \frac{k_i k_j}{2L} \right] \delta\big(g(i), g(j)\big) \,, \tag{5}$$

where $a_{i,j}$ is the element of the network's adjacency matrix A; $g(i)$ is the community to which the i-th node belongs ($1 \leq g(i) \leq N_c$), and $\delta\big(g(i), g(j)\big)$ is the Kronecker delta. A large value of Q implies a good partitioning. The Louvain method seeks the partitioning that maximises the modularity. Note that we obtain $Q \approx 0$ for random assignment of nodes to communities and that we obtain $Q \approx 1$ when the network is made of perfectly disjoint communities.

The other community detection method that we used is the stochastic block model[67]. It uses the probabilities $\mathcal{P} = \{p_{C_i,C_j}\}$ with which there exists an edge (a, b) connecting an arbitrarily selected node a in community C_i (i.e., $g(a) = C_i$) and an arbitrarily selected node b in community C_j (i.e., $g(b) = C_j$). Different instances of probabilities $\mathcal{P}$ allow the description of different mixing patterns. When the diagonal entries of $\mathcal{P}$ predominate, we obtain the most usual community structure, whereas other instances yield other structures such as bipartite or core-periphery structure.

To find the optimal partition, one maximises the likelihood function with respect to $\{p_{C_i,C_j}\}$ corresponding to the partitioning $\mathcal{C} = \{C_i\}$, where $i, j \in 1, \ldots, N_c$. The unnormalised log-likelihood, $\mathfrak{L}$, with which a partition of network G into N_c communities, $\mathcal{C}$, is reproduced reads

$$\mathfrak{L}(G \,|\, \mathcal{C}) = \sum_{i,j=1}^{N_c} e_{ij} \log\left(\frac{e_{ij}}{m_i\, m_j}\right), \tag{6}$$

where e_{ij} is the number of edges connecting community C_i and community C_j, and m_i is the number of nodes belonging to C_i.

The above formulation, however, has one major limitation: it assumes that the degrees of the nodes are distributed according to a Poisson-like function. To account for the degrees' heterogeneity, Karrer et al. have implemented the so-called degree corrected stochastic block model, in which the expected degree of each node is kept constant via the introduction of additional parameters[40]. Let e_i be the sum of the node's degree over all nodes in community C_i. Then, the unnormalised log-likelihood for the degree-corrected stochastic block model reads

$$\mathfrak{L}_{\mathrm{DC}}(G \,|\, \mathcal{C}) = \sum_{i,j=1}^{N_c} e_{ij} \log\left(\frac{e_{ij}}{e_i\, e_j}\right). \tag{7}$$

Equations (6) and (7) depend on the number of communities N_c. Because the value of N_c is not known a priori, it is inferred through the minimisation of a quantity called the description length. The minimum description length principle describes how much a model compresses the data and allows us to find the optimal number of communities while avoiding overfitting[68]. In the present work we use the degree-corrected stochastic block model and its implementation available in the Python Graph-tool package[69], which we refer to as `SBM` for brevity.

Data availability

The data sets on Cookpad™analysed in the current study are not publicly available due to exclusive ownership of Cookpad Limited. All the other data sets are available from the corresponding repositories listed in the bibliography.

Received: 2 May 2020; Accepted: 7 August 2020
Published online: 07 September 2020

References

1. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D. Complex networks: Structure and dynamics. *Phys. Rep.* **424**, 175–308. https://doi.org/10.1016/j.physrep.2005.10.009 (2006).
2. Barabási, A.-L. The network takeover. *Nat. Phys.* **8**, 14–16. https://doi.org/10.1038/nphys2188 (2011).
3. Vespignani, A. Modelling dynamical processes in complex socio-technical systems. *Nat. Phys.* **8**, 32–39. https://doi.org/10.1038/nphys2160 (2012).
4. Lambiotte, R., Rosvall, M. & Scholtes, I. From networks to optimal higher-order models of complex systems. *Nat. Phys.* **15**, 313–320. https://doi.org/10.1038/s41567-019-0459-y (2019).
5. Benson, A. R., Gleich, D. F. & Leskovec, J. Higher-order organization of complex networks. *Science* **353**, 163–166. https://doi.org/10.1126/science.aad9029 (2016).
6. Fortunato, S. & Hric, D. Community detection in networks: a user guide. *Phys. Rep.* **659**, 1–44. https://doi.org/10.1016/j.physrep.2016.09.002 (2016).
7. Gfeller, D. & De Los Rios, P. Spectral coarse graining of complex networks. *Phys. Rev. Lett.* **99**, 038701. https://doi.org/10.1103/PhysRevLett.99.038701 (2007).
8. Borgatti, S. P. & Everett, M. G. Models of core/periphery structures. *Soc. Netw.* **21**, 375–395. https://doi.org/10.1016/S0378-8733(99)00019-2 (2000).
9. Csermely, P., London, A., Wu, L.-Y. & Uzzi, B. Structure and dynamics of core/periphery networks. *J. Complex Netw.* **1**, 93–123. https://doi.org/10.1093/comnet/cnt016 (2013).
10. Rombach, P., Porter, M. A., Fowler, J. H. & Mucha, P. J. Core-periphery structure in networks (revisited). *SIAM Rev.* **59**, 619–646. https://doi.org/10.1137/17M1130046 (2017).
11. Mariani, M. S., Ren, Z.-M., Bascompte, J. & Tessone, C. J. Nestedness in complex networks: observation, emergence, and implications. *Phys. Rep.* **813**, 1–90. https://doi.org/10.1016/j.physrep.2019.04.001 (2019).
12. Zhou, S. & Mondragon, R. J. The rich-club phenomenon in the Internet topology. *IEEE Commun. Lett.* **8**, 180–182. https://doi.org/10.1109/LCOMM.2004.823426 (2004).
13. Colizza, V., Flammini, A., Serrano, M. Á & Vespignani, A. Detecting rich-club ordering in complex networks. *Nat. Phys.* **2**, 110–115. https://doi.org/10.1038/nphys209 (2006).
14. Erdős, P. & Hajnal, A. On chromatic number of graphs and set-systems. *Acta Math. Acad. Sci. Hung.* **17**, 61–99. https://doi.org/10.1007/BF02020444 (1966).
15. Seidman, S. B. Network structure and minimum degree. *Soc. Netw.* **5**, 269–287. https://doi.org/10.1016/0378-8733(83)90028-X (1983).
16. Kong, Y.-X., Shi, G.-Y., Wu, R.-J. & Zhang, Y.-C. k-core: theories and applications. *Phys. Rep.* **832**, 1–32. https://doi.org/10.1016/j.physrep.2019.10.004 (2019).
17. Malliaros, F. D., Giatsidis, C., Papadopoulos, A. N. & Vazirgiannis, M. The core decomposition of networks: theory, algorithms and applications. *VLDB J.* **29**, 61–92. https://doi.org/10.1007/s00778-019-00587-4 (2020).

18. Medina, A., Lakhina, A., Matta, I. & Byers, J. Brite: an approach to universal topology generation. In *MASCOTS Proceedings Ninth International Symposium on Modeling. Anal. Simul. Comput. Telecommun. Syst.* **346–353**, 2001. https://doi.org/10.1109/MASCOT.2001.948886 (2001).
19. de Solla Price, D. Networks of scientific papers. *Science* **149**, 510–515. https://doi.org/10.1126/science.149.3683.510 (1965).
20. Barabàsi, A. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512. https://doi.org/10.1126/science.286.5439.509 (1999).
21. Baur, M., Gaertler, M., Görke, R., Krug, M. & Wagner, D. Augmenting *k*-core generation with preferential attachment. *Netw. Heterogeneous Media* **3**, 277–294. https://doi.org/10.3934/nhm.2008.3.277 (2008).
22. Hébert-Dufresne, L., Allard, A., Young, J.-G. & Dubé, L. J. Percolation on random networks with arbitrary *k*-core structure. *Phys. Rev. E* **88**, 062820. https://doi.org/10.1103/PhysRevE.88.062820 (2013).
23. Allard, A. & Hébert-Dufresne, L. Percolation and the effective structure of complex networks. *Phys. Rev. X* **9**, 011023. https://doi.org/10.1103/PhysRevX.9.011023 (2019).
24. Young, J.-G., St-Onge, G., Desrosiers, P. & Dubé, L. J. Universality of the stochastic block model. *Phys. Rev. E* **98**, 032309. https://doi.org/10.1103/PhysRevE.98.032309 (2018).
25. Riolo, M. A. & Newman, M. E. J. Consistency of community structure in complex networks. *Phys. Rev. E* **101**, 052306. https://doi.org/10.1103/PhysRevE.101.052306 (2020).
26. Meunier, D., Lambiotte, R. & Bullmore, E. T. Modular and hierarchically modular organization of brain networks. *Front. Neurosci.* **4**, 200. https://doi.org/10.3389/fnins.2010.00200 (2010).
27. Huttlin, E. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505–509. https://doi.org/10.1038/nature22366 (2017).
28. Gilarranz, L. J., Rayfield, B., Liñán-Cembrano, G., Bascompte, J. & Gonzalez, A. Effects of network modularity on the spread of perturbation impact in experimental metapopulations. *Science* **357**, 199–201. https://doi.org/10.1126/science.aal4122 (2017).
29. Orman, K., Labatut, V. & Cherifi, H. *Complex Networks*, vol. 424 of *Studies in Computational Intelligence*, chap. An empirical study of the relation between community structure and transitivity, 99–110 (Springer, Berlin, Heidelberg, 2013).
30. Lotfi, N., Rodrigues, F. A. & Darooneh, A. H. The role of community structure on the nature of explosive synchronization. *Chaos* **28**, 033102. https://doi.org/10.1063/1.5005616 (2018).
31. Fotouhi, B., Momeni, N., Allen, B. & Nowak, M. A. Evolution of cooperation on large networks with community structure. *J. R. Soc. Interface* **16**, 20180677. https://doi.org/10.1098/rsif.2018.0677(2019).
32. Giatsidis, C., Thilikos, D. M. & Vazirgiannis, M. Evaluating cooperation in communities with the k-core structure. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, 87–93, https://doi.org/10.1109/ASONAM.2011.65 (IEEE, 2011).
33. Salathé, M. & Jones, J. H. Dynamics and control of diseases in networks with community structure. *PLoS Comput. Biol.* **6**, 1–11. https://doi.org/10.1371/journal.pcbi.1000736 (2010).
34. Mistry, D., Zhang, Q., Perra, N. & Baronchelli, A. Committed activists and the reshaping of status-quo social consensus. *Phys. Rev. E* **92**, 042805. https://doi.org/10.1103/PhysRevE.92.042805 (2015).
35. Masuda, N. Voter model on the two-clique graph. *Phys. Rev. E* **90**, 012802. https://doi.org/10.1103/PhysRevE.90.012802 (2014).
36. Alvarez-Hamelin, J. I., DallAsta, L., Barrat, A. & Vespignani, A. K-core decomposition of internet graphs: hierarchies, self-similarity and measurement biases. *Netw. Heterogeneous Media* **3**, 371. https://doi.org/10.3934/nhm.2008.3.371 (2008).
37. Fosdick, B. K., Larremore, D. B., Nishimura, J. & Ugander, J. Configuring random graph models with fixed degree sequences. *SIAM Rev.* **60**, 315–355. https://doi.org/10.1137/16M1087175 (2018).
38. Massey, F. J. Jr. The Kolmogorov–Smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **46**, 68–78 (1951).
39. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech: Theory Exp.* **2008**, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008 (2008).
40. Karrer, B. & Newman, M. E. J. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**, 016107. https://doi.org/10.1103/PhysRevE.83.016107 (2011).
41. Fortunato, S. & Barthélemy, M. Resolution limit in community detection. *Proc. Nat. Acad. Sci. U.S.A.* **104**, 36–41. https://doi.org/10.1073/pnas.0605965104 (2007).
42. Lambiotte, R., Delvenne, J. C. & Barahona, M. Random walks, Markov processes and the multiscale modular organization of complex networks. *IEEE Trans. Net. Sci. Eng.* **1**, 76–90. https://doi.org/10.1109/TNSE.2015.2391998 (2014).
43. Lancichinetti, A., Fortunato, S. & Radicchi, F. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* **78**, 046110. https://doi.org/10.1103/PhysRevE.78.046110 (2008).
44. Faqeeh, A., Osat, S. & Radicchi, F. Characterizing the analogy between hyperbolic embedding and community structure of complex networks. *Phys. Rev. Lett.* **121**, 098301. https://doi.org/10.1103/PhysRevLett.121.098301 (2018).
45. Osat, S., Radicchi, F. & Papadopoulos, F. *k*-core structure of real multiplex networks. *Phys. Rev. Research* **2**, 023176. https://doi.org/10.1103/PhysRevResearch.2.023176 (2020).
46. Olhede, S. C. & Wolfe, P. J. Network histograms and universality of blockmodel approximation. *Proc. Nat. Acad. Sci. U.S.A.* **111**, 14722–14727. https://doi.org/10.1073/pnas.1400374111 (2014).
47. Mozafari, M. & Khansari, M. Improving the robustness of scale-free networks by maintaining community structure. *J. Complex Netw.* **7**, 838–864. https://doi.org/10.1093/comnet/cnz009 (2019).
48. Govindan, P., Wang, C., Xu, C., Duan, H. & Soundarajan, S. The k-peak decomposition: Mapping the global structure of graphs. In *Proceedings of the 26th International Conference on World Wide Web*, 1441–1450, https://doi.org/10.1145/3038912.3052635 (International World Wide Web Conferences Steering Committee, 2017).
49. Hébert-Dufresne, L., Grochow, J. A. & Allard, A. Multi-scale structure and topological anomaly detection via a new network statistic: the onion decomposition. *Sci. Rep.* **6**, 31708. https://doi.org/10.1038/srep31708 (2016).
50. Mastrandrea, R., Squartini, T., Fagiolo, G. & Garlaschelli, D. Enhanced reconstruction of weighted networks from strengths and degrees. *New J. Phys.* **16**, 043022. https://doi.org/10.1088/1367-2630/16/4/043022 (2014).
51. Yang, Z., Perotti, J. I. & Tessone, C. J. Hierarchical benchmark graphs for testing community detection algorithms. *Phys. Rev. E* **96**, 052311. https://doi.org/10.1103/PhysRevE.96.052311 (2017).
52. Kumpula, J. M., Onnela, J.-P., Saramäki, J., Kaski, K. & Kertész, J. Emergence of communities in weighted networks. *Phys. Rev. Lett.* **99**, 228701. https://doi.org/10.1103/PhysRevLett.99.228701 (2007).
53. Bianconi, G., Darst, R. K., Iacovacci, J. & Fortunato, S. Triadic closure as a basic generating mechanism of communities in complex networks. *Phys. Rev. E* **90**, 042806. https://doi.org/10.1103/PhysRevE.90.042806 (2014).
54. Shang, K.-k., Yang, B., Moore, J. M., Ji, Q. & Small, M. Growing networks with communities: a distributive link model. *Chaos: An Interdiscip. J. Nonlinear Sci.***30**, 041101, https://doi.org/10.1063/5.0007422 (2020).
55. McAuley, J. & Leskovec, J. Learning to discover social circles in ego networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems—Volume 1*, NIPS'12, 539–547 (Curran Associates Inc., Red Hook, NY, USA, 2012).
56. Rossi, R. A. & Ahmed, N. K. The network data repository with interactive graph analytics and visualization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, 4292–4293 (AAAI Press, 2015).
57. Traud, A. L., Mucha, P. J. & Porter, M. A. Social structure of Facebook networks. *Phys. A* **391**, 4165–4180. https://doi.org/10.1016/j.physa.2011.12.021 (2012).

58. Traud, A. L., Kelsic, E. D., Mucha, P. J. & Porter, M. A. Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev.* **53**, 526–543. https://doi.org/10.1137/080734315 (2011).
59. Kunegis, J. KONECT – The Koblenz Network Collection. In *Proceedings of the International Conference on World Wide Web Companion*, 1343–1350 (2013).
60. Newman, M. E. J. Network data repository: Political blogs dataset. http://www-personal.umich.edu/~mejn/netdata/. Accessed on 01/10/2019.
61. The OpenFlights database. https://openflights.org/data.html. Accessed on 01/10/2019.
62. Cookpad: Make everyday cooking fun! https://cookpad.com/. Accessed on 01/10/2019.
63. Kitsak, M. *et al.* Identification of influential spreaders in complex networks. *Nat. Phys.* **6**, 888–893. https://doi.org/10.1038/nphys1746 (2010).
64. Bollobás, B. *Modern graph theory*. Graduate Texts in Mathematics 184 (Springer-Verlag New York, 1998).
65. Fagin, R., Kumar, R. & Sivakumar, D. Comparing top k lists. *SIAM J. Discrete Math.* **17**, 134–160. https://doi.org/10.1137/S0895480102412856 (2003).
66. McCown, F. & Nelson, M. L. Agreeing to disagree: search engines and their public interfaces. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, 309–318, https://doi.org/10.1145/1255175.1255237 (2007).
67. Holland, P. W., Laskey, K. B. & Leinhardt, S. Stochastic blockmodels: first steps. *Soc. Netw.* **5**, 109–137. https://doi.org/10.1016/0378-8733(83)90021-7 (1983).
68. Peixoto, T. P. Nonparametric bayesian inference of the microcanonical stochastic block model. *Phys. Rev. E* **95**, 012317. https://doi.org/10.1103/PhysRevE.95.012317 (2017).
69. Peixoto, T. P. The graph-tool python library. *figshare*. https://doi.org/10.6084/m9.figshare.1164194 (2014).
70. Oliphant, T. *Guide to NumPy* (Trelgol Publishing, 2006).
71. van der Walt, S., Colbert, S. C. & Varoquaux, G. The numpy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30. https://doi.org/10.1109/MCSE.2011.37 (2011).
72. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference* (eds. Varoquaux, G., Vaught, T. & Millman, J.) 11 – 15 (Pasadena, CA USA, 2008).
73. Hunter, J. D. Matplotlib: a 2d graphics environment. *Comput. Sci. Eng.* **9**, 90–95. https://doi.org/10.1109/MCSE.2007.55 (2007).
74. Stanford Network Analysis Project (SNAP): "social circles: Facebook" dataset. http://snap.stanford.edu/data/ego-Facebook.html. Accessed on 01/10/2019.
75. Network Repository: "American75" dataset. http://networkrepository.com/socfb-American75.php. Accessed on 01/10/2019.
76. Network Repository: "Amherst41" dataset. http://networkrepository.com/socfb-Amherst41.php. Accessed on 01/10/2019.
77. Network Repository: "Cal65" dataset. http://networkrepository.com/socfb-Cal65.php. Accessed on 01/10/2019.
78. Network Repository: "FSU53" dataset. http://networkrepository.com/socfb-FSU53.php. Accessed on 01/10/2019.
79. Stanford Network Analysis Project (SNAP): "social circles: Twitter" dataset. http://snap.stanford.edu/data/ego-Twitter.html. Accessed on 01/10/2019.
80. Adamic, L. A. & Glance, N. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd International Workshop on Link discovery*, 36–43, https://doi.org/10.1145/1134271.1134277 (ACM, 2005).
81. Stanford Network Analysis Project (SNAP): "email-EU-core network" dataset. http://snap.stanford.edu/data/email-Eu-core.html. Accessed on 01/10/2019.
82. Yin, H., Benson, A. R., Leskovec, J. & Gleich, D. F. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 555–564, https://doi.org/10.1145/3097983.3098069 (ACM, 2017).
83. Leskovec, J., Kleinberg, J. & Faloutsos, C. Graph evolution: densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data (TKDD)* **1**, 2 (2017).
84. Stanford Network Analysis Project (SNAP): "Condensed Matter collaboration network" dataset. https://snap.stanford.edu/data/ca-CondMat.html. Accessed on 01/10/2019.
85. The KONECT Project: DBLP co-authorship network dataset. http://konect.cc/networks/com-dblp (2017). Accessed on 01/10/2019.
86. Yang, J. & Leskovec, J. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Min. Data Semant.*, 3, https://doi.org/10.1007/s10115-013-0693-z (2012).
87. The KONECT Project: WordNet network dataset. http://konect.cc/networks/wordnet-words (2017). Accessed on 01/10/2019.
88. Fellbaum, C. (ed.) *WordNet: an Electronic Lexical Database* (MIT Press, Cambridge, 1998).

Acknowledgements

The authors thank A. Barrat for helpful comments on a preliminary version of this work. A.C. acknowledges the support of the Spanish Ministerio de Ciencia e Innovacion (MICINN) through Grant IJCI-2017-34300. I.M., A.C., and N.M. acknowledge the support of Cookpad Limited. This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol – http://www.bristol.ac.uk/acrc/. Numerical analysis has been carried out using the NumPy, NetworkX, and Graph-tool Python packages[69–72]. Graphics have been prepared using the Matplotlib Python package[73].

Author contributions

I.M. analysed data, designed the experiments, and performed simulations, A.C. analysed the results and designed the experiments. All authors discussed the methods and results, and wrote and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-71426-8.

Correspondence and requests for materials should be addressed to A.C. or N.M.